Lecture Notes on Data Engineering
and Communications Technologies 10

Georgios Skourletopoulos
George Mastorakis
Constandinos X. Mavromoustakis
Ciprian Dobre · Evangelos Pallis *Editors*

Mobile Big Data

A Roadmap from Models to Technologies

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 10

Series editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain
e-mail: fatos@cs.upc.edu

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

More information about this series at http://www.springer.com/series/15362

Georgios Skourletopoulos
George Mastorakis
Constandinos X. Mavromoustakis
Ciprian Dobre · Evangelos Pallis
Editors

Mobile Big Data

A Roadmap from Models to Technologies

Springer

Editors
Georgios Skourletopoulos
Department of Computer Science
University of Nicosia
Nicosia
Cyprus

George Mastorakis
Department of Business Administration
Technological Educational Institute Crete
Agios Nikolaos
Greece

Constandinos X. Mavromoustakis
Department of Computer Science
University of Nicosia
Nicosia
Cyprus

Ciprian Dobre
Department of Computer Science and Engineering
University Politehnica of Bucharest
Bucharest
Romania

Evangelos Pallis
Department of Informatics Engineering
Technological Educational Institute of Crete
Heraklion
Greece

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-3-319-67924-2 ISBN 978-3-319-67925-9 (eBook)
https://doi.org/10.1007/978-3-319-67925-9

Library of Congress Control Number: 2017954873

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The usage of mobile devices steadily grows causing an enormous rise in the mobile data traffic over the Internet. Data is produced by handheld, pervasive and wearable devices, which are configured for collecting and delivering data to related servers that host online and mobile social networks. Notwithstanding, the enormous amount, multi-source collection of data motivates the need to investigate novel access methods, mobile big data collecting techniques, methods to improve the integration of resources' availability through the 3As ('Anywhere, Anything, Anytime') paradigm, distributed big data storing methodologies, and intra- or inter-big data processing mechanisms.

The main target of this book is to give an overview of the great emerging advances and challenges in mobile technologies for collecting, storing and processing mobile big data from an engineering perspective, discussing a wide range of applications and scenarios where mobile big data can be applied. Future directions on theories, practices, standards and strategies that are related to this research domain are discussed. This timely volume includes thirteen rigorously refereed chapters from prominent international researchers and serves as a source of new schemes in the mobile big data field for students, researchers, scientists and practitioners. It may be used in undergraduate and graduate courses on the design and development of mobile big data-driven systems and applications. Researchers and scientists will find the book useful as it provides a current state-of-the-art guide and future trends in mobile big data analytics and management. Finally, practitioners will broaden their expertise on particular topics and methodologies, such as the transition of mobile big data to the Cloud, context-awareness in the mobile big data paradigm and the processing of real-time streaming events on-the-move considering the need for high-velocity processing and low latency response.

Nicosia, Cyprus — Georgios Skourletopoulos
Agios Nikolaos, Greece — George Mastorakis
Nicosia, Cyprus — Constandinos X. Mavromoustakis
Bucharest, Romania — Ciprian Dobre
Heraklion, Greece — Evangelos Pallis

Acknowledgements

This book has been made possible by the great efforts and contributions of many people. First of all, we would like to thank all the contributors for putting together excellent chapters that are very comprehensive and informative. Second, we would like to thank all the reviewers for their valuable suggestions and comments, which have greatly enhanced the quality of this book. Finally, we would like to thank the staff members from Springer International Publishing for putting this book together and assisting us throughout the process.

Contents

Part I
Introduction to Mobile Big Data Paradigm

Big Data Analytics: Applications, Prospects and Challenges

Konstantinos Vassakis, Emmanuel Petrakis and Ioannis Kopanakis

Abstract In the era of the fourth industrial revolution (Industry 4.0), big data has major impact on businesses, since the revolution of networks, platforms, people and digital technology have changed the determinants of firms' innovation and competitiveness. An ongoing huge hype for big data has been gained from academics and professionals, since big data analytics leads to valuable knowledge and promotion of innovative activity of enterprises and organizations, transforming economies in local, national and international level. In that context, data science is defined as the collection of fundamental principles that promote information and knowledge gaining from data. The techniques and applications that are used help to analyze critical data to support organizations in understanding their environment and in taking better decisions on time. Nowadays, the tremendous increase of data through the Internet of Things (continuous increase of connected devices, sensors and smartphones) has contributed to the rise of a "data-driven" era, where big data analytics are used in every sector (agriculture, health, energy and infrastructure, economics and insurance, sports, food and transportation) and every world economy. The growing expansion of available data is a recognized trend worldwide, while valuable knowledge arising from the information come from data analysis processes. In that context, the bulk of organizations are collecting, storing and analyzing data for strategic business decisions leading to valuable knowledge. The ability to manage, analyze and act on data ("data-driven decision systems") is very important to organizations and is characterized as a significant asset. The prospects of big data analytics are important and the benefits for data-driven organizations are

K. Vassakis (✉) · E. Petrakis
Department of Economics, University of Crete, Gallos Campus,
Rethymno, Crete 74100, Greece
e-mail: k.vassakis@e-bilab.gr

E. Petrakis
e-mail: petrakis@uoc.gr

I. Kopanakis
Department of Business Administration, Technological Educational Institute of Crete, Agios Nikolaos, Crete 72100, Greece
e-mail: i.kopanakis@teicrete.gr

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_1

significant determinants for competitiveness and innovation performance. However, there are considerable obstacles to adopt data-driven approach and get valuable knowledge through big data.

Keywords Big data · Big data analytics · Performance · Enterprises Knowledge management · Internet of things (IoT)

1 Introduction

Data is characterized as the *lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time, designing, monitoring and evaluating effective policies becomes almost impossible* [1]. In that context, an ongoing attention to data and data-driven approaches from academics and professionals exists, since the knowledge arising from data analysis processes leads to the promotion of innovative activity, transforming organizations, enterprises and national economies.

Nowadays, in the 4th Industrial revolution era, organizations and governments focus on the development of capabilities that provide knowledge extracted from large and complex data sets, commonly known as "big data". Big data is a buzzword in the last years in the business and economics fields, since it plays an essential role in economic activity and has strengthened its role in creating economic value by enabling new ways to spur innovation and productivity growth. Hence, the ability of management, analysis and acting is significant under the context of knowledge-based capital (KBC) that is associated with digital information, innovative capacity and economic aspects [2].

In that era, many enterprises independent size, from start-ups to large organizations, attempt to obtain data-driven culture struggling for competitive advantage against rivals. Enterprises aim to leverage data generated within organizations through their operations to gain valuable insights for better, faster and more accurate decisions in crucial business issues.

The advent of the Web 2.0 allows users interacting with each other on social media platforms, enabled companies getting access to big amounts of data easier and cheaper. In addition, the appearance of Web 3.0 provides considerably increased opportunities for external data collection. Mobile devices (smart phones and tablets) that facilitate companies to measure even more precisely, since those devices, both Internet and mobile-enabled, have the capability to promote e.g. highly mobile, location-aware and person-centered processes and transactions. This capability will continue offering unique research challenges and opportunities through the years [3].

Digital enterprises like Google, Amazon and Facebook highlight the significance of big data, indicating the various ways that can be used from supply chain to customer satisfaction highlighting the benefits of enterprises. Many enterprises started to benefit from those opportunities offered by the immense development of big data technologies. Today, enterprises in every industry sector and not limited to

ICT sector, are focused on data exploitation to gain a competitive advantage, while managerial decisions rely on data-based analytics and less on the leader's experience [4]. Nonetheless, exploitation of big data needs people with skills and expertise who will able to capture value from data insights providing significant knowledge to managers and decision-makers.

1.1 Defining Big Data

The tremendous generation of data, expected to reach 180 ZB in 2025, give data a leading role in change and growth of the 21st-century shaping a new "digital universe" with the transformation of markets and businesses [5]. Digital information from complex and heterogeneous data coming from anywhere and at any time introducing a new era, the era of "Big Data" [6].

Big data refers to large datasets that are not able to be captured, stored, managed and analyzed by typical software tools [7]. These data sets that are huge -not only in size- but also in heterogeneity and complexity (structured, semi-structured and unstructured data) including operational, transactional, sales, marketing and other data. In addition, big data includes data that comes in several formats including text, sound, video, image and more. This unstructured data is growing faster than structured and have captured the 90% of all the data [8]. Therefore, new forms of processing capabilities are required for getting data insights that lead to better decision making.

On the data life cycle the challenges can be divided into three categories: data, process and management challenges (Fig. 1) [6]. Data challenges refer to characteristics of big data including volume, velocity, variety and veracity. Process challenges are related with the techniques needed for big data acquisition, integration, transformation and analysis in order to gain insights from the big data. The data management challenges include challenges regarding data security, privacy, governance and cost/operational expenditures.

Big data can be characterized by the seven Vs: volume, variety, veracity, velocity, variability, visualization and value.

Volume refers to the large size of the datasets. It is fact that Internet of Things (IoT) through the development and increase of connected smartphones, sensors and other devices, in combination with the rapidly developing Information and Communication Technologies (ICTs) including Artificial Intelligence (AI) have contributed to the tremendous generation of data (counting records, transactions, tables, files etc.). The speed of data is surpassing Moore's law and the volume of data generation introduced new measures for data storage i.e. exabytes, zettabytes and yottabytes.

Variety represents the increasing diversity of data generation sources and data formats. Web 3.0 leads to growth of web and social media networks leading to the generation of different types of data. From messages, updates, photos and videos that are posted in social media networks like Facebook or Twitter, SMS, GPS

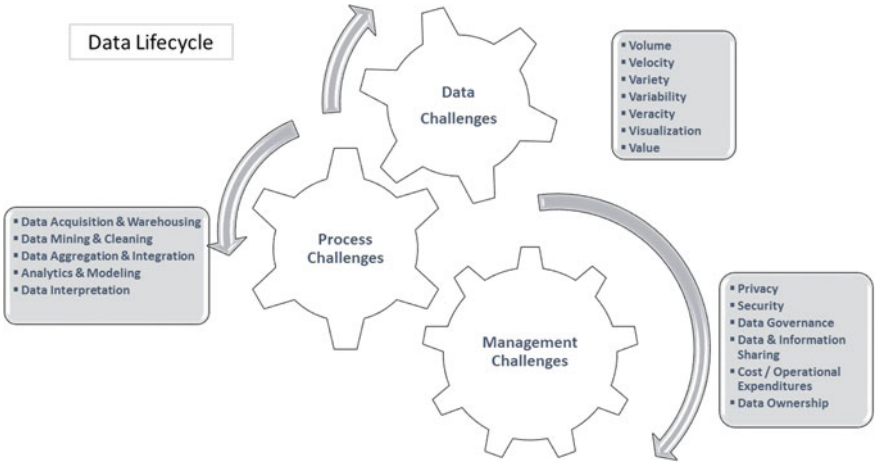


Fig. 1 Challenges in data lifecycle

signals from smartphones, customers transactions in banking, e-business and retail, voice data in call centers etc. Many of the crucial sources of big data are comparatively novel, including mobile devices that supply huge streams of data that are connected with human behavior through their activities and locations; or web sources supplying data through comprising logs, click-streams and social media actions. Additionally, big data also differs in data types that are generated, thus big data consists on structured data (tables, records), unstructured data (text and voice), semi-structured data (XML, RSS feeds) and other data that is difficult to classify like data deriving from audio, video and other appliances.

Variability is often confused with variety, but variability is related with rapid change of meaning. For instance, words in a text can have a different meaning according to context of a text, thus for an accurate sentiment analysis, algorithms need to find out the meaning (sentiment) of a word taking into account the whole context.

Velocity. Big data is characterized by the high speed of data generation. Data generated by connected devices and web arriving in enterprises in real-time. This speed is extremely significant for enterprises in taking various actions that enable them to be more agile, gaining competitive advantage against competitors. Despite the fact that some enterprises have already exploited big data (click-streams data) to offer their customers purchase recommendations, nowadays enterprises though big data analytics have the ability to analyze and understand data taking actions in real-time.

Veracity of data refers to data reliability and accuracy. The data collection has data that are not clean and accurate, thus data veracity refers to the data uncertainty and the level of reliability correlated with some type of data.

Visualization. Data visualization is the science of visual representation of data and information. It presents quantitative and qualitative information in some schematic form, indicating patterns, trends, anomalies, constancy, variation, in ways that cannot be presented in other forms like text and tables [9].

The leverage of big data can provide valuable knowledge and thus the value offered by the data analysis process can benefit enterprises, organizations, communities and consumers.

Enterprises that overcome challenges and exploit big data efficiently have more precise information and are able to create new knowledge by which they can improve their strategy and business operations regarding well-defined targets like productivity, financial performance and market value [10], while big data plays a major role in digital transformation of enterprises introducing innovations. Therefore, an increasing interest in exploitation of big data among enterprises and organizations exists (Fig. 2).

The economic benefits of big data in UK private and public-sector businesses will increase from £25.1 billion in 2011 to £216 billion in 2017 [11]. Big data can provide more value in enterprises in various ways and is able to enhance productivity and competitiveness of enterprises. Big data is referred to the continuous growth of data and technologies that are necessary for collection, storage, management and analysis of data. The way of thinking about businesses has changed with big data, since it changes major elements of organizations and not only management. Big data can be a key resource for enterprises obtaining new knowledge, added value and fostering new products, processes and markets, thus data is characterized as an asset from enterprises' executives indicating the significance of data-driven approach within enterprises [12]. Enterprises gathered data for ages, however, nowadays more and more enterprises are actually analyzing the data instead of just keeping them. Hence, data-driven enterprises perform better in financial and operational terms, 5% more productive and 6% more profitable than no data-driven, gaining significant competitive precedence against their competitors [13].

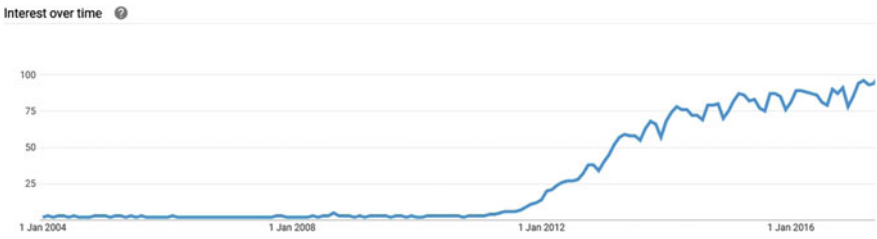


Fig. 2 Big data trend https://trends.google.com/trends/explore?date=all&q=%2Fm%2F0bs2j8q

1.2 Big Data Analytics

The analysis of large data-sets in enterprises, the term of big data analytics is associated with data science, business intelligence and business analytics. Data science is defined as a collection of fundamental principles that promotes taking information and knowledge from data [4]. Over the last years, data-driven approaches like Business Intelligence (BI) and Business Analytics are characterized indispensable to operating enterprises. BI is defined as the methodologies, systems and applications for collecting, preparing and analyzing data to provide information helping decision makers. In other words, BI systems are data-driven decision making systems [14], while Business Analytics are the techniques, technologies, systems and applications that are used to analyze critical business data for supporting them to understand their business environment and take business decisions on time. The power of Business Analytics is to streamline vast amounts of data to enhance its value, while BI mainly concentrates historical data in graphs and data table reports as a way to provide answers to queries without streamlining data and enhancing its value.

Business Analytics was commenced to outline the principal analytical element in BI in the late 2000s. Afterwards, the terms of big data and big data analytics have been utilised to describe analytical techniques for data- sets that are so large and complex, needing advanced data storage, management, analysis and visualization technologies. In that rapidly growing environment, the velocity of data makes the conversion of data into valuable knowledge quickly a necessity. The differences between conventional analytics and fast analytics with Big data are in analytics characteristics (type, objective and method), data characteristics (type, age/flow, volume) and primary objective (Table 1) [15, 16].

The development of the Internet and later on the connectivity coming from the web has contributed in the increase of the volume and speed of data. Since the early 2000s, Internet and Web technologies have been offering unique data collection and

Table 1 Conventional and big data analytics

	Conventional analytics	Big data analytics
Analytics type	Descriptive, Predictive	Predictive, Prescriptive
Analysis methods	Hypothesis-based	Machine learning
Primary objective	Internal decision support and performance management	Business processes driver and data-driven Products
Data type	Structured and defined (formatted in rows & columns)	Unstructured and undefined (unstructured formats)
Data age/flow	>24 h Static pool of data	<Min Constant flow of data
Data volume	Tens of terabytes or less	100 terabytes to petabytes

analysis for enterprises. Web 1.0 systems enable enterprises to establish a web presence and offer their products/services online interacting with their customers. Web 2.0 systems, including the introduction of social media networks like Facebook, provide enterprises more data with information about enterprises, products and customers. The ongoing increase of mobile devices against the number of computers introduced a new era of business analytics, including the analysis of user-generated content by social media channels. Mobile devices have the capability to promote e.g. highly mobile, location-aware and person-centered processes and transactions. Therefore, Data-driven decision making is on data coming from all the sources of enterprises, while predictions and machine learning are based on traditional data and new innovative sources like IoT and AI.

Data analysis is the process of inspecting, cleaning, transforming and modeling data gaining useful information for suggestions and support in decision-making. It has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science and social science schemes, while "Big Data Analytics" refers to advanced analytic techniques, considering large and various types of datasets to examine and extract knowledge from big data, constituting a sub-process in gaining insights from big data process. Using advanced technologies, Big Data Analytics (BDA) includes data management, open-source programming like Hadoop, statistical analysis like sentiment and time-series analysis, visualization tools that help structure and connect data to uncover hidden patterns, undiscovered correlations and other actionable insights.

The process of BDA is a resource for strategic decisions leading to significant improvements in operations performance, new revenue streams and competitiveness against rivals. In that context, the process of getting insights from big data can be divided into two phases: data management and data analysis. Data management is related with the processes and technologies for data generation, storage, mining and preparation for analysis, while data analysis refers to the methods and techniques for analysis and interpretation of the insights coming from big data [17] (Fig. 3).

Analytics can be divided into four categories, ranging from descriptive and diagnostic analytics to the more advanced predictive and prescriptive analytics.

Fig. 3 Process of leveraging big data

Descriptive analytics, based on historical and current data, is a significant source of insights about what happened in the past and the correlations between various determinants identifying patterns using statistical measures like mean, range and standard deviation. Descriptive analytics using techniques like online analytical processing (OLAP) exploits knowledge from the past experience to provide answers in what's happening in the organizations. Common examples of descriptive analytics include data visualization, dashboards, reports, charts and graphs presenting key metrics of enterprises including sales, orders, customers, financial performance etc.

Diagnostic analytics based also in historical data provide insights about the root-cause of some outcomes of the past. Thus, organizations can take better decisions avoiding errors and negative results of the past.

Predictive analytics is about forecasting and providing an estimation for the probability of a future result, defining opportunities or risks in the future. Using various techniques including data mining, data modeling and machine learning, the implementation of predictive analytics is significant for any organization's segment. One of the most known applications of that type of analytics is the prediction of customer behavior, determining operations, marketing and preventing risk. Using historical and other available data, predictive analytics are able to uncover patterns and identify relationships in data that can be used for forecasting [17]. Predictive analytics in the digital era is a significant weapon for organizations in the competitive race. Therefore, organizations exploiting predictive analytics can identify future trends and patterns, presenting innovative products/services and innovations in their business models.

Prescriptive analytics provide a forecasting of the impact of future actions before they are taken, answering "what might happen" as outcome of the organization's actions. Therefore, the decision-making is improved taking under consideration the prediction of future outcomes. Prescriptive analytics using high level modeling tools is able to contribute remarkably to the performance and efficiency of organizations, through smarter and faster decision with lower cost and risk and identifying optimal solutions for resource allocation [18].

The advanced predictive and prescriptive analytics can play crucial role in efficient strategic decision making dealing with significant problems of organizations like design and development of products/services, supply chain formation etc. [19].

1.2.1 Big Data Analytics Applications

Nowadays, as the growing generation of available data is a recognized trend across enterprises, countries and market segments, the majority of enterprises regardless industry is collecting, storing and analyzing data in order to capture value. Digital economy through the tremendous use of internet and digital services has transformed almost all the industry sectors, including agriculture and manufacturing, to more service-centered [20]. There are many and different sectors, like e-commerce,

politics, science & technology, health, government services etc., where big data analytics are applied. Data-driven companies from various industries clarify the power of big data, making more accurate predictions leading on better decisions.

The large streams of data generated everyday need better infrastructures in order to be captured, stored and analyzed. A market with a wide supply of new products and tools designed to cover all the needs of big data has been created and it is developing rapidly [21]. There is a wide variety of analytic tools that can be used to perform BDA, among others on the basis of SQL queries, statistical analysis, data mining, fast clustering, natural language processing, text analytics, data visualization and artificial intelligence (AI). These techniques and tools provide easily and rapidly exploitation of big data.

The knowledge derived from exploitation of big data provides enterprises added value through new ways of productivity, growth, innovation and consumer surplus [7], thus big data becomes a major determinant of competitiveness and enterprises are in need of data analysis capacity to exploit the full potential of data.

Enterprises that learn to capitalize big data utilizing real-time information coming from various sources like sensors, connected devices etc. can understand in more detail their environment and define new trends, create new and innovative products/services, respond quickly in changes and optimize their marketing actions. The leverage of big data is able to contribute to the efficient resources' allocation and supervision, waste reduction, facilitation of new insights and higher level of transparency in different sections of enterprises from production to sales.

Therefore, BDA applications in almost every business sector exist. Applications also in politics and e-government, science and technology, security and safety, smart health and well-being exist [3]. In addition, there are plenty and various types of big data applications among enterprises and industry sectors. BDA can be employed in e-commerce and marketing applications like online advertising and cross-selling, while it helps enterprises to analyze customer behavior in shaping 360-degree customer profile for implementation of targeted and optimized marketing actions to impact customer acquisition and satisfaction. It offers better understanding of customers' behavior and preferences and thus improve customer service.

Some examples of the ways BDA are exploited showing the significance of analytics in various themes [22]:

Marketing	Market basket analysis	Recommendation systems	Customer Intelligence	Retention modeling	Customer churn prediction
Processes	Supply chain analytics	Demand and supply forecasting	Business Processes analytics	HR analytics	
Government	Fraud detection	Terrorism Detection	Tax avoidance	Cost reduction	Social security

(continued)

(continued)

Risk Management	Credit risk modeling	Market risk modeling	Fraud detection		
Web and Social media	Web analytics	Social media analytics	Multivariate testing		

Enterprises and organizations collect large amounts of security-relevant data such as software application events, network events, people's action events. The generation of data coming from these actions are increasing rapidly per day as organizations enable logging in more sources, running more software programs, have more working employees and move to cloud solutions. Unfortunately, the volume and variety of security data quickly become overwhelming and existing analytical techniques cannot work efficiently and trustworthily. BDA applications become part of security management and monitoring, since it contributes to cleaning, preparation and analysis of various complex and heterogeneous datasets efficiently [23]. One of the most common uses of BDA is fraud detection, thus financial institutions, governments and phone companies use big data technologies to eliminate risk and enhance their efficacy.

In addition, BDA is widely applied in supply chain and logistics operations playing a significant role in developing supply chain strategies and supply chain operations management. BDA can support decision making through the understanding of changes of marketing conditions, identification of supply chain risks and exploiting supply chain capabilities to model innovative supply chain strategies, thereby improving the flexibility and profitability of supply chain. BDA contributes also in decision making at operational level, since it measures and analyses supply chain performance taking into account demand planning, supplies, production, inventory and logistics. It thus improves efficiency of operations, measures supply chain performance, reduces process alterability and contributes to the implementation of the best supply chain strategies at operational level [24].

Talking about digital and data-driven enterprises, the firsts coming in mind are Google, Amazon, Apple and Facebook. Amazon that was born digital, exploited big data achieving to disrupt traditional book market and became the leader in digital shopping. Another example of a famous born-digital firm is Google that harness data from engine search to digital marketing in order to provide and personalize search to its users, while Google and Facebook collect data providing opportunities for personalized and customized marketing.

Nevertheless, traditional non-technological enterprises are also attempting to gain data-driven benefits. General Electric (GE) has developed a cloud-based platform for Industrial Internet application named "Predix" that provides real-time insights for engineers to schedule maintenance checks, improves machine efficiency and reduces downtime. GE this way provided new service value propositions in the conservative market of the oil and gas industry, while it faces its most pressing challenges: improving assets and operations productivity and eliminating the cost of tacit knowledge from aging workforce [25].

Walmart and other major retailers using BDA in the entire business process, from supply-chain management to marketing, gained benefits from data. Applications of BDA are everywhere and not only in digital sectors, but also in no web-based sectors including manufacturing, agriculture, health care, energy, traveling and others. In healthcare sectors, various applications of BDA exist, from quality of treatment services and cost efficiency of hospitals to improvement and predictions of patient health condition. In traveling and retail, BDA applications are able to provide customer intelligence through web and social media analytics, thus enterprises can offer personalized products/services. Additionally, in energy management the majority of the enterprises use data analytics to track and control devices achieving a more efficient energy management without services deviation.

1.2.2 Big Data Analytics Prospects

Analytics in decision making procedure is not something new, since business analytics appeared as early as in the mid1950s—Analytics 1.0 era—with the advent of tools that were able to generate and capture larger amounts of data in enterprises data warehouses and discover patterns more quickly than human minds with business intelligence tools. In that first era, managers gained a data-based comprehension going beyond intuition in decision making. Until mid-2000s, the rapid growth of data generation and the arrival of big data have signaled a new era—Analytics 2.0—where enterprises have the opportunity to leverage that data with new more powerful tools. The need of new innovative technologies appeared and enterprises moved quickly to acquire the necessary capabilities and knowledge for gaining insights from big data, with the major difference between eras being in skills required for data analysis [26]. In the next era, analytics is an integral part of enterprises supporting decision making and enterprises move to creation of analytics-based products/services. Moving ahead, the next era—Analytics 3.0 or "data economy"—is characterized by the tremendous increase of data generation coming from the growth of Internet of Things (IoT) with 8.4 billions connected devices in 2017 globally and 20.4 billion by 2020 [27].

The most recent era—Analytics 4.0—includes cognitive technologies including machine learning, where actions and decision making are shifted to augmentation with dynamic machine automation. The main characteristics of all these eras are appeared in Fig. 4 [28].

In the current era of analytics, the emerging new technologies will increase the generation of data, thus enterprises and organizations have to face up technical challenges in order to have access to more and better data. The worldwide revenues of big data and business analytics (BDA) will be more than $203 billion in 2020 and banking, manufacturing, government and professional services will be the top industries in BDA investments according to International Data Corporation (IDC) [29].

Therefore, enterprises should focus on capturing value from data using analytical techniques and tools. BDA can help enterprises to examine trends and discover new

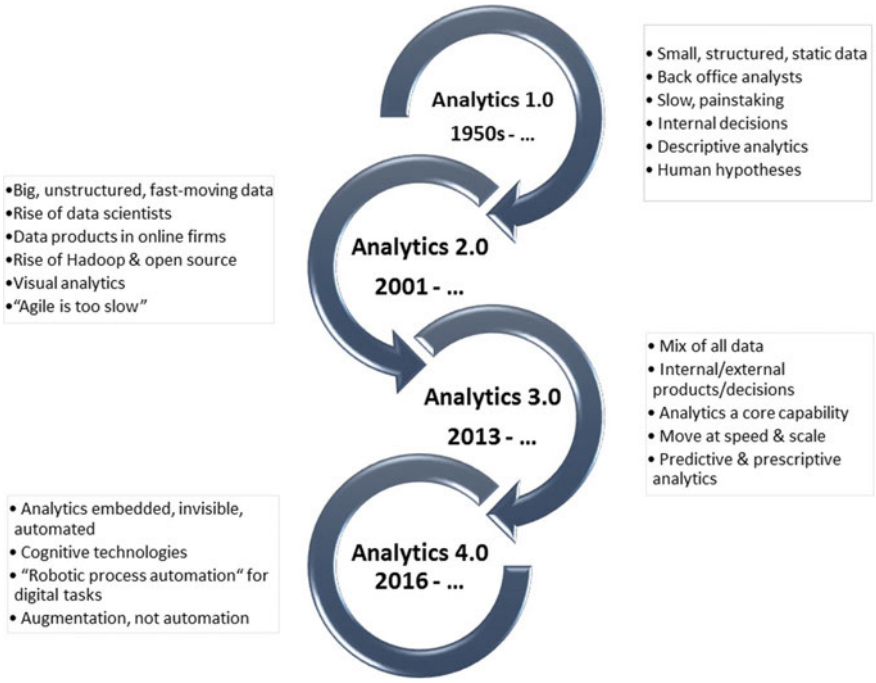


Fig. 4 The evolution of analytics eras

ones for gaining competitive advantage, introducing new and improved products. Among others, data visualization and process simulation, text and voice analytics, social media analysis, predictive and prescriptive techniques can provide valuable knowledge to enterprises, while they are able to make insights more transparent and impact any enterprise's section.

Data science and big data technologies—techniques promote data-driven decision making and thus contribute in better enterprise's performance, since the ultimate goal of data science is the improvement of decision making. Therefore, whether organizations couldn't capture value from applying data-driven decision making as their strategy, they have failed [4]. There is evidence that data-driven decision making contributes significantly and positively to enterprise's performance in terms of productivity and profitability [30]. Data-driven approach can provide great opportunities for gaining competitive advantage, as measuring and managing more precisely business analytics can enable organizations to make better predictions and smarter decisions also to target more-effective interventions [13].

Moving to a whole new era in data analytics, organizations and enterprises are exploring new innovative strategies and techniques to remain competitive in their market. Using BDA help them to introduce new and/or improved products/services, manage more efficiently their supply chains and processes, eliminate risk through fraud detection and security improvement and exploit customer intelligence.

Applications of BDA can provide several advantages in organizations and enterprises that have an efficient data-driven approach. Big data analysis is able to provide in-depth knowledge about the different departments of an organization and thus using big data analytics for prediction making will contribute to increased performance and higher returns on investments with lower cost and risk, while more transparency is achieved.

Some of the prospects of big data analytics are:
Gaining insights from big data analytics of all the departments of an organization to develop a comprehensive business strategy, or the entire organization. This strategy will be able to contribute to higher level of productivity and efficiency, within the departments, but also in the whole organization with cost reduction and elimination of processes.

Organizations will exploit more artificial intelligence (AI) technologies that are able to reinvent organizations in various ways. However, organizations should develop automations and structured analytics, before they move on the adoption of advanced AI. The integration of structured and unstructured data analytics with AI systems makes it possible to examine, explain and predict customer preferences and behavior [31].

Data-driven innovation (DDI) relying on the knowledge-based capital, refers to innovations arising from data-driven decision processes [2] that lead to the discovery of new and disruptive business models, the enhancement of customer intelligence [32] and the introduction of new/improved products or services. The potential of data-driven innovation big data in UK private and public sector businesses will lead to £24.1 billion contribution to UK economy during 2012–2017 [11].

Real- time analytics is a big trend that enterprises need to pay attention at in the near future. Despite the challenges and issues that are addressed, it is proven that analytics-driven management has significant implications on enterprises, whether they are looking for growth, efficiency or competitive differentiation. Therefore, Big data analytics have seemingly unlimited potential to help an enterprise to grow and reveal its data potential.

The rapid growth of the demand for data analytics in combination with the lack of talent lead on collaborations and initiatives between academia and industry in order to bridge the talent gap. In that context, many universities are preparing and starting academic courses related with data science. In addition, companies realizing the potential of big data, provide training to their employees. Recently AirBnB started its own internal university called “Data University” to democratize data science and help to drive data-informed decision making.

There are different expectations from enterprises regarding big data analytics. Organizational leaders want to exploit analytics to be smarter and innovative like never before, while senior executives want to use data-driven decision making for their efficient operations [33]. Managers using a data-driven decision system (DSS), have access to historical and new data supporting them to gain insights for organization processes and resources’ performance. DDS are significant not only for

global organizations but also for small and medium organizations that can exploit them to their benefit [10].

1.2.3 Big Data Analytics Challenges and Barriers

The major challenges in adopting big data analytics from enterprises are more managerial and cultural than associated with data and technology, while the main barriers are the lack of comprehension of how to utilize big data analytics to enhance the business and the lack of management spectrum from competing priorities [33]. Studies among different industry sectors indicate that organizations use less than half of their structured data in decision making process, while less than 1% of their unstructured data is analyzed or exploited, 70% of employees have access to data they should not and 80% of analysts' time is to discover and prepare data [34].

Leadership. According to management challenges, enterprises that achieve to be successful in the data-driven era have leadership teams that determine aims, modulate achievements and ask the right questions to be answered by data insights. Despite its technological approach, the power of big data cannot be exploited without vision or human insight. Therefore, leaders of enterprises with vision and ability of revealing the future trends and opportunities, will have the ability to act innovative, motivate their teams work efficiently to achieve their targets.

Talent management. Enterprises in order to leverage data through big data analytics need human capital with high level of technical skills to use and exploit these systems in order to achieve exploitable knowledge for end users, mainly C-suite. People's specific skills include statistics, big data mining, master visualization tools, business oriented mindset and machine learning. These are required to get valuable insights from big data contributing in decision making procedure [13]. However, these people (data scientists, data analysts etc.) are extremely difficult to be found and thus demand for them is high. There is a challenge in finding data scientists with skills both in analytics and in domain knowledge. In general, there are existing fewer data scientists than needed [35].

Decision making procedure. In efficient enterprises, decision makers and knowledge derived from data exploitation are in the same place. Nonetheless, it is difficult for decision makers to handle huge amounts of data. Therefore, there is need of decision-makers having problem-solving skills and the ability to provide answers to problems with the right data or cooperation of different people in problem solving through leveraging big data [13].

Decision making Quality. The quality of decision making adopting a data-driven approach is a significant factor for taking advantage of the possibilities that big data analytics are offering. In that context, ensuring decision making quality is correlated with factors like data quality of big data sources, big data analytics capabilities, staff and decision-maker quality [36]. The accuracy of big data sources is significant in providing high value in decision making eliminating wrong actions, while big data

analytics capabilities are related with the utilization of the right techniques and tools from specialists with knowledge of big data analytics.

Data-driven culture. Another significant challenge for adopting data-driven approach is enterprise culture. The basis in obtaining data-driven culture is the capabilities to quickly condense, analyze and distribute crucial business information to decision makers. That basis is extremely significant for enhancement of business performance, while development and improvement of that capabilities empower enterprises leading to improvements in all business segments and higher returns on investments. In that context, enterprises have to adopt data-driven decision making in all issues and stop acting solely on hunches and instinct. Therefore, management must fully understand the significance of getting insights from data exploitation. In addition, for a data-driven enterprise, people who are involved in the process of data-driven decision making need to meet some requirements. Managers should be able to manage efficient data-analytics teams and projects, while marketers should be able to understand metrics and analytics in order to manage efficiently marketing activities.

New technology utilization. Many enterprises conceiving the power of data, have developed technology skills in business intelligence and/or data warehousing, but technologies of big data analytics are different and new. Therefore, enterprises have to utilize techniques and technologies that are available in order to capture value from big data. As these technologies are evolving rapidly, IT departments should be able to develop their capacity and be up to dated to that ongoing innovation. For instance, problems will emerge when database software does not support big data analytics options.

Data privacy. The collection of data is considered to be deeply suspicious by many people. For them, big data is an invasion of their privacy. Marketers are struggling with consumers' perception of data, as the 71% of them believe that brands with access to their personal data are using it unethically, while the 58% of them have not used any digital service due to privacy concerns that lead to decision-making about the applications they download, the email addresses they share and the social media sites to use in order to connect to other websites [37]. Therefore, enterprises need to use safeguards in order to ensure that data are not used to violate the customers' personal privacy [7]. In that direction, data policies including privacy, security, intellectual property and liability issues, should be addressed in order to exploit big data value.

2 Conclusions

The growth of Internet with the beginning of Web 2.0 era enabled companies getting access to big amounts of data easier and cheaper, while the opportunities for external data collection have even increased with the appearance of the Web 3.0.

Enterprises and organizations from all sectors began to focus on data exploitation for gaining competitive advantage.

Nowadays, the big data era has quietly settled down on almost every company, because they realized that data-driven decisions tend to be better and more accurate decisions. However, that many companies in several industries are applying business analytics including big data analytics, it doesn't mean that they all take benefit from it by getting valuable insights and real business value from the available data.

Becoming a data-driven company is more than using analytical techniques and tools. The companies need to hire people equipped with systematic thinking to promote the success in data-driven decision making. Success in the data-oriented business environment today includes being able to think data-analytically. Since the amount of data is continuously growing, domain knowledge and analysis can't be considered as separate areas. Both academic and applied professionals of the companies are expected to have the analytical skills and to understand business processes.

Employees, who don't have the basic understanding of data-analytic thinking, do not really know how the business of an organization is working. If they are able to understand the process and its steps, it will be easier for them to find suitable solutions for the weaknesses of the concerning process step. But to be able to perform data-driven, organizations have to face some challenges, both managerial and technical.

Big data is not just about data volume, but also about variety and velocity. Big data analytics have the ability to help enterprises understanding their business environments, their customers' behavior and needs and their competitors' activities. Thanks to big data analytics enterprises are able to form their products and actions in order to fulfill customers' needs and innovate against rivals through better predictions and smarter decisions on basis of evidence instead of intuition. Organizations that achieve to manage the challenges and adopt a data-driven culture, they can expect good prospects. There is strong evidence that business performance can be improved via data-driven decision making, big data technologies analytical tools and techniques on big data. As more companies learn the essential skills of using big data and how to engage with current technologies, which are continuously developing, may soon stand out from their competitors and have a decisive competitive advantage.

References

1. United Nations: A world that counts. Mobilizing the data revolution for sustainable development. United Nations, New York (2014)
2. OECD: Data-driven innovation big data for growth and well-being: big data for growth and well-being. OECD Publishing (2015)
3. Chen, H., Chiang, R., Storey, V.C.: Business intelligence and analytics: from big data to big impact. Miss. Q. **36**(4), 1165–1188 (2012)

4. Provost, F., Fawcett, T.: Data science and its relationship to big data and data-driven decision making. Big Data **1**(1), 51–59 (2013)
5. Economist, T.: Data is giving rise to a new economy. In: The Economist, 05 Jun 2017. https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy. Accessed 06 Oct 2017
6. Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V.: Critical analysis of big data challenges and analytical methods. J. Bus. Res. **70**, 263–286 (2017)
7. Manyika, J., et al.: Big data: the next frontier for innovation, competition, and productivity (2011)
8. Gantz, J., Reinsel, D.: Extracting Value from Chaos, IDC (2011)
9. Friendly, M.: The golden age of statistical graphics. Stat. Sci. **23**(4), 502–535 (2008)
10. Power, D.J.: Understanding data-driven decision support systems. Inf. Syst. Manag. **25**(2), 149–154 (2008)
11. Cebr: Data equity: unlocking the value of big data Report for SAS, April (2012). https://www.cebr.com/wp-content/uploads/2013/03/1733_Cebr_Value-of-Data-Equity_report.pdf. Accessed 06 Nov 2017
12. Website. https://www.news.microsoft.com/europe/2016/04/20/go-bigger-with-big-data/sm.0008u654e19yueh0qs514ckroeww1/XmqRHQB1Gcmde4yb.97. Accessed 15 Jun 2017
13. McAfee, A., Brynjolfsson, E.: Big data: the management revolution. Harv. Bus. Rev. **90**(10) 60–66, 68, 128 (2012)
14. Burstein, F., Holsapple, C.: Handbook on Decision Support Systems 1: Basic Themes. Springer Science & Business Media (2008)
15. Larson, D., Chang, V.: A review and future direction of agile, business intelligence, analytics and data science—Science Direct. Int. J. Inf. Manage. **36**(5), 700–710 (2016)
16. Davenport, T.: Big Data at Work: Dispelling the Myths. Harvard Business Review Press, Uncovering the Opportunities (2014)
17. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. **35**(2), 137–144 (2015)
18. How to leverage the power of prescriptive analytics to maximize the ROI. In: IBM Big Data and Analytics Hub. http://www.ibmbigdatahub.com/blog/how-leverage-power-prescriptive-analytics-maximize-roi. Accessed 16 Jun 2017
19. Demirkan, H., Delen, D.: Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud. Decis. Support Syst. **55**(1), 412–421 (2013)
20. Lodefalk, M.: Servicification of manufacturing—evidence from Sweden. Int. J. Econom. Bus. Res. **6**(1), 87 (2013)
21. Davenport, T.H., Barth, P., Bean, R.: How 'big data' is different. MIT Sloan Manag. Rev **54**(1), 22–24 (2012)
22. Baesens, B.: Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley (2014)
23. Big Data Analytics for Security—IEEE Xplore Document. http://ieeexplore.ieee.org/abstract/document/6682971/?reload=true. Accessed 18 Jun 2017
24. Wang, G., Gunasekaran, A., Ngai, E.W.T., Papadopoulos, T.: Big data analytics in logistics and supply chain management: certain investigations for research and applications—science direct. http://www.sciencedirect.com/science/article/pii/S0925527316300056?via%3Dihub. Accessed 18 Jun 2017
25. GE's big bet on data and analytics|MIT sloan management review. In: MIT Sloan Management Review. http://sloanreview.mit.edu/case-study/ge-big-bet-on-data-and-analytics/. Accessed 14 Jun 2017
26. Analytics 3.0: Harvard Business Review, 01 Dec 2013. https://hbr.org/2013/12/analytics-30. Accessed 21 Jun 2017
27. Gartner Says 8.4 Billion Connected. http://www.gartner.com/newsroom/id/3598917. Accessed 21 Jun 2017
28. Davenport, T.: Analytics and IT new opportunity for CIOs. In: Harvard Business Review (2016)

29. Double-digit growth forecast for the worldwide big data and business analytics market through 2020 led by banking and manufacturing investments, according to IDC. http://www.idc.com, http://www.idc.com/getdoc.jsp?containerId=prUS41826116. Accessed 21 Jun 2017
30. Brynjolfsson, E., Hitt, L.M., Kim, H.H.: Strength in numbers: how does data-driven decision making affect firm performance?. SSRN Electron. J.
31. If your company isn't good at analytics, it's not ready for AI. In: Harvard Business Review, 07 Jun 2017. https://www.hbr.org/2017/06/if-your-company-isnt-good-at-analytics-its-not-ready-for-ai. Accessed 22 Jun 2017
32. Ryan, L.: The Visual Imperative: Creating a Visual Culture of Data Discovery. Morgan Kaufmann (2016)
33. Lavalle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. **52**(2), 3–22 (2010)
34. The 2 types of data strategies every company needs. In: Harvard Business Review, 01 May 2017. https://hbr.org/2017/05/whats-your-data-strategy. Accessed 18 Jun 2017
35. Waller, M.A., Fawcett, S.E.: Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J. Bus. Logist. **34**(2), 77–84 (2013)
36. Janssen, M., van der Voort, H., Wahyudi, A.: Factors influencing big data decision-making quality. J. Bus. Res. **70**, 338–345 (2017)
37. Chahal, M., et al.: Marketers overestimate consumers' attitude to data—Marketing Week. In: Marketing Week, 23 Jun 2016. https://www.marketingweek.com/2016/06/23/marketers-overestimate-consumers-attitude-to-data/. Accessed 18 Jun 2017

Levering Mobile Cloud Computing for Mobile Big Data Analytics

Yongxin Liu and Houbing Song

Abstract Mobile devices are becoming an indispensable tool for daily life and a considerable number of services are delivered via mobile devices. However, the capacity of mobile devices is constrained for complex interactive and computationally intensive applications (such as Siri on iOS), and therefore, cloud computing is needed to improve user experience. This results in mobile cloud computing. In this chapter, we first review the architectures of popular cloud computing platforms used in enterprise level application scenarios, then we present the requirements and challenges of cloud computing enabled service oriented intelligent mobile applications. After analyzing those challenges on both client side and cloud architecture, we propose the cloud computing architecture for mobile big data analytics and present several application cases.

1 Introduction

In this chapter, we will address several closely related concepts, i.e., cloud computing, mobile cloud computing, and mobile big data. The relationship among these three concepts is given in Fig. 1. As shown in Fig. 1, mobile cloud computing and general cloud computing, are closely related, since both mobile cloud computing and cloud computing are solutions of mobile big data collection, processing, and customized service delivery. Big data provide users with the ability to get insights and discover knowledge from the ocean of data while cloud computing provides the necessary engine.

Y. Liu
School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, Guangdong, China
e-mail: yongxin_liu@foxmail.com

H. Song (✉)
Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
e-mail: h.song@ieee.org

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_2

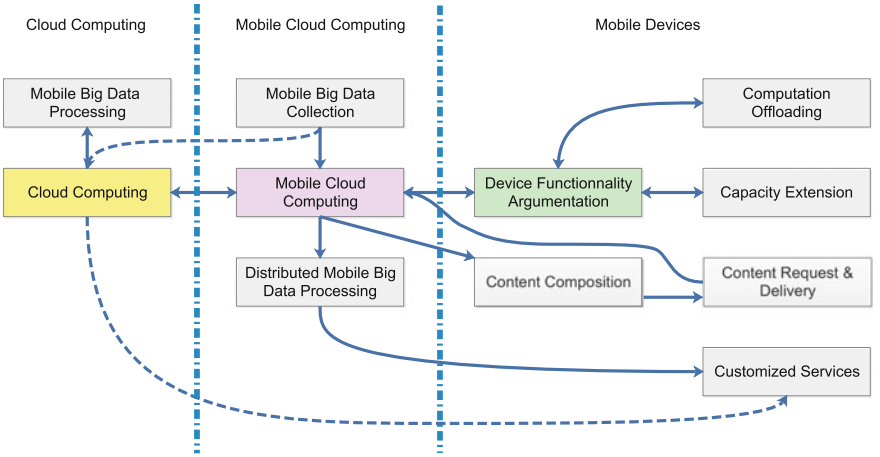


Fig. 1 Relationship among three concepts in this chapter

Mobile big data mining and harvesting is an application paradigm for mobile cloud computing. However, the mobile cloud computing must be specially enhanced from architectures to services for delivering services over resource-constrained mobile devices. With the help of mobile cloud computing, the functionalities of our mobile devices have been significantly enhanced.

The objective of this chapter is to present the opportunities and challenges of applying mobile cloud computing in mobile big data analytics. This chapter is organized as follows: we first provide an overview of general cloud computing application paradigm and its service models along with a logical architecture in Sect. 2. The key technologies of cloud computing, i.e., virtualization and middleware, are presented in Sect. 2.1 and Sect. 2.2, respectively. In Sect. 3, we introduce the challenges (Sect. 3.1) and feasible solutions (Sect. 3.2) for using cloud computing platform to provide service and enhance functionality for mobile devices. The application framework of mobile big data analytics is presented in Sect. 4.

2 An Overview of General Cloud Computing

Cloud computing is an emerging field in information technology that moves computing and data away from desktop and portable PCs into large data centers. Pervasive cloud computing is the antecedent of mobile cloud computing. The word cloud is a metaphor for describing the Web as a space where operating systems [26], applications, storage, data, and processing capacity all have been preinstalled and exist as a service, ready to be shared among users.

The main objective of cloud computing is to make better use of online resources and solve large-scale computation problems [10] (e.g. big data mining). An example

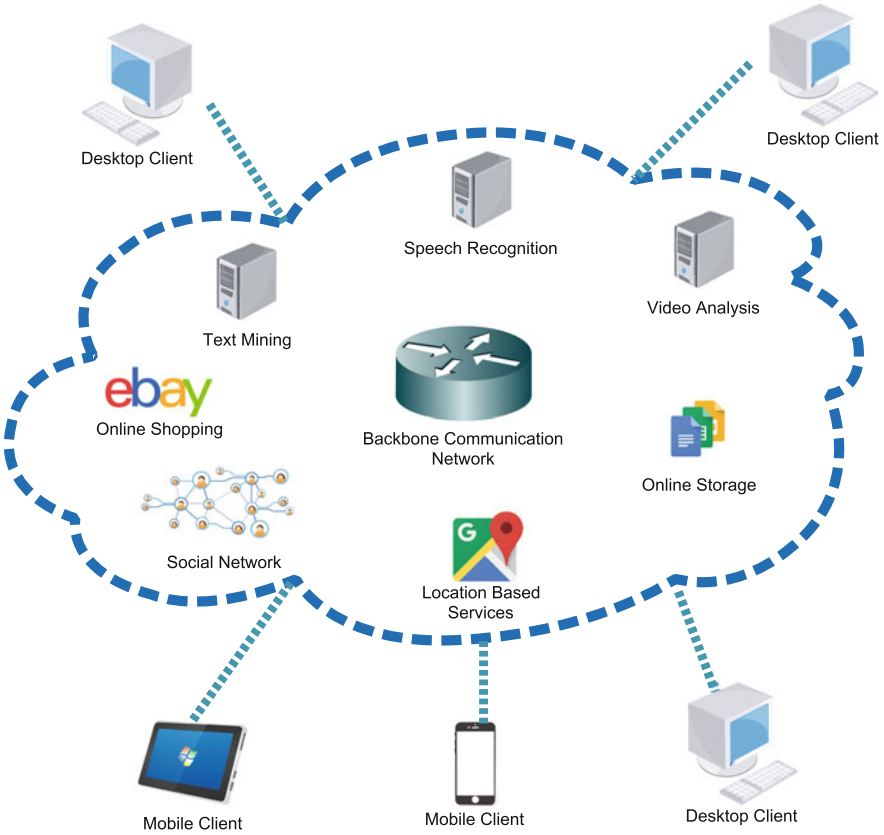


Fig. 2 An overview of cloud computing service

for cloud computing application is illustrated in Fig. 2. Cloud computing enhances the capacity of distributed computers to solve large scale computation problems [26], likewise, resources in the cloud provide transparent resource accessibility to a large number of users who do not need to know their exact locations and specifications [14]. In this scheme, on one hand, cloud applications and data are accessible to authenticated users from anywhere at any time. On the other hand, cloud computing providers offer their "services" in various forms.

According to the National Institute of Standards and Technology (NIST), three standard models are defined, i.e., Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (SaaS) [11, 20, 22]. These cloud service models are explained as follows (their hierarchical relationship is illustrated in Fig. 3).

- **Infrastructure as a Service (IaaS)**. Examples include Flexiscale and Amazon's EC2. The service is usually provided in the form of virtualized PC (also called node) where the consumer is able to deploy and run arbitrary software, which can

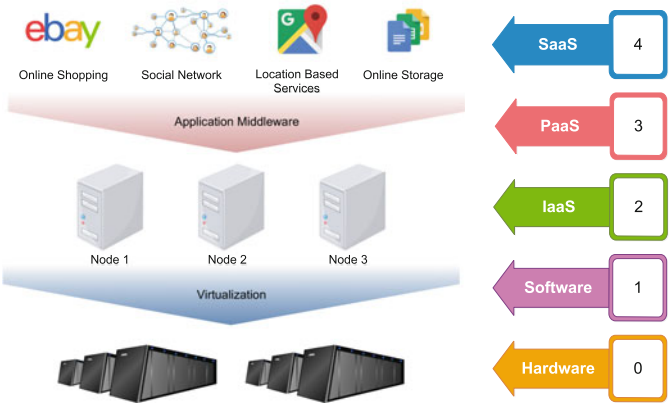


Fig. 3 Architecture of general cloud computing service

include operating systems and applications. Consumers do not manage or control the underlying cloud infrastructure but have full control over the virtualized operating systems, networking components (e.g., host firewalls), storage, and their own applications.

- **Platform as a service (PaaS)**. Examples include Google's Apps Engine, Salesforce.com, Force platform, and Microsoft Azure. In this mode cloud providers deliver a computing platform, which typically includes operating system, programming-language execution environment, database, and web server for end users. In such environments, users develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with hardware and low level operation system APIs.
- **Software as a Service (SaaS)**. Examples include Google Docs, Gmail, Salesforce.com, and Online Payroll. This mode enables users to access providers' applications running on a cloud infrastructure. The applications are accessible from various client devices through either web browser (e.g., web-based email), or a program interface (e.g., navigation service). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, but sends requests and receives responses from the cloud.

In Fig. 3 we can find that there are two supporting layers in cloud computing architecture, i.e., virtualization and application middleware, respectively. A principle of IaaS is that virtualization layer allows one physical server to run several individual computing environments [8]. In practice, it resembles multiple servers for each physical server user. In other words, cloud providers have large data centers which consist of servers to power their cloud offerings, but they cannot devote a single server to

each customer [10, 37]. Likewise, application middleware encapsulates the details of virtual nodes and provides higher level developers a unified interface for deploying elastic services. These two supporting layers are discussed in the following two sections.

2.1 Virtualization Framework

An overview of virtualization architecture for cloud computing is illustrated in Fig. 4. This architecture can be divided into three layers.

- **Computing and networking infrastructures**. In this layer, physical servers are connected with high speed network, at the same time hypervisor operating system (e.g. ESX by VMWare or XenSever by Citrix) is installed to manipulate the hardware and network interface on each physical server node. The specialized operation system runs on "bare metal" with its own kernel and provides components for virtualization [8].
- **Resource pooling and management**. In this layer, resource pooling server aggregates resources from individual physical servers and provides unified management interfaces (e.g. XenCenter by Citrix or vCenter by VMWare) for clients to allocate resources and install operation system for virtual PCs [9]. On the other hand, this resource pooling server collaborates with physical severs to achieve the functionality of individual PCs. In some circumstances, where there is only one physical server, resource pooling sever and management client are combined.
- **Virtual nodes and network**. This layer consist of virtual machines simulated by resource pooling server and virtual networks connecting them. Those virtual machines are real objects providing specific services. It's notable that, network virtualization has played an important role in facilitating the flexibility in topology of these virtual nodes [13]. Thus, cloud providers can provide users with a cluster of nodes along with desired networking structure.

The virtualization of physical infrastructure stimulates and provides a consolidated foundation for IaaS, where users interact with the seamlessly simulated virtual nodes rather than the low level hypervisory systems [24].

For certain purposes, specialized hardware, such as Graphical Processing Unit (GPU) can be installed in physical servers from which they can be specially allocated to virtual nodes enabling the parallel computing capacity.

2.2 Middlewares

In cloud computating, middleware refers to the software framework that connects service resource to the application. In many cases, middleware is a major concept of

PaaS that provides users with an encapsulated environment with unified programming interfaces.

Although cloud service providers and subscribers have developed various service oriented middlewares for their applications, traditional middlewares for data manip-

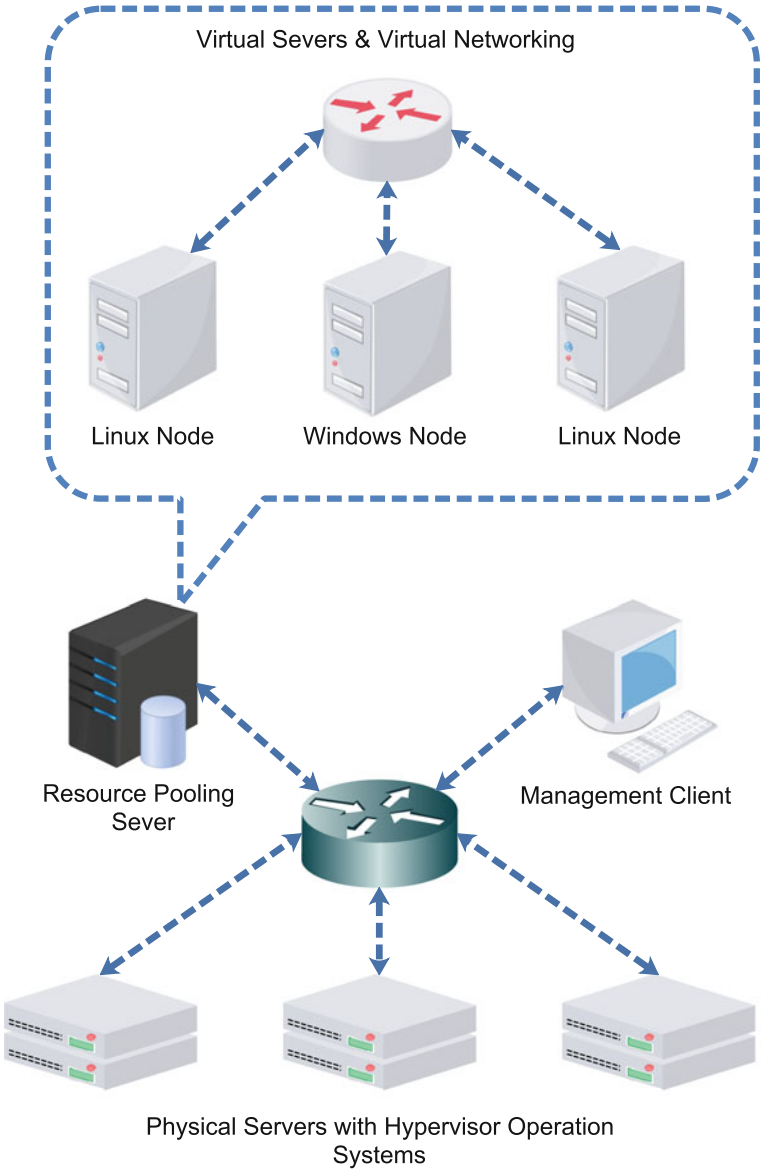


Fig. 4 Architecture of virtualization in cloud computing

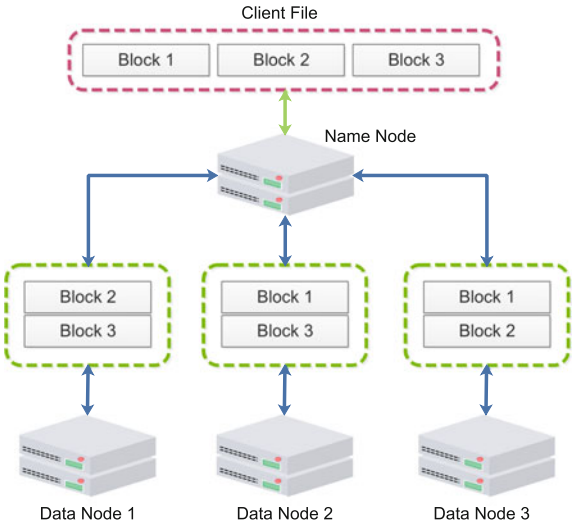


Fig. 5 An overview of mechanism of HDFS

ulation are inadequate for the era of mobile big data. For big data processing, specially for manipulating unstructured data from mobile devices and Internet of Things (IoT), advanced and versatile middlewares are needed to efficiently organize virtualized and even distributed resources (IaaS) to provide subscribers with unified API development and management interfaces. In this manner, the role of middleware for big data processing in cloud resembles a combination of API libraries and resource scheduling engines. In the following subsection, two state-of-art middlewares for cloud based big data storage, management and processing, i.e., Hadoop and Spark, are introduced respectively.

2.2.1 Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models [2]. The core of Apache Hadoop consists of two major abstractions: a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce programming model.

HDFS: The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware [3]. It is similar to existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

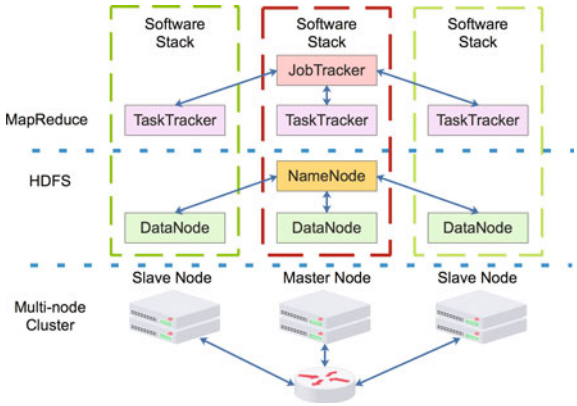


Fig. 6 Logical architecture of MapReduce along with HDFS

A brief architecture of HDFS is given in Fig. 5. HDFS consists of two types of nodes, namely, a NameNode called master and several DataNoes called slaves. HDFS can also include secondary NameNodes. The NameNode manages the hierarchy of file systems and directory namespace (i.e., metadata). File systems are presented in the form of NameNode that registers attributes, such as access time, modification, permission, and disk space quotas. The file content is split into large blocks (in the figure, the client's file is split into 3 blocks), and each block of the file is independently replicated across DataNoes for redundancy. This approach takes advantage of data locality, where nodes, in most of the processing period, manipulate the data within their vicinity. The feature of data locality reduces the dependence on high speed network for information exchange as in conventional parallel computing architectures.

MapReduce: MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster (a group of computers) [6]. Processing can occur on data stored either in a filesystem (unstructured) or in a database (structured). MapReduce can take advantage of the locality of data, processing it near the place it is stored in order to minimize communication overhead.

The logical architecture of MapReduce along with HDFS is given in Fig. 6. MapReduce engine relies on the HDFS, which consists of one JobTracker, to which client applications submit MapReduce jobs. The JobTracker pushes work to available TaskTracker nodes in the cluster, striving to keep the work as close to the data as possible. With a rack-aware file system, the JobTracker knows which node contains the data, and which other machines are nearby. If the work cannot be hosted on the actual node where the data resides, priority is given to nodes in the same rack. This reduces network traffic on the main backbone network. If a TaskTracker fails or times out, that part of the job is rescheduled.

In most cases, using MapReduce, programmers are required to specify two functions only: the map function (mapper) and the reduce function (reducer), respectively

[17]. To illustrate the mechanism of MapReduce, we suppose an application scenario where we want to calculate the occurrence frequency of each word within one million articles. In this scenario, these articles have been merged into text blocks and stored in DataNodes. The MapReduce based application then works in the following manner:

1. **Map**: Each slave node counts the occurence frequency of words within the text block in its memory and generates an intermediate word count table. The master node ensures that only one copy of redundant text blocks is processed.
2. **Shuffle**: Reschedule the work load for merging the intermediate word count table generated by each node, e.g. words starting with letter `a` to `e` may be assigned to node 1, likewise, words starting with `f` to `h` can be assigned to node 2.
3. **Reduce**: Corresponding proportion of intermediate word count table is transmitted and processed to generate the overall statistic of word count. In this scenario, the overall count of words starting with letter `a` to `e` are derived by node 1 whilst node 2 provides the overall count of words starting with letter `f` to `h`. Finally, the master node collects the output results of each slave node and merges into our overall word count table.

Although HDFS and MapReduce are most critical components of Hadoop, several other current open-source Apache projects are related to the Hadoop ecosystem. These components can greatly boost the users to implement certain SaaS applications. It's also notable that all the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework [33, 40].

2.2.2 Spark

Compared with Hadoop, Spark is regarded as a more accessible, powerful and capable big data tool for tackling various big data challenges, because Spark enables applications in Hadoop clusters to run up to 10 times faster either in memory and on disk [27, 31]. Spark runs on top of existing HDFS to provide enhanced and additional functionalities, and therefore it is considered as a powerful complement to Hadoop. The architecture of Apache Spark is based on two main abstractions: Resilient Distributed Datasets (RDD) and Directed Acyclic Graph (DAG).

RDD: Resilient Distributed Datasets (RDDs) are collection of data blocks that are split into read-only partitions and can be stored in-memory on workers nodes (similar to the slave nodes in Hadoop cluster) of the spark cluster. In terms of datasets, apache spark supports two types of RDDs: Hadoop Datasets which are created from the files stored on HDFS and parallelized collections which are based on existing Scala collections. Spark RDDs support two different types of operations: Transformations and Actions. Transformations don't return a single value, since RDDs are immutable. The transformation functions just reads in an RDD and return a new RDD. An Action operation evaluates and returns a new value. When an Action function is called on a

RDD object, all the data processing queries are computed at that time and the result value is returned [36].

DAG: Directed Acyclic Graph (DAG) is a sequence of computations performed on data represented as graph, in which each node is an RDD partition and edge is a transformation on top of data. The DAG abstraction helps eliminate the Hadoop MapReduce multi-stage execution model and provides performance enhancements. In conventional hadoop platforms, when dealing with complicated tasks, developers have to connect together a series of MapReduce jobs and execute them in sequence. Each of those jobs is of high-latency. The job output data between each step has to be stored in the HDFS before other procedures can begin. The feature of DAG as well as the RDD, on one hand, replace the disk IO with in-memory operations and supports in-memory data sharing across DAGs, so that different jobs can work with the same data enabling complex work flows [15].

Spark is highly compatible with the Hadoop cluster. However, the logical definitions of nodes are slightly different although both Hadoop and Spark cluster follow a master-slaver hierarchy. An overview of Spark architecture over HDFS is illustrated in Fig. 7. In Fig. 7, the architecture consists of three type of nodes: master, slave, and resource manager. In small clusters, resource manager and master are combined. HDFS is deployed in the cluster, and the Spark's master node within this cluster can be the NameNode of Hadoop. When a task is submitted to the master node, the following steps are executed:

1. When task driver submits a task, it sends the request to the resource manager.
2. Resource manager checks data locality and finds the best available slave nodes for task scheduling.

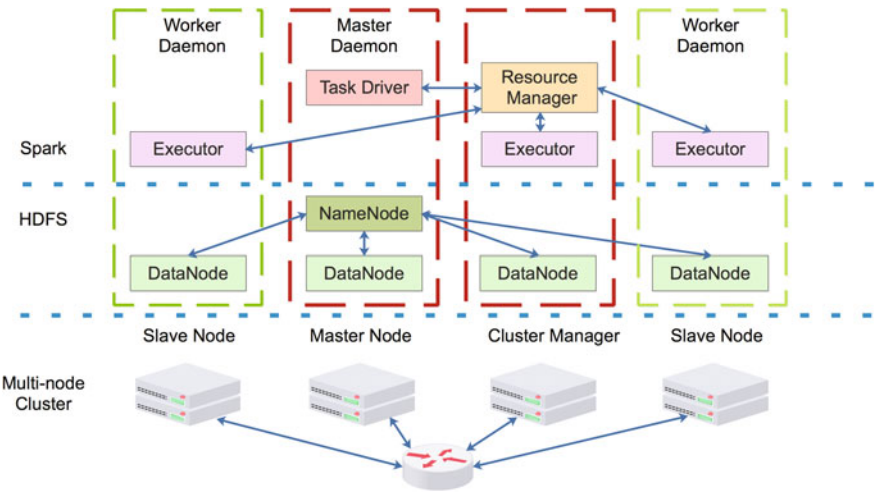


Fig. 7 Logical architecture of spark over HDFS

3. The task is split into different stages based on data locality and resources. Detailed information of the task is sent to slave nodes by the task driver in master node.
4. Task driver keeps track of currently executing task and updates the job monitoring status on master node.
5. Once the task is completed, all the nodes share the aggregated results to the master node.

Nowadays in big data processing applications, specially for mobile big data analytics [1], the fusion of Hadoop and Spark is regarded as an optimal solution.

3 Challenges and Solutions of Mobile Cloud Computing

Mobile devices have become an indispensable tool of daily life. Delivering services and gathering information through mobile devices have also become an inevitable trend in the era of mobile Internet. However, the capacity of mobile devices is constrained where rich media and computationally intensive applications can not be carried out by mobile terminals directly. Hence, as shown in Fig. 1, the cloud computing services with full capacity is considered as a powerful complement to resource-constrained mobile devices to alleviate their burden from heavy load tasks and deliver optimum experience [16, 34]. This scheme has led to the emergence of a Mobile Cloud Computing (MCC). MCC is the combination of cloud computing, mobile computing and wireless networks. The ultimate goal of MCC is to enable the execution of rich mobile applications on capacity limited mobile devices [12].

In this section, we first present the challenges of mobile cloud computing. Then, we introduce solutions for tackling such challenges using enhanced cloud computing infrastructure.

3.1 Challenges

Smartphones have been improved in various aspects such as capability of processor, storage, wireless connectivity and sensory integration. There are still apparent bottlenecks for developing and deploying complicated applications on mobile devices. For instance, 3D and Argumented Reality games may require intense GPU assisted computation which may quickly exhaust the power of batteries. Although MCC is a feasible solution for such resource intensive applications, several challenges exist, resulting in the application development and deployment on mobile devices more complicated than on the desktop cloud clients.

- **Elasticity**: With the increase in the number of mobile users, cloud providers may encounter the phenomenon that in peak period, the amount of service requests may exceed the capacity of their computation resource while in valley period,

their resources are far beyond abundance. This scenario may require automatic scheduling of virtual machines (VMs) to achieve the elasticity of services [28].

- **Wireless connectivity**: Wireless networks are supposed to be bandwidth-saving, less-reliable compared with the wired networks. Establishing and maintaining seamless connectivity between mobile MCC users and clouds in a wireless channel with various ISP and protocols are difficult [19]. For instance, mobile terminals have to re-establish their connections due to their roaming, resulting in disconnections of sessions. Hence, the quality of cloud computing services can be significantly degraded. On the other hand, the over crowded wireless channel or signal interference may even disable the access for cloud computing services.
- **Network latency**: Latency impacts the energy efficiency and user experience of cloud-mobile applications by causing delays. Especially in cellular networks, the capacity of cellular base stations, shadowing effect of buildings, and channel inference can all become possible causes of the latency of MCC connections [18]. On the other hand, there will be bottlenecks if mobile applications have to establish long distance connections to the remote cloud servers through the Wide Area Networks (WAN) where latency is an ineligible aspect. To reduce interaction latency, solutions such as Cloudlet have been proposed.
- **Battery duration**: The purpose of cloud computing is to enable resource-constrained mobile devices to deliver computational density services, however, the frequent service requests from wireless networks and inquiries for response from cloud may in contrast become a major source of power consumption, specially for communication network with low QoS. Therefore, it's important to eliminate the abuse of cloud services, in other words, the mobile devices and cloud should collaborate to provide users with better services.
- **Privacy**: Mobile devices are closely related to users' daily life. Information security and user privacy must be addressed when user related information is processed and transmitted through wireless networks. For instance, from a users' location itineraries, data scientists can easily mine their home and work location or other sensitive information [21, 38].

To wrap up, the major challenges for introducing cloud computing to enhance the functionality of mobile devices are: (a) the latency caused by communication networks or resource inadequacy of cloud providers; (b) Extra power consumption of mobile devices caused either by communication latency or response latency.

3.2 Solutions

In this section, we present the solutions to address the challenges of mobile cloud computing identified in the last subsection and provide an overview of the state-of-art architectures for mobile cloud computing.

- **CloudLet**: In conventional scenarios, mobile devices obtain services from a remote cloud infrastructure. In such process, mobile devices usually access the

wide area network via WiFi access points (APs) or cellular networks where, as discussed in previous subsection, issues such as latency, connectivity may occur, degrading users experience. To mitigate this problem, a series of dedicated high speed networks connected cloud servers are distributed in the vicinity of potential users [30], and therefore, low latency can easily be achieved because there are less packet forwarding in congested backbone networks. However, the deployment of Cloudlet is not so straightforward. Service providers need to balance the cost of subscribing spatially separated server infrastructure, the expense of data traffic between remote servers and the profit they can obtain from mobile users. A possible solution is to use data driven approaches to derive mobile users' locations and preference so as to optimize the distribution of either CloudLet infrastructures and distribution of contents iteratively [5].

- **5G Networks**: In the paradigm of 5G networks, the users' handover is the key component, where users' mobility is provisioned and traffic flow is moved to the next point of attachment [23], i.e. the next base station, with no handover request from the mobile devices. This may boost the development of mobile cloud computing with better connectivity.
- **Dynamic partitioning**: Partitioning technology is introduced to automatically decompose applications to mobile devices and cloud servers so that they can be processed optimally. In the current stage, most partitioning strategies try to distribute applications either to mobile devices and cloud servers without an application controller or a user interface [35, 39]. In this paradigm, the capacity of cloud services is considered to be limited, and therefore, applications may automatically coordinate with remote cloud servers and assess the quality of user experience and decide whether to execute tasks remotely or locally. By using dynamic partitioning, we may reduce the burden of mobile cloud servers by collaborating mobile devices with redundant capacities.
- **Access authorization**: Customized services along with precision decision making rely on highly sensitive data, such as location, contact list or physiological sensory data. This type of data could bring unpredictable harms to mobile users if it is used inappropriately [7]. Therefore, it's necessary for mobile operating systems to block unauthorized access for sensitive data in any case. On the other hand, users should be aware of either data security or misappropriation usage.

Based on these solutions, the architecture of mobile cloud computing is depicted in Fig. 8. The MCC uses Cloudlet to send requests and deliver services to mobile users while the center cloud service is responsible for resource management and service deployment. It's notable that Cloudlets and center cloud infrastructure are connected by dedicated network (e.g. VPN). The architecture of single Cloudlet or center cloud are similar to ordinary cloud computing platforms.

4 Application Framework of Mobile Big Data Analytics

As illustrated in the concept diagram (Fig. 1) of this chapter, mobile big data analytics is an important application scenario of cloud computing. A general framework for big data and mobile big data analytics is illustrated in Fig. 9. By using Internet infrastructure in interconnection layer, mobile big data and other sources of data are sent to the data management layer, where the cloud infrastructure (e.g. HDFS, high capacity servers) are deployed. In this layer, data is divided in two types: the structured data (e.g. data tables) and unstructured data (e.g. sounds, videos). The above layer is the cloud computation layer, where data are analyzed using a series of methods as depicted. Finally, knowledge and patterns from data mining are applied for service support and future strategy optimization.

In the following subsection, several application cases for mobile big data analytics are introduced.

4.1 Case Study: Smart Recommendation

Accurate recommendation is difficult in many aspects due to the lack of preference information of their customers. Recommendation systems can benefit from the integration of mobile big data and context-aware data mining techniques. An example is provided by Sun et al. in [29]. They proposed a case study of IoT and big data analytics for smart tourism and sustainable cultural heritage in the city of Trento, Italy. Their system, called TreSight, integrated wearable sensors, open data, and participatory sensing enhances the services in the area of tourism and cultural heritage with a context-aware recommendation system.

The target users are cities that want to offer innovative services for citizens and visitors in a cost effective way such as cultural heritage, tourism-related companies

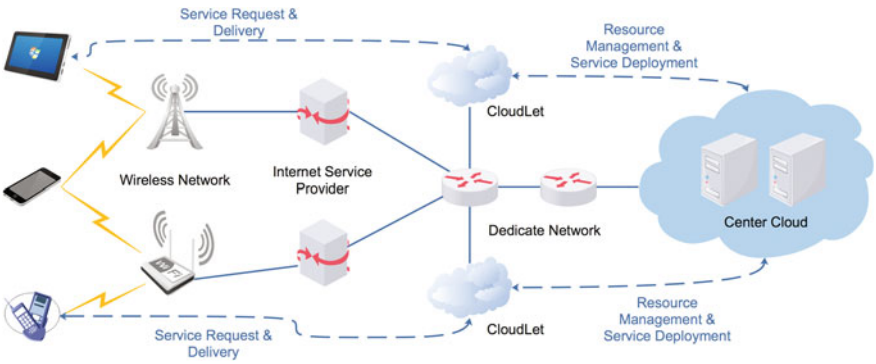


Fig. 8 Logical architecture for mobile cloud computing

that want to promote themselves (hotels, museums, bars, restaurants, etc.) adding their advertisements, promotion codes, coupons etc. in the mobile app that will be offered to the users for the recommendation system.

Their solution first designed a wearable bracelet for each visitor, which is a crowd-sensing device, interacting with the mobile phone, and the Points of Interest, in order to provide the recommendation system the required data to make it context-aware.

The solution requires the deployment of a hotspot for each relevant place that want to be considered a Point of Interest. The hotspot is required to: (a) Gather the data about how many tourists have attended. (b) Update the data repository for a tourist indicating that he/she has visited this place. (c) Collect the sensed data about surrounding humidity, temperature, noise; (d) Provide additional information and content such as the real-time availability, reservations of a restaurant to the visitors through Bluetooth. (e) Finally, a mobile app will be used by visitors to interact with the bracelet, obtain the recommendations, get promotions (discounts, offers, and coupons from the promoted places and sponsors), and obtain more details about the points of interest (pictures, comments, statistics, open hours, current status information such as availability etc.).

Integration of mobile big data strengthen their approach from these aspects: (a) Better understanding the underlying preference from mobile big data. (b) From the cloud level, their approach manipulates nearly every piece of useful information, this

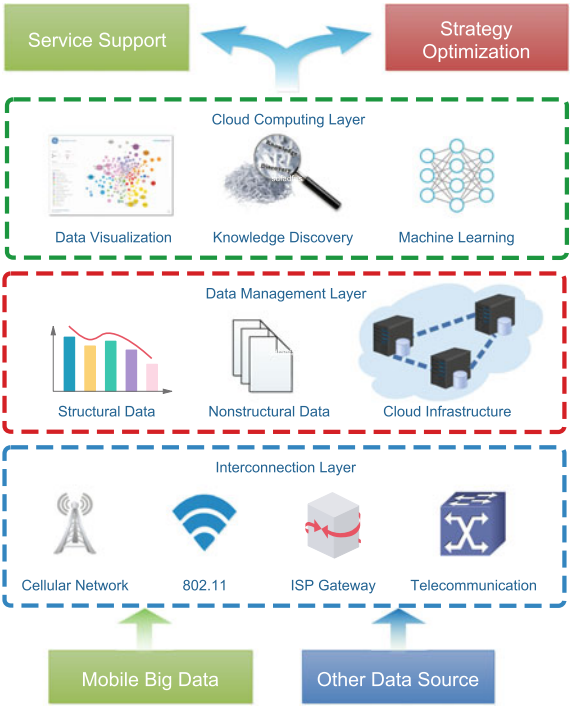


Fig. 9 General framework for pervasive big data analytics

is an essential quality of a good recommender. (c) Successful interaction with visitors via wearable devices provides real-time recommendation and useful information.

4.2 Case Study: Intelligent Healthcare

Mobile devices are integrating more sensors than ever before and continuous information about its owner is collected as time goes by. By leveraging such data, service providers can get a thorough understanding of the user from various aspects [4] which may potentially stimulate better healthcare services. On the other hand, by leveraging the cloud, medical service providers, the potential patients or companies can be connected together. In this way, we may monitor users' health conditions. An example application is provided by Wan et al. in [32] with a framework for a pervasive healthcare system with MCC capability to provide three types of scenarios (home, hospital, or outdoor environment) for ambulatory monitoring, and support a point of care to patients, the elderly, and infants in different environments.

Their system is composed of four main components: WBANs (Wireless Body Area Networks), wired/wireless transmission, cloud services, and users. WBANs collect various vital signals such as body temperature or heart rate information from wearables or implantable sensors. The collected monitoring data are processed in the cloud and then selectively transmitted to the users. The medical video stream from cameras are transmitted to the adjacent routing equipment via wired or wireless transmission, and then to the cloud server via the Internet. Cloud servers possess powerful VM resources such as CPU, memory, and network bandwidth in order to provide all kinds of cloud services such as automatic diagnosis and alarm, geographical information system (GIS) services, location-based services, and medical decision making (MDM). Different users such as hospitals, clinics, researchers, and even patients ubiquitously acquire multiple cloud services by a variety of interfaces such as personal computers, TVs, and mobile phones. This enables the sharing of monitoring data to authorized social networks or medical communities to search for personalized trends and group patterns, offering insights into disease evolution, the rehabilitation process, and the effects of drug therapy.

In their system, patients' profile and medical history data are maintained by the management center of the local private cloud. According to a user's service priority and/or doctor's availability, the doctor may access the user's information as needed. At the same time, automated notifications can be issued to his/her relatives based on this data via various telecommunication means. Besides these basic services, the cloud services also provide GIS deployment, medical data storage, MDM, virtual resource optimization management, and so on. With cloud support, the mobile devices of medical staff will easily exhibit richer mobile video streaming from remote cameras.

4.3 Case Study: Urban Analytics

The ubiquity of mobile big data for urban analytics can be categorized in two ways: first, in microscopic level, mobile trajectory collection on client side provides individuals' coordinates as well as related timestamps, as the gradual accumulation of data, all types of information including individuals' frequency pattern becomes available. From the frequency pattern set, users' mobility pattern can be derived and their behavior in a short range of future time can be predicted; second, in macroscopic level, the aggregation of mobile trajectories from different groups of users results in a dynamic and insightful image of the flow of crowds, from which we are able to assess the occupancy or quality of service of transportation infrastructure while this kind of work in traditional transportation engineering is undertaken with tiny amount of samples from manual survey.

Qiao et al. in [25] introduced a mobility analytical framework for mobile big data, based on real data traffic collected from second-, third- and fourth-generation networks, which covered nearly 7 million people. To construct a user's historical trajectories, they applied different rules to extract users' locations from different data sources and reduce the noise in their data.

They further explore human movement behavior in densely populated areas. They employ a parameter-free method to identify city hotspots from the view of population, apply a modified version of the Apriori algorithm to mine maximal sequential pattern, discover similar users based on their historical trajectories, and predict users' future movements from both temporal and spatial perspectives. These functionalities are of significance for improving the user experience of location-based service (LBS), for optimizing network resources, and for advising city planning.

5 Closing Remarks

Mobile big data analytics has the potential to benefit our society by enabling the move from data to knowledge to action. In this move, mobile cloud computing, which combines cloud computing, mobile computing, and wireless networks, to bring rich computational resources to mobile users, network operators, as well as cloud computing providers, plays an important role. This chapter presents the opportunities and challenges of leveraging mobile cloud computing for mobile big data analytics. We expect that the mobile big data analytics enabled by mobile cloud computing coudl reduce data transfer times, remove potential performance bottlenecks, and increase data security and enhance privacy while enabling advanced applications.

References

1. Alsheikh, M.A., Niyato, D., Lin, S., Tan, H.P., Han, Z.: Mobile big data analytics using deep learning and apache spark. IEEE Netw. **30**(3), 22–29 (2016)
2. Awadallah, A.: Introducing apache hadoop: the modern data operating system. Lecture given at Stanford University. Retrieved Feb 16, **2014** (2011)
3. Borthakur, D., et al.: HDFS architecture guide. Hadoop Apache Project **53** (2008)
4. Chen, M., Zhang, Y., Li, Y., Mao, S., Leung, V.C.: EMC: Emotion-aware mobile cloud computing in 5g. IEEE Netw. **29**(2), 32–38 (2015)
5. Cheng, X., Fang, L., Yang, L., Cui, S.: Mobile big data: the fuel for data-driven wireless. IEEE Internet Things J. (2017)
6. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
7. Dinh, H.T., Lee, C., Niyato, D., Wang, P.: A survey of mobile cloud computing: architecture, applications, and approaches. Wireless Commun. Mobile Comput. **13**(18), 1587–1611 (2013)
8. García-Valls, M., Cucinotta, T., Lu, C.: Challenges in real-time virtualization and predictable cloud computing. J. Syst. Arch. **60**(9), 726–740 (2014)
9. Gulati, A., Holler, A., Ji, M., Shanmuganathan, G., Waldspurger, C., Zhu, X.: Vmware distributed resource management: design, implementation, and lessons learned. VMware Tech. J. **1**(1), 45–64 (2012)
10. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of big data on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
11. Hogan, M., Liu, F., Sokol, A., Tong, J.: Nist cloud computing standards roadmap. NIST Special Publication **35** (2011)
12. Huang, D., Xing, T., Wu, H.: Mobile cloud computing service models: a user-centric approach. IEEE Netw. **27**(5), 6–11 (2013)
13. Jain, R., Paul, S.: Network virtualization and software defined networking for cloud computing: a survey. IEEE Commun. Mag. **51**(11), 24–31 (2013)
14. Jula, A., Sundararajan, E., Othman, Z.: Cloud computing service composition: a systematic literature review. Expert Syst. Appl. **41**(8), 3809–3824 (2014)
15. Karau, H., Konwinski, A., Wendell, P., Zaharia, M.: Learning spark: lightning-fast big data analysis. "O'Reilly Media, Inc." (2015)
16. Kumar, K., Liu, J., Lu, Y.H., Bhargava, B.: A survey of computation offloading for mobile systems. Mobile Netw. Appl. 1–12 (2013)
17. Lämmel, R.: Googles mapreduce programming model revisited. Sci. Comput. Program. **70**(1), 1–30 (2008)
18. Lei, L., Zhong, Z., Zheng, K., Chen, J., Meng, H.: Challenges on wireless heterogeneous networks for mobile cloud computing. IEEE Wireless Commun. **20**(3), 34–44 (2013)
19. Liu, F., Shu, P., Jin, H., Ding, L., Yu, J., Niu, D., Li, B.: Gearing resource-poor mobile devices with powerful clouds: architectures, challenges, and applications. IEEE Wireless Commun. **20**(3), 14–22 (2013)
20. Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., Leaf, D.: Nist cloud computing reference architecture. NIST Spec. Publ. **500**, 292 (2011)
21. Liu, H., Zhou, Y., Zhang, Y.: Estimating users' home and work locations leveraging large-scale crowd-sourced smartphone data. IEEE Commun. Mag. **53**(3), 71–79 (2015)
22. Mell, P., Grance, T., et al.: The nist definition of cloud computing (2011)
23. Morales, A.C., Aijaz, A., Mahmoodi, T.: Taming mobility management functions in 5g: handover functionality as a service (FAAS). In: Globecom Workshops (GC Wkshps), 2015 IEEE, pp. 1–4. IEEE (2015)
24. Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., Zagorodnov, D.: The eucalyptus open-source cloud-computing system. In: Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 124–131. IEEE Computer Society (2009)

25. Qiao, Y., Cheng, Y., Yang, J., Liu, J., Kato, N.: A mobility analytical framework for big mobile data in densely populated area. IEEE Trans.Veh. Technol. **66**(2), 1443–1455 (2017)
26. Sadiku, M.N., Musa, S.M., Momoh, O.D.: Cloud computing: opportunities and challenges. IEEE Potentials **33**(1), 34–36 (2014)
27. Samadi, Y., Zbakh, M., Tadonki, C.: Comparative study between hadoop and spark based on hibench benchmarks. In: 2016 2nd International Conference on, Cloud Computing Technologies and Applications (CloudTech), pp. 267–275. IEEE (2016)
28. Sanaei, Z., Abolfazli, S., Gani, A., Buyya, R.: Heterogeneity in mobile cloud computing: taxonomy and open challenges. IEEE Commun. Surv. Tutor. **16**(1), 369–392 (2014)
29. Sun, Y., Song, H., Jara, A.J., Bie, R.: Internet of things and big data analytics for smart and connected communities. IEEE Access **4**, 766–773 (2016)
30. Verbelen, T., Simoens, P., De Turck, F., Dhoedt, B.: Cloudlets: Bringing the cloud to the mobile user. In: Proceedings of the third ACM workshop on Mobile cloud computing and services, pp. 29–36. ACM (2012)
31. Verma, A., Mansuri, A.H., Jain, N.: Big data management processing with hadoop mapreduce and spark technology: a comparison. In: Symposium on, Colossal Data Analysis and Networking (CDAN), pp. 1–4. IEEE (2016)
32. Wan, J., Zou, C., Ullah, S., Lai, C.F., Zhou, M., Wang, X.: Cloud-enabled wireless body area networks for pervasive healthcare. IEEE Netw. **27**(5), 56–61 (2013)
33. White, T.: Hadoop: the definitive guide. "O'Reilly Media, Inc." (2012)
34. Xu, Y., Mao, S.: A survey of mobile cloud computing for rich media applications. IEEE Wireless Commun. **20**(3), 46–53 (2013)
35. Yang, L., Cao, J., Yuan, Y., Li, T., Han, A., Chan, A.: A framework for partitioning and execution of data stream applications in mobile cloud computing. ACM SIGMETRICS Perform. Eval. Rev. **40**(4), 23–32 (2013)
36. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, pp. 2–2. USENIX Association (2012)
37. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
38. Zhang, X., Yi, Z., Yan, Z., Min, G., Wang, W., Elmokashfi, A., Maharjan, S., Zhang, Y.: Social computing for mobile big data. Computer **49**(9), 86–90 (2016)
39. Zhu, C., Leung, V.C., Hu, X., Shu, L., Yang, L.T.: A review of key issues that concern the feasibility of mobile cloud computing. In: Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, pp. 769–776. IEEE (2013)
40. Zikopoulos, P., Eaton, C., et al.: Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media (2011)

Game Theoretic Approaches in Mobile Cloud Computing Systems for Big Data Applications: A Systematic Literature Review

Georgios Skourletopoulos, Constandinos X. Mavromoustakis, George Mastorakis, Jordi Mongay Batalla, Ciprian Dobre, John N. Sahalos, Rossitza I. Goleva and Nuno M. Garcia

Abstract The constant technological innovations in wireless communications and network technologies as well as the increasing number of smart mobile devices generate an enormous volume of data stemming from a set of user equipments (UEs). Since an exponential growth of data and analytics is witnessed, new technical and application challenges emerge associated with underlying models that exploit cloud computing technologies, such as the Big Data-as-a-Service (BDaaS) or Analytics-as-a-Service (AaaS). In this context, this survey chapter summarizes

G. Skourletopoulos (✉) · C.X. Mavromoustakis
Mobile Systems Laboratory (MoSys Lab), Department of Computer Science, University of Nicosia, Nicosia, Cyprus
e-mail: skourletopoulos.g@unic.ac.cy

C.X. Mavromoustakis
e-mail: mavromoustakis.c@unic.ac.cy

G. Mastorakis
Department of Informatics Engineering, Technological Educational Institute of Crete, Heraklion, Crete, Greece
e-mail: gmastorakis@staff.teicrete.gr

J.M. Batalla
National Institute of Telecommunications and Warsaw University of Technology, Szachowa Str. 1 and Nowowiejska Str. 15/19, Warsaw, Poland
e-mail: jordim@interfree.it

C. Dobre
Faculty of Automatic Control and Computers, Department of Computer Science and Engineering, University Politehnica of Bucharest, Bucharest, Romania
e-mail: ciprian.dobre@cs.pub.ro

J.N. Sahalos
Radio-Communications Laboratory (RCLab), Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: sahalos@auth.gr

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_3

and establishes to what extend existing research studies have progressed towards applying game theoretic approaches in mobile cloud computing systems for big data applications. We identify and critically evaluate the findings of relevant works addressing this research problem by shedding light on contradictions and gaps in the literature. We therefore propose a cost-benefit model formulation in mobile cloud computing environments and a new game theoretic conceptualization, which accounts for the dynamic storage allocation in cloud systems formulated as a benefit optimization problem. Diverse experimental scenarios are adopted to verify and evaluate the optimality and effectiveness of the developed theory in real-world scenarios.

Keywords Game theory · Cloud computing · Mobile computing
Big data · Data analytics · Risk analysis

1 Introduction

The rapid advancements in wireless communications, embedded systems and big data [1, 2] enabled the development of the Internet of Things (IoT) and Internet of Everything (IoE) paradigms [3] as the intersection to connect and interact between the cyber and the physical world [4]. Objects and devices are linked within the physical space and, therefore, location data is a critical component of mobile big data, which can be harnessed to optimize and personalize mobile services [5]. Localization is also divided into outdoor or indoor with respect to the application's use case scenario. Data is gathered from a range of different sources, such as digital sensors, communications, streaming and multimedia applications [6] or even computations [7]. A huge number of user equipments (UEs) and devices produce vast amounts of data in different locations daily related to traffic and flight information, social media or multimedia content, including digital pictures and videos [8]. Hence, the efficient transmission, processing and analysis of big data are essential to extract useful knowledge [9]. In this context, big data analytics constitute a solution concept to offer meaningful information by analyzing big data that can benefit in decision making and problem solving for application domains such as science, engineering or commerce. Big data mining and analytics are considered emerging, interdisciplinary

R. I. Goleva
Faculty of Telecommunications, Department of Communication Networks, Technical University of Sofia, Sofia, Bulgaria
e-mail: rig@tu-sofia.bg

N. M. Garcia
Assisted Living Computing and Telecommunications Laboratory (ALLab), Faculty of Engineering, Department of Computer Science, Instituto de Telecomunicações, University of Beira Interior, Covilhã, Portugal
e-mail: ngarcia@di.ubi.pt

research areas with various applications, such as the intrusion detection or outlier detection in massive data sets, which can prevent from credit card fraud or can be exploited in the medical care and image processing research fields [10]. In addition, data processing systems require more computing power and storage to capture, store and analyze huge and complex data sets [11]. Cloud computing technology minimizes these restrictions providing network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour [12].

On the other hand, game theory has been extensively used to analyze diverse problems in computer science, having many applications in cloud [13] and mobile computing [14, 15]. Mobile cloud computing (MCC) systems are cloud-based systems that can be accessed by end-users through their own mobile devices. Since non-cooperative game theory focuses on predicting each player's actions and payoffs along with analyzing Nash equilibria, this study investigates an optimal strategy for managing costs and benefits in cloud-centric systems and environments by effectively reducing the monetary units to achieve higher payoffs. The proposed approach coupled with a game theoretic perspective strengthens our argument that the modelling can be used to resolve resource allocation problems in real-world applications. Following this introductory section, this tutorial article is a comprehensive, critical review of existing research works in the literature, which investigate game theoretic approaches in cloud and mobile cloud computing systems for big data applications presented in Sect. 2. Section 3 elaborates on a cost-benefit model formulation in mobile cloud computing environments and an original game theoretic analysis of the dynamic storage allocation problem in cloud systems. In Sect. 4, we adopt experimental case scenarios towards the verification and validation of our scheme. Section 5 indicates challenges and issues in this domain and Sect. 6 concludes this survey paper.

2 Exploitation of Mobile Cloud Computing for Big Data Applications: Game Theoretic Perspectives

The proliferation of mass market applications and smart devices drives the explosive growth of wireless sensor data traffic in recent years [16]. Cloud computing and virtualization technologies [17] provide the opportunity for mobile devices to offload computation and execute code remotely in public clouds in order to achieve optimal battery energy saving conditions [18]. Public clouds might use either hypervisors or containers [19]; container technology was introduced within the cloud computing context as a paradigm for cloud-based execution and can achieve near native speeds in processing, memory and network throughput compared to traditional Virtual Machines (VMs), establishing more effective resource management frameworks [20]. In addition, the level of computing capacity introduced between users and the data center-based clouds, the combination of distributed capacity, and the range and complexity of cloud-supported applications run on

end-user devices (coupled with mobility) require new resource allocation, management and scheduling methodologies [21]. In this context, we present previous related research contributions and existing works that study the cloud [22], mobile [23, 24] and mobile cloud technologies coupled with the big data paradigm [25, 26] based on game theory. The goal of this literature review is to discover how the cloud- and mobile-related technologies benefit and improve the use and performance of big data applications and how the application of game theoretic frameworks contributes towards this direction.

In the Big Data era, the users take up the opportunities to act and interact on the internet as their online activities include, among others, web browsing, chatting, online gaming, downloading, uploading or video watching [27]. On the other hand, cloud computing is a highly scalable and cost-effective infrastructure for running high-performance computing, enterprise and web or mobile applications [28]. Cloud-centric big data analytic applications reduce application cost by elastically provisioning resources based on user requirements; efficient scheduling of cloud resources is a key element to guarantee quality of service requirements of budget according to the data analytic requests [29]. Since the cloud-based services become dynamic, the challenges in the resource provisioning domain are crucial in order to ensure that the performance of mobile cloud computing applications is not affected by the time-varying nature of the availability of resources [30]. The survey in [25] points out the strengths that the mobile cloud paradigm brings to the big data field in terms of network partitioning, while scaling out remains still a field for further contribution. Since mobile cloud computing is gaining ground as the ideal environment to run computationally intensive and ubiquitous mobile applications, the resource management issue in cloud computing environments has been attempted to be resolved as a Bayesian Nash equilibrium allocation algorithm in [31], considering different conditions such as heterogeneous distribution of resources, exchange of behavior between the end-users in the cloud or partial dynamic successive allocation. The novelty of this work compared to the literature is that the experimental tests show that the users are able to receive Nash equilibrium allocation solutions by gambling stage by stage even though the player's information is uncertain. From a cloud storage service selection point of view, an automated technique that best matches each dataset of a given application is presented in [32], which is based on a machine-readable description of the capabilities of each storage system along with the user's requirements. Authors in [13] attempt to resolve the uncertainty problem on cloud storage service selection level exploiting fuzzy logic, theory of evidence and game theory. More specifically, the use of fuzzy sets theory for service selection is elaborated in order to express vagueness in subjective preferences, supported by the fuzzy inference and Dempster-Shafer theory of evidence approaches. Another approach in [33] deals with the computation offloading issue in mobile cloud environments using game theory. The problem is formulated as decentralized computation offloading game and a relevant mechanism is proposed to quantify the efficiency ratio over the centralized optimal solution.

The mapping between devices and heterogeneous servers is another critical issue when it comes to energy sustainability challenges [34]. Towards the development

of effective energy optimization mechanisms in mobile cloud computing systems, a congestion game is proposed in [35], which achieves to reduce the overall energy consumption of the mobile systems and the cloud infrastructure. To strengthen this argument, the authors prove that the minimization of the cloud energy consumption will also reduce the cost of mobile cloud users. Furthermore, a nested two-phase game in a mobile cloud computing interaction system is examined in [36]; initially, each mobile device determines the portion of the service requests for remote processing in the cloud in order to minimize the power consumption and the service request response time. In the second phase, a cloud computing controller allocates a portion of the total resources for the service request processing, aiming to maximize the benefits. Another work in [37] elaborates on a coalition game model within a mobile cloud environment where the service providers create a common resource pool to support mobile applications. Additionally, a game model is presented towards a more effective short-term capacity expansion of the resource pool such that the profits of the providers are maximized [38]. Another research work in [39] attempts to resolve the resource provisioning problem across cloud-based networks exploiting game theory. To further elaborate, the authors examine the sophisticated parallel computing problem by requesting the usage of resources and the cost of each computational service. The proposed evolutionary mechanism considers optimization and fairness when investigating the multiplexed strategies of the initial optimal solutions of different participants. Finally, similar research effort is made in [40] where the authors propose a coalition-oriented, uncertainty-focused resource allocation mechanism on cloud service level, comparing the obtained results with existing schemes in the literature and achieving better resource utilization.

3 A Game Theoretic Analysis for Managing Costs and Benefits in Cloud Computing Environments

The cost-effective management of cloud resources motivates the need for advanced allocation strategies with minimum wastage [41]. Cloud computing and data centers provide computing and data storage services and capabilities at large scale ubiquitously, enabling vendors to own data centers for cloud-hosting purposes [42]. However, these solutions are associated with high costs [43, 44], technical debt [45, 46] and environmental impact due to the high-energy consumption at various levels of the computational and data storage processes [47]. Energy consumption is a key issue for the normal operation and maintenance of cloud computing platforms and data centers [48] and, thus, research trends, such as the mobile cloud computing paradigm [2], aim to minimize energy consumption and costs towards greener cloud computing environments [49].

Limited research efforts have been devoted investigating the storage allocation issue in cloud-oriented systems from a cost-benefit viewpoint [50]. Most research works deal with decision procedures for data storage [51], storage allocation challenges in mobile social networks [52] or energy-aware resource allocation for

efficient management of data centers [53]. Since the exploitation of a Nash equilibrium game [54] to resolve the problem is challenging, this section introduces a cost-benefit model to enable the evaluation of different cloud-centric mobile services and a game theoretic conceptualization of the storage and resource allocation issue in cloud-oriented systems from a profit optimization viewpoint. We particularly describe an original scheme for resource allocation in cloud systems based on game theory and we formulate the problem as a cost-benefit game where each player selects the strategy under the necessity of additional resources on storage and computing capacity level. Game theory helps to analyze such systems in real-world scenarios in terms of minimizing the risk for storage upgradation in the long run, avoiding service-level agreement violations and optimizing the predicted payoffs.

3.1 A Cost-Benefit Model Formulation in Mobile Cloud Computing

Since the selection criteria of cloud-centric mobile services might introduce accumulated costs, we initially elaborate on a novel cost-benefit model that predicts benefits and costs on mobile cloud-based service level towards the increase of the return on investment. The mathematical formula is developed to evaluate different cloud-centric mobile services taking into consideration the cost that stems from the

Table 1 Abbreviations and variable/parameter descriptions

Abbreviations	Variable/parameter descriptions
λ	The prediction period of time that is examined
i	The index of the year
U_{max}	The maximum number of end-users
U_{curr}	The initial number of end-users
$\beta\%$	The annual increase in the demand
ppm	The monthly subscription price
$\Delta\%$	The average increase in the monthly subscription price over the period of λ-years
$C_{u/m}$	The monthly service cost for a cloud-supported end-user in the mobile cloud system
$\alpha\%$	The average increase in the document storage cost per month over the period of λ-years
$\gamma\%$	The increase in the data storage cost per month over the period of λ-years
$\mu\%$	The average increase in the maintenance cost per month over the period of λ-years
$\sigma\%$	The increase in the monthly network bandwidth cost over the period of λ-years
$\eta\%$	The average increase in the server cost per month over the period of λ-years
CBA	The cost-benefit measurement result

unused capacity and a linear growth in the number of end-users, which might result in the overutilization of a service and possible service-level agreement violations. We assume that the cloud-supported mobile services are subscription-oriented and there are also charges for servicing the end-users in a mobile cloud system. Since the decrease of the monetary units is of significant importance, two possible types of estimates are encountered: (a) positive results, revealing the underutilization of a cloud-centric mobile service, and (b) negative results, demonstrating the overutilization of a service and possible service-level agreement violations. In this context, the modelling for predicting benefits and costs in mobile cloud computing systems takes the following form [55] (the variables of the formula are explained in Table 1):

$$\begin{aligned} \mathrm{CBA_i} = 12 * \Bigg\{ & \left(1+\frac{\Delta\%}{\lambda}\right)^{\mathrm{i}-1} * \mathrm{ppm} * \left[\mathrm{U_{max}} - (1+\beta\%)^{\mathrm{i}-1} * \mathrm{U_{curr}}\right] \\ & - \left(1+\frac{\alpha\%}{\lambda}+\frac{\gamma\%}{\lambda}+\frac{\mu\%}{\lambda}+\frac{\sigma\%}{\lambda}+\frac{\eta\%}{\lambda}\right)^{\mathrm{i}-1} * \mathrm{C_{u/m}} \\ & * \left[\mathrm{U_{max}} - (1+\beta\%)^{\mathrm{i}-1} * \mathrm{U_{curr}}\right] \Bigg\} \\ = 12 & * \left[\mathrm{U_{max}} - (1+\beta\%)^{\mathrm{i}-1} * \mathrm{U_{curr}}\right] \\ & * \left[\left(1+\frac{\Delta\%}{\lambda}\right)^{\mathrm{i}-1} * \mathrm{ppm} - \left(1+\frac{\alpha\%+\gamma\%+\mu\%+\sigma\%+\eta\%}{\lambda}\right)^{\mathrm{i}-1}\right. \\ & \left. * \mathrm{C_{u/m}}\right], \text{ with } \mathrm{i}=1,\ 2,\ \ldots\lambda \end{aligned} \tag{1}$$

3.2 A Game Theoretic Formulation of the Dynamic Storage Allocation Problem in Cloud Systems

The creation of advanced, automated mechanisms for assigning datasets to storage systems gives the opportunity to meet performance requirements and estimates cost, while customers-companies express their storage needs using high-level concepts. In this context, we propose the cost-benefit modelling perspective in cloud-oriented environments from a big data-as-a-service point of view, considering a set of companies $N = \{1, 2, \ldots, N\}$. The cost (CA) is measured from the data warehouse appliance viewpoint as [56]

$$CA_i = 12 * (C_{s/m} * S_{max}),\ 0 < i \leq l \text{ and } S_{curr} \leq S_{max} \tag{2}$$

where, $C_{s/m}$ is the monthly cost for leasing storage, S_{max} refers to the maximum storage capacity and S_{curr} is the storage currently used. In data warehouse appliances where there are no actual profits ($B = 0$), incremental capacity is added to the storage systems in case of an unpredicted increase in the demand for storage and computing capacity resulting in overhead and downtime. In this direction, two

possible types of benefit calculations are encountered: (a) positive, and (b) negative results. In addition, the cost-benefit modelling from the cloud storage service perspective has been introduced from both non-linear [56] and linear [57] point of views. From a non-linear viewpoint, the cost (CA) and benefits (B) are calculated in year 1 (i.e., Eqs. (3) and (5)) and from year 2 and onwards (i.e., Eqs. (4) and (6)) as [56]

$$CA_1 = 12 * (C_{s/m} * S_{curr}) \tag{3}$$

$$CA_i = 12 * (\Delta_{i-2} * K_{i-2}),\ i \geq 2 \tag{4}$$

$$B_1 = 12 * [C_{s/m} * (S_{max} - S_{curr})] \tag{5}$$

$$B_i = 12 * [\Delta_{i-2} * (S_{max} - K_{i-2})],\ i \geq 2 \tag{6}$$

with,

$$C_{s/m} = C_{s/m_{(curr)}}$$
$$\Delta_0 = (1 + \delta_1\%) * C_{s/m}$$
$$\Delta_i = (1 + \delta_{i+1}\%) * \Delta_{i-1},\ i \geq 1$$
$$\delta_i\% = a_i\% + \gamma_i\% + \eta_i\% + \theta_i\% + \kappa_i\% + \lambda_i\% + \mu_i\% + \sigma_i\%,\ i \geq 1$$
$$K_0 = (1 + \beta_1\%) * S_{curr}$$
$$K_i = (1 + \beta_{i+1}\%) * K_{i-1},\ i \geq 1$$

where, Δ_0 is the formation of the cost for leasing cloud storage in year 2, Δ_i refers to the formation of the cost for leasing storage from year 3 and onwards, $\delta_i\%$ is the total variation in the cost for leasing cloud storage, K_0 indicates the storage used in year 2, K_i is related to the storage used from year 3 and onwards, and $\beta_i\%$ the variation in the demand per year for storage and computing capacity.

We then introduce the cost and benefit measurement models from a linear point of view (i.e., Eqs. (7) and (8)) given as [57]

$$CA_i = 12 * \left[\left(1 + \frac{\Delta\%}{l} \right)^{i-1} * C_{s/m} * (1 + \beta\%)^{i-1} * S_{curr} \right] \tag{7}$$

$$B_i = 12 * \left\{ \left(1 + \frac{\Delta\%}{l} \right)^{i-1} * C_{s/m} * \left[S_{max} - (1 + \beta\%)^{i-1} * S_{curr} \right] \right\} \tag{8}$$

with $0 < i \leq l$, $\Delta\%$ is the variation in the cost for leasing cloud storage, and $\beta\%$ reveals the increase in the demand per year for storage and computing capacity.

In the sequel, we elaborate on a new game theoretic conceptualization towards the benefits optimization in cloud systems by introducing a throttling-oriented storage allocation control mechanism. The proposed scheme enables to allocate

additional resource allocation requests on storage and computing capacity level into a mutually satisfactory condition with respect to the benefit numerical results derived by a storage system. The scheme intends to maximize the profits and guarantees that storage upgradation will not occur either short- or long-term. The charges in cloud-service level are also restructured according to the additional storage capacity and overall resources to be leased off. In this direction, the storage allocation problem is investigated within a benefit prediction period of time and we consider $a_{-n} = (a_1, \ldots, a_{n-1}, a_{n+1}, \ldots, a_N)$ be the storage allocation selection decisions by all other companies-players except new company n. Given the other company's decisions a_{-n}, company n selects a decision $a_n \in \{0, 1\}$ towards the benefit optimization, i.e.,

$$\min_{a_n \in \{0,1\}} B_n(a_n, a_{-n}), \forall n \in N$$

Since the benefit results differ between the storage systems due to the different pool of companies that each one is able to accommodate, the benefit optimization problem is not resolved in the same manner for all storage systems. According to (5), (6) and (8), the benefits function takes the following form with respect to the company n, i.e.,

$$B_n(a_n, a_{-n}) = \begin{cases} B_1, & \text{if} \quad a_n = 0 \\ B_2, & \text{if} \quad a_n = 1 \end{cases} \tag{9}$$

where, B_1 the benefit formula for system 1 once company n is allocated, and B_2 the benefit formula for system 2.

The storage allocation selection problem is now defined as a cost-benefit game $G = \left(N, \{A_n\}_{n \in N}, \{B_n\}_{n \in N}\right)$, where N the set of companies-players, $A_n \triangleq \{0, 1\}$ the set of strategies for the company n and $B_n(a_n, a_{-n})$ the cost-centric benefit function of each new customer-player n. In the sequel, the concept of Nash equilibrium is introduced [54].

Definition 1 *A strategy profile* $a^* = \left(a_1^*, \ldots, a_N^*\right)$ *is a Nash equilibrium of the cost-benefit game if at the equilibrium* a^**, no new player can be allocated to a storage system to further achieve benefit optimization by unilaterally changing its strategy, i.e.,*

$$B_n\left(a_n^*, a_{-n}^*\right) \leq B_n\left(a_n, a_{-n}^*\right), \forall a_n \in A_n, n \in N \tag{10}$$

The Nash equilibrium organizes the increasing cloud storage and computing capacity requests and enables the lease optimization of resources. The game property analyzes the existence of Nash equilibrium in the game, motivating the concept of best response [54].

Definition 2 *Given the strategies* a_{-n} *of the other players, company n's strategy* $a_n^* \in A_n$ *is a best response if*

$$B_n\left(a_n^*, a_{-n}\right) \leq B_n(a_n, a_{-n}),\ \forall\ a_n \in A_n \tag{11}$$

As per (10) and (11), all companies-players play the best response strategies towards each other at the Nash equilibrium, concluding to the following lemma.

Lemma 1 *Given the strategies a_{-n} of the other players in the cost-benefit game, the best response of a new company n is given as the benefit status strategy, i.e.,*

$$a_n^* = \begin{cases} 1, & \text{if} \quad B_i > 0 \\ 0, & \text{if} \quad B_i \leq 0 \end{cases}$$

In this context, Lemma 1 elaborates on the case when the benefit results are greater than zero, which reveals that a storage system has cloud storage and computing capacity left and the company *n* can be added to the pool of current companies-players. The profits are further maximized and the risk of entering in a storage upgradation status in the long run does not occur. However, in case that we get benefit results less than or equal to zero, there is no remaining storage and computing capacity in that storage system and, thus, the additional storage requests will not be satisfied. In this direction, our resource scheduling control mechanism achieves to avoid accumulated costs by allocating the new request to another storage system in the data center where the benefits will be further maximized.

4 Verification and Validation of Theorem

Since the power of sensing devices is drained due to multiple data requests, an effective and secure data access control framework for mobile cloud computing systems is imperative in terms of computation, communication and storage [58]. In this section, we elaborate on a benefit optimization mechanism, which achieves to avoid service-level agreement violations and the risk of entering in a storage upgradation status. Throughout the experimental testing, we prove that the game always admits a pure strategy Nash equilibrium. The results demonstrate the effectiveness of the proposed model and game theoretic approach in real-world operations as the scheme manages to allocate the new requests for resources on cloud storage and computing capacity level in a cost-optimal manner.

Concerning the simulation environment, a quantification tool has been developed as a proof of concept, which quantifies and manages the costs and benefits in cloud systems. The simulation tool achieves to allocate the new requests for additional resources in different storage systems aiming to optimize the benefits and avoid service-level agreement violations. From the technical viewpoint, the web application is targeted to be deployed in the Google Cloud Platform supported by the Google App Engine and it was implemented using the Java programming language.

4.1 Evaluation Analysis of Cost-Benefit Methodology for Cloud-Centric Mobile Services

Towards the evaluation of the proposed model in Sect. 3.1, the experimental results enable to do a risk analysis on whether the linear growth in the number of cloud-supported users will risk the overutilization of a service and the incurrence of accumulated costs in the long run. An indicative usage scenario emphasizes on the need to lease a cloud-centric mobile service and three different services are investigated during a 5-year cost-benefit prediction period. The case scenario examines the cost-benefit flow for these services based on a fifty per cent (50%) annual growth in the number of end-users (see Fig. 1), while Table 2 presents the values that work as inputs to the formula.

Towards the interpretation of the numerical results about this use case scenario, the corporate and premium services are underutilized over the 5-year period and an increase in the return on investment is witnessed due to the constant decrease in the number of monetary units. Despite the fact that the basic service is also underutilized the first four years, the cost-benefit calculations become negative until the end of the prediction period, indicating that this service is overutilized. In this context, service-level agreement violations will occur and the need for abandoning the existing cloud-centric mobile service will be faced in the long run; on this occasion, new accumulated cost will occur hard to be managed. To conclude, the lease of the premium service would be the most cost-effective option for that case scenario, because the numerical results are closer to the minimum positive values and the optimal condition, not to mention that the problem of overutilization does not lurk.

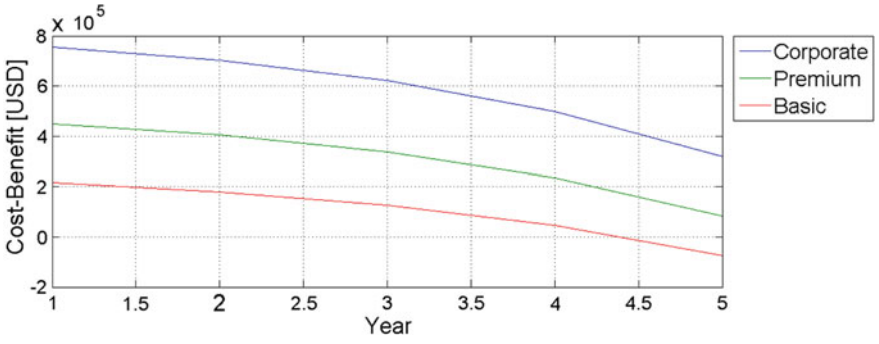


Fig. 1 Case scenario: the cost-benefit flow

Table 2 Inputs to formula (1)

Variable descriptions	Corporate	Premium	Basic
Period of λ-years	$\lambda = 5$	$\lambda = 5$	$\lambda = 5$
Maximum number of active end-users	$U_{max} = 12000$	$U_{max} = 9000$	$U_{max} = 6000$
Initial number of end-users	$U_{curr} = 1500$	$U_{curr} = 1500$	$U_{curr} = 1500$
Monthly subscription price	$ppm = 10$	$ppm = 7$	$ppm = 5$
Increase in the monthly subscription price	$\Delta\% = 2\%$	$\Delta\% = 2\%$	$\Delta\% = 2\%$
Monthly service cost for a cloud-supported user in the mobile cloud system	$C_{u/m} = 4$	$C_{u/m} = 2$	$C_{u/m} = 1$
Increase in the monthly document storage cost	$\alpha\% = 1.5\%$	$\alpha\% = 2\%$	$\alpha\% = 2\%$
Increase in the monthly data storage cost	$\gamma\% = 0.5\%$	$\gamma\% = 1\%$	$\gamma\% = 1.5\%$
Increase in the monthly maintenance cost	$\mu\% = 1\%$	$\mu\% = 2\%$	$\mu\% = 3\%$
Increase in the monthly network bandwidth cost	$\sigma\% = 1\%$	$\sigma\% = 1.5\%$	$\sigma\% = 2.5\%$
Increase in the monthly server cost	$\eta\% = 1\%$	$\eta\% = 1.5\%$	$\eta\% = 3\%$

Table 3 Case scenario

Terms	Variations in the demand
Year 1 to 2	$\beta_1\% = 5\%$
Year 2 to 3	$\beta_2\% = 25\%$
Year 3 to 4	$\beta_3\% = 30\%$

4.2 Experimental Analysis of Game Theoretic Conceptualization for Dynamic Storage Allocation in Cloud-Oriented Systems

Towards the validation of our game theoretic conceptualization in Sect. 3.2, we examine two storage systems in the cloud environment, where non-linear demand for cloud storage and computing capacity is predicted in a 4-year benefit prediction period (see Table 3). The efficient provisioning of cloud resources is a challenging task especially when fluctuations in the resource requirements occur. A storage allocation-inspired throttling controller is targeted to be embedded to regulate and balance the rates at which resource consumption is conducted, either statically or dynamically. We provide improved experimental results compared to our work in [59]. We prove that the proposed game and control mechanism achieves profits and resource management optimization, avoiding the risk for storage upgradation in the long run. Given the Eqs. (5) and (6), we investigate whether a storage upgradation status will occur in either system 1 or 2 for the given period. In this direction, we observe that service-level agreement violations will not occur as far as it concerns system 1 due to the positive calculations, while the profits are also increased (inputs to formulas (5) and (6) are presented in Tables 4 and 5).

Table 4 System characteristics

Variable definitions	Storage system 1	Storage system 2
Maximum storage capacity	$S_{max} = 8$	$S_{max} = 5$
Storage currently used	$S_{curr} = 4$	$S_{curr} = 3$
Monthly cost for leasing cloud storage and computing capacity	$C_{s/m} = 420$	$C_{s/m} = 400$

Table 5 Variations in the cost for leasing resources

Variable definitions	Storage system 1	Storage system 2
Cost variation for leasing additional cloud storage and computing capacity	$\delta_1\% = 2\%$ $\delta_2\% = 8\%$ $\delta_3\% = 9\%$	$\delta_1\% = 3\%$ $\delta_2\% = 10\%$ $\delta_3\% = 12\%$

As far as it concerns system 2, the need for storage upgradation will be faced in the long run, since the increase in the demand in year 4 should have been 26.9%, i.e.,

$$5 - (1 + x) * 3.9375 = 0 \Rightarrow x = 0.26984 \approx 26.9\% \approx 1.06 \; terabytes$$

In this context, the cost-effective, resource scheduling control mechanism achieves to allocate the remaining terabytes in system 1, contributing towards the benefit optimization for this system. Once the mechanism is initiated, the storage and computing capacity currently used for system 1 is approximately

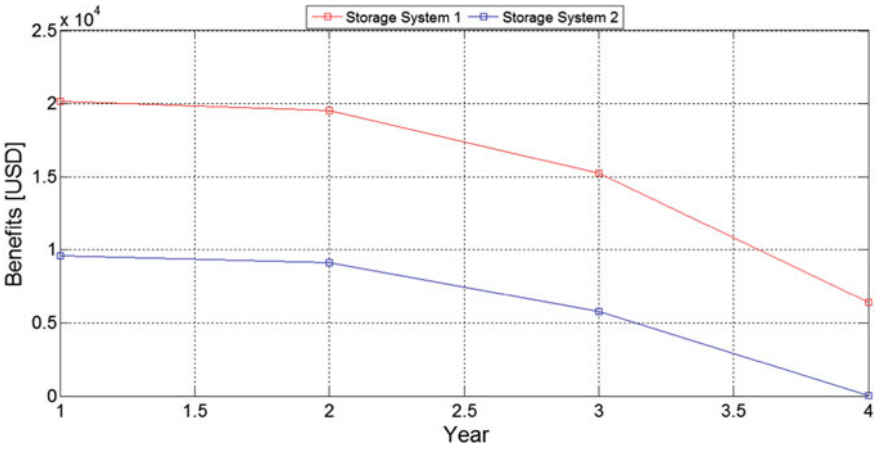


Fig. 2 Case scenario: the benefits flow

Table 6 Benefit results, once the throttling-oriented storage allocation and resource mechanism is motivated

	Year 1	Year 2	Year 3	Year 4
Storage system 1	20,160	19,535.04	15,268.18	6,385.91
Storage system 2	9,600	9,146.4	5,778.3	0.03

$6.825 + 0.12 = 6.945$ TB in year 4 and the increase in the demand is approximately 32.3%, i.e.,

$$(1 + y) \ * \ 5.25 \ = \ 6.945 \Rightarrow y \ = \ 0.32286 \ \approx \ 32.3\%$$

The variation in the total cost for leasing additional resources for system 1 is now 9.1% and the benefit calculation will be approximately 6,385.91 monetary units in year 4. Likewise, the variation in the total cost for the system 2 is 11% and the benefit result is approximately 0.03 monetary units in year 4 (see Fig. 2 and Table 6).

5 Research Challenges and Issues

The problem of cloud storage service selection within a cloud platform or a federation of heterogeneous clouds has become an issue of great importance over recent years [32, 60] as these services are characterized by different features, limitations and prices to meet the customer's requirements in terms of quality of service, quality of experience and costs. Most of the existing works available in the literature do not manage to deal with uncertainty in the sense of subjective preferences from the stakeholders and might result in falsified service selection with respect to the cloud providers, revealing untrustworthy indications concerning the quality of service and pricing [13]. A research gap is also observed when dealing with the complexity of service selection processes for scalability reasons due to multi-objective nature [22].

On the contrary, the features and characteristics that big data era brings coupled with cloud computing, has led to further complexities and challenges associated with the available infrastructures in both hardware and software [61]. Novel software engineering methodologies for developing big data applications are imperative as most related works deal with the effective management, storage, extraction, transformation, processing and analysis issues of large-scale data [62] or the delivery of cloud-based big data analytics solutions [63]. Therefore, it is critical to elaborate on the types of applications, requirements and limitations that big data-oriented software engineering brings, considering both real-time analysis and delay-tolerance, and develop effective models and frameworks for the design of these solutions along with the infrastructure and software architectures that best satisfy these restrictions [64]. Finally, some of the key constraints in the big data

field that attracted the attention of both the research community and the industry are outlined below [28]:

- **Scalability:** NoSQL remains the dominant technology for the deployment of large-scale applications in the cloud due to the lack of cloud computing features to support Relational Database Management Systems (RDBMS) related to enterprise solutions [65].
- **Data Availability and Integrity:** Since mobile end-users require vast amounts of data within short intervals, it is critical for cloud service providers to meet these demands even in case of a security breach. In addition, the correctness of user data remains an issue for cloud-centric applications that require personalized data storage and management [66].
- **Data Quality and Heterogeneity:** The sources and types of data may vary making the data quality quite poor and questionable. The data formats can be inconsistent and not appropriately represented to extract value as a consequence of these heterogeneous sources [67].
- **Data Privacy:** Although there is related work in data privacy [68], there are still limitations about the security and privacy of the data on cloud storage level. The development of new big data analytics solutions requires personalized, location-based information towards more targeted results, which can be potentially exposed to scrutiny, misuse or loss. In this direction, most effective encryption methodologies should be developed [69].

6 Conclusions and Future Work

This tutorial article constitutes a critical review and assessment of the existing research works in the area of game theory applied to cloud and mobile computing from a big data application point of view. Besides our findings in the literature, we also highlight gaps and challenges in this domain. In this direction, we initially elaborate on a cost-benefit model formulation for measuring costs and benefits on mobile cloud-based service level. This viewpoint enables the evaluation of cloud-centric mobile services examining the probability of overutilization. We additionally provide a game theoretic analysis intending to achieve resource allocation optimization in cloud systems by formulating the benefit optimization problem as a cost-benefit game towards the minimization of the risk for storage upgradation either short- or long-term. Our study considers one customer at a time assigning one storage system to satisfy the customer requests. The experimental testing proves the effectiveness of the theorem in real-world operations, achieving to dynamically allocate the new requests for cloud storage and computing capacity in order to obtain higher payoffs. In the future, more complex scenarios with various fluctuations in the demand for resources on storage and computing capacity level or interactions between the players for different systems will be further investigated. A coalition of storage services assigned to new customers will be also

researched along with the selection process when multiple customer service needs have to be taken into account to realize a multi-tenant storage provisioning framework.

Acknowledgements The authors would like to acknowledge networking support by the EU ICT COST Action IC1303 on 'Algorithms, Architectures and Platforms for Enhanced Living Environments (AAPELE)' and the EU ICT COST Action IC1406 on 'High-Performance Modelling and Simulation for Big Data Applications (cHiPSet)'.

References

1. Skourletopoulos, G. et al.: Big data and cloud computing: a survey of the state-of-the-art and research challenges. In: Advances in Mobile Cloud Computing and Big Data in the 5G Era, 1st edn, vol. 22, pp. 23–41. Springer International Publishing AG, Cham, Switzerland (2016)
2. Skourletopoulos, G. et al.: Towards mobile cloud computing in 5G mobile networks: applications, big data services and future opportunities. In: Advances in Mobile Cloud Computing and Big Data in the 5G Era, 1st edn, vol. 22, pp. 43–62. Springer International Publishing AG, Cham, Switzerland (2016)
3. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
4. Weber, R.H., Weber, R.: Internet of Things, vol. 12. Springer, Berlin, Heidelberg (2010)
5. Bello-Orgaz, G., Jung, J.J., Camacho, D.: Social big data: recent achievements and new challenges. Inf. Fusion **28**, 45–59 (2016)
6. Kryftis Y., Mavromoustakis C.X., Batalla J.M., Mastorakis G., Pallis E., Skourletopoulos G.: Resource usage prediction for optimal and balanced provision of multimedia services. In: 2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD 2014), Athens, Greece, pp. 255–259 (2014)
7. Chen, C.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
8. Leskovec J.: Social media analytics: tracking, modeling and predicting the flow of information through networks. In: Proceedings of the 20th international conference companion on World wide web, Hyderabad, India, pp. 277–278 (2011)
9. Emani, C.K., Cullot, N., Nicolle, C.: Understandable big data: a survey. Comput. Sci. Rev. **17**, 70–81 (2015)
10. Hawkins D.M.: Identification of Outliers, vol. 11. Springer (1980)
11. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
12. Mastorakis G., Markakis E., Pallis E., Mavromoustakis C.X., Skourletopoulos G.: Virtual network functions exploitation through a prototype resource management framework. In: 2014 IEEE 6th International Conference on Telecommunications and Multimedia (TEMU 2014), Heraklion, Crete, Greece, pp. 24–28 (2014)
13. Esposito, C., Ficco, M., Palmieri, F., Castiglione, A.: Smart cloud storage service selection based on fuzzy logic, theory of evidence and game theory. IEEE Trans. Comput. **65**(8), 2348–2362 (2015)
14. Charilas, D.E., Panagopoulos, A.D.: A survey on game theory applications in wireless networks. Comput. Netw. **54**(18), 3421–3430 (2010)
15. Han Z.: Game Theory In Wireless And Communication Networks: Theory, Models, And Applications. Cambridge University Press (2012)

16. Pantazis, N.A., Nikolidakis, S.A., Vergados, D.D.: Energy-efficient routing protocols in wireless sensor networks: a survey. IEEE Commun. Surv. Tutor. **15**(2), 551–591 (2013)
17. Batalla J.M., Kantor M., Mavromoustakis C.X., Skourletopoulos G., Mastorakis G.: A novel methodology for efficient throughput evaluation in virtualized routers. In: 2015 IEEE International Conference on Communications (ICC 2015), Communications Software, Services and Multimedia Applications (CSSMA) Symposium, London, UK, pp. 6899–6905. (2015)
18. Kumar, K., Lu, Y.-H.: Cloud computing for mobile users: Can offloading computation save energy? Computer **43**(4), 51–56 (2010)
19. Wood T., Cherkasova L., Ozonat K., Shenoy P.: Profiling and modeling resource usage of virtualized applications. In: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, Leuven, Belgium, pp. 366–387 (2008)
20. Higgins, J., Holmes, V., Venters, C.: Orchestrating Docker Containers in the HPC Environment. Int. Conf. High Perform. Comput. **9137**, 506–513 (2015)
21. Bittencourt, L.F., Diaz-Montes, J., Buyya, R., Rana, O.F., Parashar, M.: Mobility-aware application scheduling in fog computing. IEEE Cloud Comput. **4**(2), 26–35 (2017)
22. Garg, S.K., Versteeg, S., Buyya, R.: A framework for ranking of cloud computing services. Future Gener. Comput. Syst. **29**(4), 1012–1023 (2013)
23. Posnakides D., Mavromoustakis C.X., Skourletopoulos G., Mastorakis G., Pallis E., Batalla J. M.: Performance analysis of a rate-adaptive bandwidth allocation scheme in 5G mobile networks. In: 20th IEEE Symposium on Computers and Communications (ISCC 2015), 2nd IEEE International Workshop on A 5G Wireless Odyssey: 2020, Larnaca, Cyprus, pp. 955–961 (2015)
24. Papadopoulos M., Mavromoustakis C.X., Skourletopoulos G., Mastorakis G., Pallis E.: Performance analysis of reactive routing protocols in mobile ad hoc networks. In: 2014 IEEE 6th International Conference on Telecommunications and Multimedia (TEMU 2014), Heraklion, Crete, Greece, pp. 104–110 (2014)
25. Stergiou C., Psannis K.E.: Recent advances delivered by mobile cloud computing and Internet of Things for big data applications: a survey. Int. J. Netw. Manag. (2016)
26. Mavromoustakis C.X., Mastorakis G., Dobre C. (eds.): Advances in Mobile Cloud Computing and Big Data in the 5G Era, vol. 22, 1st edn. Springer International Publishing AG: Cham, Switzerland (2016)
27. Tan, W., Blake, M.B., Saleh, I., Dustdar, S.: Social-network-sourced big data analytics. IEEE Internet Comput. **17**(5), 62–69 (2013)
28. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of 'big data' on cloud computing: Review and open research issues. Inf. Syst. **47**, 98–115 (2015)
29. Zhao Y., Calheiros R.N., Gange G., Ramamohanarao K., Buyya R.: SLA-based resource scheduling for big data analytics as a service in cloud computing environments. In: 2015 44th International Conference on Parallel Processing (ICPP), Beijing, China, pp. 510–519 (2015)
30. Kosta S., Aucinas A., Hui P., Mortier R., Zhang X.: Thinkair: dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In: 2012 Proceedings IEEE INFOCOM, Orlando, FL, USA, pp. 945–953 (2012)
31. Teng F., Magoulès F.: A new game theoretical resource allocation algorithm for cloud computing. In: 5th International Conference on Grid and Pervasive Computing (GPC 2010), Hualien, Taiwan, pp. 321–330 (2010)
32. Ruiz-Alvarez A., Humphrey M.: An automated approach to cloud storage service selection. In: Proceedings of the 2nd international workshop on Scientific cloud computing, San Jose, California, USA, pp. 39–48 (2011)
33. Chen, X.: Decentralized computation offloading game for mobile cloud computing. IEEE Trans. Parallel Distrib. Syst. **26**(4), 974–983 (2015)
34. Garg, S.K., Yeo, C.S., Anandasivam, A., Buyya, R.: Environment-conscious scheduling of HPC applications on distributed cloud-oriented data centers. J. Parallel Distrib. Comput. **71** (6), 732–749 (2011)

35. Ge Y., Zhang Y., Qiu Q., Lu Y.-H.: A game theoretic resource allocation for overall energy minimization in mobile cloud computing system. In: Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design, Redondo Beach, California, USA, pp. 279–284 (2012)
37. Wang Y., Lin X., Pedram M.: A nested two stage game-based optimization framework in mobile cloud computing system. In: 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), Redwood City, USA, pp. 494–502 (2013)
37. Niyato D., Wang P., Hossain E., Saad W., Han Z.: Game theoretic modeling of cooperation among service providers in mobile cloud computing environments. In: 2012 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, China, pp. 3128–3133 (2012)
38. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Sahalos J.N., Batalla J.M., Dobre C.: A Game theoretic formulation of the technical debt management problem in cloud systems. In: Proceedings of the 14th IEEE International Conference on Telecommunications (ConTEL 2017), 4th International Workshop (Special Session) on Enhanced Living Environments (ELEMENT 2017), Zagreb, Croatia (2017)
39. Wei, G., Vasilakos, A.V., Zheng, Y., Xiong, N.: A game-theoretic method of fair resource allocation for cloud computing services. J. Supercomput. **54**(2), 252–269 (2010)
40. Pillai, P.S., Rao, S.: Resource allocation in cloud computing using the uncertainty principle of game theory. IEEE Syst. J. **10**(2), 637–648 (2016)
41. De Assunção M.D., Di Costanzo A., Buyya R.: Evaluating the cost-benefit of using cloud computing to extend the capacity of clusters. In: Proceedings of the 18th ACM international symposium on High performance distributed computing, Garching, Germany, pp. 141–150 (2009)
42. Lin, M., Wierman, A., Andrew, L.L.H., Thereska, E.: Dynamic right-sizing for power-proportional data centers. IEEEACM Trans. Netw. TON **21**(5), 1378–1391 (2013)
43. Skourletopoulos G., Bahsoon R., Mavromoustakis C.X., Mastorakis G., Pallis E.: Predicting and quantifying the technical debt in cloud software engineering. In: 2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD 2014), Athens, Greece, pp. 36–40 (2014)
44. Skourletopoulos G., Bahsoon R., Mavromoustakis C.X., Mastorakis G.: The technical debt in cloud software engineering: a prediction-based and quantification approach. In: Resource Management of Mobile Cloud Computing Networks and Environments, 1st edn, pp. 24–42. Hershey, Pennsylvania, USA, IGI Global (2015)
45. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Pallis E., Batalla J.M., Kormentzas G.: Quantifying and evaluating the technical debt on mobile cloud-based service level. In: 2016 IEEE International Conference on Communications (ICC 2016), Communications QoS, Reliability and Modelling (CQRM) Symposium, Kuala Lumpur, Malaysia, pp. 1–7 (2016)
46. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Rodrigues J.J.P.C., Chatzimisios P., Batalla J.M.: A fluctuation-based modelling approach to quantification of the technical debt on mobile cloud-based service level. In: 2015 IEEE Global Communications Conference (GLOBECOM 2015), Fourth IEEE International Workshop on Cloud Computing Systems, Networks, and Applications (CCSNA 2015), San Diego, California, USA, pp. 1–6 (2015)
47. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
48. Beloglazov A., Buyya R.: Energy efficient resource management in virtualized cloud data centers. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGRID'10), pp. 826–831 (2010)
49. Sun, X., Ansari, N.: Green Cloudlet Network: A Distributed Green Mobile Cloud Network. IEEE Netw. **31**(1), 64–70 (2017)
50. Yuan, D., et al.: A highly practical approach toward achieving minimum data sets storage cost in the cloud. IEEE Trans. Parallel Distrib. Syst. **24**(6), 1234–1244 (2013)

51. Ruiz-Alvarez A., Humphrey M.: A model and decision procedure for data storage in cloud computing. In: 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 572–579 (2012)
52. Fan, B., Leng, S., Yang, K., Zhang, Y.: Optimal storage allocation on throwboxes in mobile social networks. Comput. Netw. **91**, 90–100 (2015)
53. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. Future Gener. Comput. Syst. **28**(5), 755–768 (2012)
54. Nash, J.: Non-cooperative games. Ann. Math. **54**(2), 286–295 (1951)
55. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Batalla J.M., Sahalos J.N.: An evaluation of cloud-based mobile services with limited capacity: a linear approach. Soft Comput. J., pp. 1–8, Feb 2016
56. Skourletopoulos G.,. Mavromoustakis C.X, Mastorakis G., Pallis E., Chatzimisios P., Batalla J.M.: Towards the evaluation of a big data-as-a-service model: a decision theoretic approach. In: 35th IEEE International Conference on Computer Communications (INFOCOM 2016), First IEEE International Workshop on Big Data Sciences, Technologies, and Applications (BDSTA 2016), San Francisco, California, USA, pp. 877–883 (2016)
57. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Chatzimisios P., Batalla J.M.: A novel methodology for capitalizing on cloud storage through a big data-as-a-service framework. In: 2016 IEEE Global Communications Conference (GLOBECOM 2016), Fifth IEEE International Workshop on Cloud Computing Systems, Networks, and Applications (CCSNA 2016), Washington, D.C., USA, pp. 1–6 (2016)
58. Zhou Z., Huang D.: Efficient and secure data storage operations for mobile cloud computing. In: 2012 8th international conference on Network and service management (cnsm) and 2012 workshop on systems virtualiztion management (svm), pp. 37–45 (2012)
59. Skourletopoulos G., Mavromoustakis C.X., Mastorakis G., Sahalos J.N., Batalla J.M., Dobre C.: Cost-benefit analysis game for efficient storage allocation in cloud-centric internet of things systems: a game theoretic perspective. In: Proceedings of the 15th IFIP/IEEE International Symposium on Integrated Network Management (IFIP/IEEE IM 2017), 2017 First International Workshop on Protocols, Applications and Platforms for Enhanced Living Environments (PAPELE 2017), Lisbon, Portugal, pp. 1149–1154 (2017)
60. Maximilien, E.M., Singh, M.P.: A framework and ontology for dynamic web services selection. IEEE Internet Comput. **8**(5), 84–93 (2004)
61. Mosco V.: To The Cloud: Big Data In A Turbulent World. Routledge (2015)
63. Ji C., Li Y., Qiu W., Awada U., Li K.: Big data processing in cloud computing environments In: 2012 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), San Marcos, TX, USA, pp. 17–23 (2012)
63. Talia D.: Toward cloud-based big-data analytics. IEEE Comput. Sci. 98–101 (2013)
64. Simmhan, Y., et al.: Cloud-based software platform for big data analytics in smart grids. Comput. Sci. Eng. **15**(4), 38–47 (2013)
65. Cattell, R.: Scalable SQL and NoSQL data stores. ACM SIGMOD Rec. **39**(4), 12–27 (2011)
66. Zissis, D., Lekkas, D.: Addressing cloud computing security issues. Future Gener. Comput. Syst. **28**(3), 583–592 (2012)
67. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (1997)
68. Agrawal D., Aggarwal C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Santa Barbara, California, USA, pp. 247–255 (2001)
69. Tene, O., Polonetsky, J.: Privacy in the age of big data: a time for big decisions. Stan. Rev. Online **64**, 63 (2012)

Author Biographies

Georgios Skourletopoulos is currently a Doctoral Researcher in the Mobile Systems Laboratory (MoSys Lab), Department of Computer Science at the University of Nicosia, Cyprus, majoring in cloud, mobile cloud and mobile computing. He also works as a Senior Bid Professional in the EU Institutions Pre-Sales & Solutions Department within the EU Institutions Business Unit at INTRASOFT International S.A., INTRACOM S.A. Holdings Group, Greece. He obtained his M.Sc. in Computer Science from the University of Birmingham, UK in 2013 and his B.Sc. in Commerce and Marketing from the Technological Educational Institute of Crete, Greece in 2012, majoring in e-Commerce and Digital Marketing (including a semester abroad as an Erasmus-Socrates exchange student at the Czech University of Life Sciences Prague, Czech Republic). In the past, he has worked as a Business Analytics and Strategy Consultant in the Business Analytics and Strategy (BA&S) Service Line within IBM's Global Business Services (GBS) Business Unit at IBM Hellas S.A., Greece and he was also a member of the IBM Big Data and Business Analytics Center of Competence in Greece. Mr. Georgios Skourletopoulos has also worked as an e-Banking Platforms Use Case and Quality Assurance Analyst at Scientia Consulting S.A., Greece and as a Junior Research Analyst at Infobank Hellastat S.A., Greece. He is author or co-author of more than fifteen (15) papers published in various international scientific journals, conference and workshop proceedings and book chapters. His research interests lie in the areas of cloud, mobile cloud and mobile computing with a focus on cloud-based software engineering, cloud-based big data and big data-as-a-service (BDaaS), cloud-inspired cost metrics development, technical debt management and risk-cost-benefit analysis in the cloud, game theory in and for the cloud, mobile communications and communication networks.

Constandinos X. Mavromoustakis is currently a Professor in the Department of Computer Science at the University of Nicosia, Cyprus. He received a 5-year Dipl. Eng in Electronic and Computer Engineering from Technical University of Crete, Greece, his M.Sc. in Telecommunications from University College of London, UK and his Ph.D. from the Department of Informatics at Aristotle University of Thessaloniki, Greece. He serves as the Chair of C16 Computer Society chapter of the Cyprus IEEE section, whereas he is the main recipient of various grants including the ESR-EU. His research interests are in the areas of spatial and temporal scheduling, energy-aware self-scheduling and adaptive behaviour in wireless and multimedia systems.

George Mastorakis received his B.Eng. in Electronic Engineering from UMIST, UK in 2000, his M.Sc. in Telecommunications from UCL, UK in 2001 and his Ph.D. in Telecommunications from the University of the Aegean, Greece in 2008. He is serving as an Associate Professor at the Technological Educational Institute of Crete and as a Research Associate in Research & Development of Telecommunications Systems Laboratory at the Centre for Technological Research of Crete, Greece. His research interests include cognitive radio networks, networking traffic analysis, radio resource management and energy efficient networks. He has more than 80 publications at various international conferences proceedings, workshops, scientific journals and book chapters.

Jordi Mongay Batalla received his M.Sc. degree from Universitat Politecnica de Valencia in 2000 and his Ph.D. degree from Warsaw University of Technology in 2009, where he still works as Assistant Professor. In the past, he has worked in Telcordia Poland (Ericsson R&D Co.) and later in the National Institute of Telecommunications, Warsaw, where he is the Head of Internet Architectures and Applications Department from 2010. He took part (coordination and/or participation) in more than 10 national and international ICT research projects, four of them inside the EU ICT Framework Programmes. His research interests focus mainly on Quality of Service (Diffserv, NGN) in both IPv4 and IPv6 infrastructures, Future Internet architectures (Content

Aware Networks, Information Centric Networks) as well as applications for Future Internet (Internet of Things, Smart Cities, IPTV). He is author or co-author of more than 100 papers published in books, international and national journals and conference proceedings.

Ciprian Dobre completed his Ph.D. at the Computer Science Department, University Politehnica of Bucharest, Romania, where he is currently working as a full-time Professor. His main research interests are in the areas of modeling and simulation, monitoring and control of large scale distributed systems, vehicular ad hoc networks, context-aware mobile wireless applications and pervasive services. He has participated as a team member in more than 10 national projects the last four years and he was member of the project teams for five international projects. He is currently involved in various organizing committees or committees for scientific conferences. He has developed MONARC 2, a simulator for LSDS used to evaluate the computational models of the LHC experiments at CERN and other complex experiments. He collaborated with California Institute of Technology in developing MonALISA, a monitoring framework used in production in more than 300 Grids and network infrastructures around the world. He is the developer of LISA, a lightweight monitoring framework that is used for the controlling part in projects like EVO or the establishment of worldwide records for data transferring at SuperComputing Bandwidth Challenge events (from 2006 to 2010). His research activities were awarded with the Innovations in Networking Award for Experimental Applications in 2008 by the Corporation for Education Network Initiatives (CENIC) and a Ph.D. scholarship by Oracle between 2006 and 2008.

John N. Sahalos received his B.Sc. degree in Physics, in 1967 and his Ph.D. degree in Physics, in 1974, from the Aristotle University of Thessaloniki (AUTh), Thessaloniki, Greece. Except of his Ph.D., during 1970-75, he studied in the School of Engineering of AUTh and he received his Diploma (BCE + MCE) in civil engineering. During 1972-74, he also studied in the Electronic Physics Department of AUTh and he received his professional Diploma of postgraduate studies in Radio-Electrology (1975). Prof. Sahalos is the director of the Radio-Communications Laboratory (RCL) since 1986. From 1971 to 1974, he was a Teaching Assistant of Physics and from 1974 to 1976, he was an Instructor at AUTh. During 1976, he worked at the ElectroScience Laboratory, the Ohio State University, Columbus, as a Postdoctoral Fellow. From 1977 to 1985, he was a Professor in the Electrical Engineering Department, University of Thrace, Greece, and Director of the Microwaves Laboratory. During 1982, he was a visiting Professor at the Department of Electrical and Computer Engineering, University of Colorado, Boulder. From 1985 to 2010, he was a Professor at the School of Sciences, AUTh, while since 2010 he is a Professor Emeritus at the same School. From 1989 to 2010, he was the director of the Postgraduate Studies in Electronic Physics. During 1989, he was a visiting Professor at the Technical University of Madrid, Spain. Since 2010, he is a Professor at the Department of Electrical & Computer Engineering, University of Nicosia (UNic), Cyprus. Prof. Sahalos is an IEEE Life Fellow, a member of the Academy of Science of New York and he has been honoured with a special investigation fellowship of the Ministerio de Education Y Siencia, Spain. He is a member of eight IEEE societies and Associate Editor in three international journals. Since 1992, he has been a member of commissions A and E of URSI and, since 1999, he is the president of the Greek committees of URSI. He is also a member of the Greek Physical Society and the Technical Chamber of Greece. He is honorary member of the Greek Society of Electronic Physics. From 2002 to 2004, he was in the Board of Directors of OTE and, from 2005 to 2008, he was a member of the National Committee of Research and Technology and the chairman of the section of Informatics, Telecommunications and Systems of the same Committee. From 2007 to 2010, he was the Vice-President of the Research Committee and Associate President of the Committee of Communications & Networks of the Aristotle University of Thessaloniki. Since 2010, he is a Member of the International Consulting Committee of the GRNET S.A. He is author of four books, seven book chapters and more than 350 articles published in the scientific literature. His research interests are in the areas of Antennas Analysis & Design, Electromagnetic Fields Measurements, Electromagnetic Diffraction,

Numerical Methods (FEM, FDTD, MoM), Microwave Engineering, Radio-Communications, Sensor Networks & RFID Systems, EMC and Biomedical Engineering.

Rossitza I. Goleva received her Ph.D. in Communication Networks in 2016 and M.Sc. in Computer Science in 1982 at Technical University of Sofia, Bulgaria. She was part of the research staff of the research Institute of Bulgarian PTT between 1982 and 1987. Since 1987, she is with the Department of Communication Networks at Technical University of Sofia. At present, she works on communication networks, communication protocols, and software engineering. Her research interests are in Quality of Service in communication networks, communication protocols, traffic engineering, cloud and fog computing, and performance analyses. She is an IEEE Member, involved in IEEE Bulgaria section activities, has more than 90 research publications, was part of more than 30 research projects including the EU ICT COST Action IC1303 AAPELE and the Advanced Systems for Prevention and Early Detection of Forest Fires 2016/PREV/03 (ASPIRES).

Nuno M. Garcia holds a Ph.D. in Computer Science Engineering from the University of Beira Interior (UBI), Covilhã, Portugal (2008) and he has a 5-year B.Sc. in Mathematics/Informatics also from UBI (1999-2004). He is Assistant Professor at the Faculty of Engineering, Computer Science Department at UBI and Invited Associate Professor at the School of Communication, Architecture, Arts and Information Technologies of the Universidade Lusófona de Humanidades e Tecnologias in Lisbon, Portugal. He was founder and is coordinator of the Assisted Living Computing and Telecommunications Laboratory (ALLab), a research group within the Instituto de Telecomunicações at UBI. He was also co-founder and is coordinator of the Executive Council of the BSAFE LAB – Law enforcement, Justice and Public Safety Research and Technology Transfer Laboratory, a multidisciplinary research laboratory in UBI. He is the coordinator of the Cisco Academy at UBI, Head of EyeSeeLab in EyeSee Lda., Lisbon, Portugal, and member of the Consultative Council of Favvus IT HR SA, Lisbon. He is Chair of the ICT COST Action IC1303 AAPELE – Architectures, Algorithms and Platforms for Enhanced Living Environments (Brussels, Belgium). He is the main author of several international, European and Portuguese patents and he is member of the Non-Commercial Users Constituency, a group within GNSO in ICANN. He is also member of ACM SIGBio, ISOC and IEEE. His main interests include Next-Generation Networks, algorithms for bio-signal processing, distributed and cooperative protocols.

Part II
Architectures, Applications and Services for Mobile Big Data

Evidence-Aware Mobile Cloud Architectures

Huber Flores, Vassilis Kostakos, Sasu Tarkoma, Pan Hui and Yong Li

Abstract The potential of mobile offloading has contributed towards the flurry of recent research activity known as mobile cloud computing. By instrumenting the mobile applications with offloading mechanisms, a mobile device can save its energy and increase its performance. However, existing offloading mechanisms lack from efficient decision models for augmenting the mobile device with cloud resources on the fly. This problem is caused by the large amount of system's parameters and their scattered values that need to be considered and characterized merely by the device depending on its contextual needs. Thus, the offloading process still suffers from deficiencies that do not allow a device to maximize the advantages of going cloud-aware. In this chapter, we explore the challenges and opportunities of a new kind of mobile architecture, namely *evidence-aware mobile cloud architecture*, which relies on crowdsensing to diagnose the optimal configuration for migrating mobile functionality to cloud. The key insight is that by using the massive parallel infrastructure of the cloud to process big data, it is possible to collect offloading evidence from large amount of devices that is later analyzed in conjunction to infer an efficient configuration to execute a smartphone app for a particular device.

H. Flores (✉) · V. Kostakos
Center of Ubiquitious Computing, University of Oulu, Oulu, Finland
e-mail: huber.flores@oulu.fi

V. Kostakos
e-mail: vassilis.kostakos@oulu.fi

S. Tarkoma
Department of Computer Science, University of Helsinki, Helsinki, Finland
e-mail: sasu.tarkoma@helsinki.fi

P. Hui
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Kowloon, Hong Kong
e-mail: panhui@cse.ust.hk

Y. Li
Department of Electronic Engineering, Tsinghua University, Beijing, China
e-mail: liyong07@tsinghua.edu.cn

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_4

1 Introduction

Mobile and cloud computing are two of the biggest forces in computer science [1]. Nowadays, a user relies either on the mobile or the cloud to perform most of the software aid activities, e.g., e-mail, video streaming, image editing, document editing, web browsing, payment, messaging, games, among many others. While the cloud provides to the user the ubiquitous computational and storage platform to process any complex task, the smartphone grants to the user the mobility features to process simple tasks, anytime and anywhere. Therefore, it is logical that the convergence of these two domains into Mobile Cloud Computing (MCC) will lead to the next generation of mobile applications [2, 3].

Generally, mobile devices are able to consume cloud services through specialized Web APIs in a service-oriented manner [4, 5], e.g., REST. A back-end server located in the cloud is a common component of a mobile application, e.g., push notification, Web service, etc. In fact, since the cloud grants dynamic features to the back-end of a mobile architecture, e.g., scalability on the fly, new paradigms such as MBaaS (Mobile Back-end as a Service) are on the rise [1, 6]. Thus, a logical question to answer is *how the cloud can assist the smartphone in creating the post-pc era?*

Since the mobility of the smartphones imposes many limitations in the mobile resources, e.g., processing, storage and energy, among others, several work proposes to offload opportunistically computational tasks from the mobile device to the cloud [7–17]. Offloading is a technique that allows a low power device, e.g., smartphones, to outsource the processing of a task, e.g., code, service, job, etc., to a higher capabilities machine [17–19], e.g., cloud. The potential of the approach for improving the performance and extending the batterylife is widely accepted and proven feasible with latest mobile technologies. However, the technique still suffers from deficiencies caused from the large amount of parameters that need to be configured correctly to optimize the binding between mobile and cloud resources [20], for instance, since last generation smartphones are as powerful as some cloud servers, it is reasonable to bind those devices with even higher capabilities machines, such that the performance is increased instead of decreased. While the offloading gains are improved even further when optimizing the configuration in which a mobile device migrates the tasks to the cloud, it is not a trivial task to find the optimal configuration to offload, mainly because the large amount of possibilities available.

To counter the deficiencies in the offloading process, we explore a new kind of mobile architecture that relies on *crowdsensing* in order to ***diagnose*** the optimal configuration to migrate tasks for a particular device. The key insight is that traces (aka evidence) from the offloading process are collected from the huge amount of devices that outsource tasks to the cloud (community). By using the massive parallel infrastructure to process *big data* [22–24], the cloud analyzes the evidence to infer the optimal configuration and injects it into each device. Naturally, since the approach relies on data, the improvements are incremental and adaptive based on the amount of data collected. Thus, this type of architecture is defined as an *Evidence-aware Mobile Cloud Architecture* (EMCA).

In this chapter, we start by providing a literature review about how the cloud can assist the smartphone to overcome the limitations imposed by mobility. We then make a comparison of existing solutions and we highlight the differences with our proposed EMCA. Next, we explore the challenges, technical problems and opportunities of an EMCA. Lastly, we discuss about the benefits and drawbacks of the architecture along with our future directions.

2 Mobile Cloud Offloading

Mobile cloud offloading (aka computational offloading, cyber-foraging) has been re-discovered as a technique to empower the computational capabilities of mobile devices with elastic cloud resources. Computational offloading refers to a technique, in which a computational operation is extracted from a local execution workflow, later, that operation is transported to a remote surrogate for being processed externally, and lastly, the result of that processing is synchronized back into the local workflow [17]. Computational offloading has evolved considerably from cloudlets to code offloading.

2.1 Cloudlets

Cloudlets [25] is one of the initial work that propose the augmentation of mobile computational resources with nearby servers in proximity, e.g. hot spots. Cloudlets overcome the problem of connecting to high latency remote servers by bridging the cloud infrastructure closer to the mobile user. The motivation of reducing the latency between mobile device and cloud is to enrich the functionality of the mobile applications without degrading its perception and interaction in environments where network communication changes abruptly. Figure 1 shows a basic cloudlet architecture. The architecture consists of two parts, a client and a server located in proximity, which means that there is no network hopping between the device and the server. A nearby server is managed by a service provider using virtual machines. A virtual machine is migrated from the cloud of the service provider to the nearby server, so that cloud service provisioning (create, launch or delete) for the mobile can occur from the nearby server. Alternatively, the service provider also can migrate a service to other types of infrastructure, e.g., base stations, in order to reduce the communication latency with the device [18].

While a cloudlet overcomes the problems that arise from high communication latency, the deployment of a cloudlet is a complex task, as involves to introduce specialized components or modify existent ones at low level of granularity, e.g., hardware. Thus, its adaptation is neither flexible nor scalable. As a result, many other solutions have been proposed [19]. The goal of these solutions is to optimize the delegation of computational tasks by relying on higher manipulation of the source

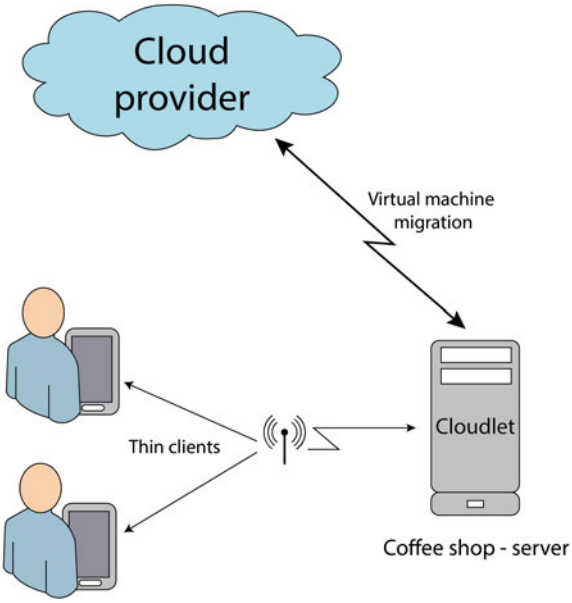


Fig. 1 Components and functionality of a cloudlet system

code of the applications. In this process, computational tasks are delegated to powerful machines at code level as explained in subsection 2.2. Notice that a cloudlet also can be equipped with strategies to offload code [18]. However, the only advantage of using code offloading techniques in a cloudlet model is that the processing of a task can be splitted to multiple devices in a fine-grained fashion, e.g., Method.

2.2 Mobile Code Offloading

Code offloading leverages the small amount of data transferred and the opportunistic high speed connectivity to cloud infrastructure for augmenting the capabilities of the mobile devices [3]. The potential of technique lies in the ability for making the battery life of the smartphones last longer and shortening the response time of mobile applications. Mobile applications are instrumented with code offloading mechanisms for moving a computational task at code level from one place to another. The decision whether to move or not the task from the device for harnessing dedicated external infrastructure is done in the device by analyzing the multiple parameters that can influence the decision to be beneficial or not for the device [10]. The evaluation of the code requires to consider different aspects, for instance, *what code to offload*, e.g., method name; *when to offload*, e.g., RTT (Round Trip Times) thresholds; *where to offload*, e.g. type of cloud server; *how to offload*, e.g. split code into *n* processes, etc.

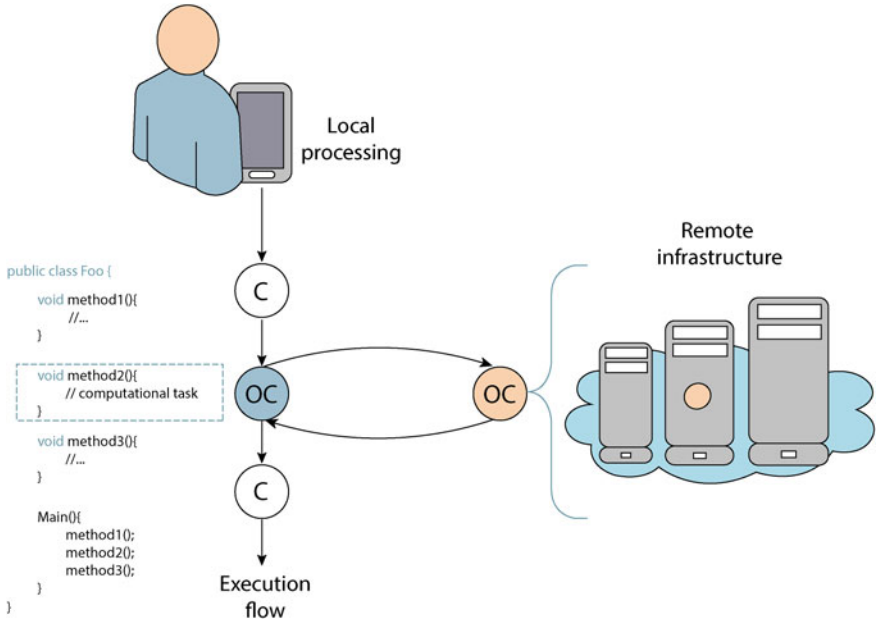


Fig. 2 A code offloading architecture: components and functionalities

Most of the proposals in the field do not cover all these aspects, and thus we describe a basic offloading architecture, which is shown in Fig. 2. The architecture consists of two parts: a client and a server. The client is composed of a code profiler, system profilers and a decision engine. The server contains the surrogate platform to invoke and execute code. Each component is described in detail as follows:

1. ***Code profiler*** is in charge of determining *what to offload*. The profiler characterizes the effort required for the device to execute a portion (*C*) of code—*Method, Thread or Class*. This includes the time of execution and amount of energy required. Based on this characterization, the profiler identifies the code (*OC*) that is candidate to offload. Code can be profiled at different development stages of a mobile application. Thus, we define two types of profilers, manual and automated. Manual profilers are the developers, who select explicitly the portions of code that can be offloaded, e.g., with a code annotation (application is not installed in the device). Automated profilers are runtime processes that analyze the code during runtime using different approaches, e.g., static analysis, history data, etc., and determine the portions of code that are intensive or not for the device (application is already installed in the device).
2. ***System profilers*** are responsible for monitoring, sampling and characterizing multiple parameters of the smartphone during runtime, such as available bandwidth, data size to transmit, energy required to execute the code, surrogate computational capabilities, etc. These parameters are utilized to quantify whether

offloading or not *OC* introduces energy or performance gains for the mobile device.

3. ***Decision engine*** is a reasoner that infers *when to offload* to cloud. The engine retrieves the characterized data obtained by the profilers, and applies certain logic over them, e.g. linear programming, fuzzy logic, markov chains, etc., so that the engine can measure whether the handset obtains or not a concrete benefit from offloading to cloud. The amount of parameters considered in the decision process define the opportunistic context in which a mobile task is offloaded. This suggests that based on the combination of multiple parameters, it is possible to obtain different gains in performance and energy [20]. Proof of this is the way in which existing frameworks characterize with different amount of parameters, the opportunistic moments in which a device offloads to cloud.
4. ***Surrogate platform*** is the computational service located in the proximity of the device or in the cloud, which contains the environment to execute the intermediate code sent by the mobile, e.g. Android-x86, .Net, etc. The computational capabilities of the server are important in an offloading architecture as determine the level in which the task is accelerated [21]. This information is critical to adjust the response time of applications based on the type of device. Ideally, a mobile application must accelerate its execution when offloading rather than slowing down performance.

3 Mobile Offloading Frameworks

In this section, we provide a literature review of frameworks to offload to cloud. Table 1 describes most relevant proposals in code offloading. The table compares the key features of the offloading architectures, namely the main goal, how code is profiled, the adaptation context, the characterization of the offloading process, and how code offloading is exploited from mobile and cloud perspectives. From the table, the main goal defines what is the actual benefit for using the associated framework. The mechanism used to profile code provides information about the flexibility and integrability of the system. The adaptation context specifies the considerations taken by the system to offload. The characterization means whether the offloading system has a priori knowledge or not about the effects of code offloading for the components of the system. Finally, the exploitation highlights the mobile benefits obtained from going cloud-aware, and the features of the cloud that are leveraged to achieve those benefits. Moreover, we can also observe that currently, most of the effort has been focused on providing the device with an offloading logic based on its local context.

MAUI [11] proposes a strategy based on code annotations to determine which methods from a Class must be offloaded. An annotation is a form of metadata that can be agregated into the source code, e.g. classes, methods, etc. An annotation allows the compiler to apply extra functionality to the code before its called, e.g. override annotations. MAUI uses annotations to identify methods that are resource-intensive for the device. Annotations are introduced within the source code by relying on the

Table 1 Code offloading approaches from a mobile and cloud perspectives

Code offloading strategies					Mobile perspective	Cloud perspective
Framework	*Main goal*	*Code profiler*	*Offloading adaptation context*	*Offloading characterization*	*Applications effect*	*Features exploited (Besides server)*
MAUI [11]	Energy-saving	Manual annotations	Mobile *(what, when)*	None	Low resource consumption, Increased performance	None
Odessa [26]	Responsiveness	Automated process	Mobile	None	Applications are up to 3x faster	None
CloneCloud [12]	Transparent code migration	Automated process	Mobile *(what, when)*	None	Accelerate responsiveness	None
ThinkAir [13]	Scalability	Manual annotations	Mobile + Cloud *(what, when, how)*	None	Increased performance	Dynamic allocation and destruction of VMs
COMET [15]	Transparent code migration (DSM)	Automated process	Mobile *(what, how)*	None	Average speed gain 2.88x	None

(continued)

Table 1 (continued)

Code offloading strategies					Mobile perspective	Cloud perspective
EMCO [14]	Energy-saving, Scalability (Multi-tenancy)	Automated process	Mobile + Cloud *(what, when, where, how, etc.)*	Based on historical crowdsourcing data	Based on context (Low resource consumption, increased responsiveness, etc.)	Dynamic allocation and destruction of VMs, Big data processing, Characterization-based utility computing
COSMOS [16]	Responsiveness	Manual process	Mobile *what*	None	Increased performance by choosing right surrogate	Resource allocation decided by user
HyMobi [28]	Energy-saving	Manual process	Mobile *what*	None	Energy saving based on social interaction	Resource allocation based on user's social context
Other work [2, 3]	Responsiveness	Manual annotations	Mobile *what, when*	None	Increased performance	None

expertise of the software developer. Once the code is annotated, MAUI transforms all the annotated methods into an offloadable format. This format equips the methods with RMI capabilities. Since MAUI targets Windows Phones, it is developed using *.NET* framework. Thus, RMI happens by using the WFC (Windows Communication Framework). During application runtime, the MAUI profiler collects contextual information, e.g. energy, RTT, etc., if the MAUI profiler detects a suitable context to offload code, then the execution of the code is delegated to a remote server instead of being performed by the device. While MAUI is successful in saving energy and shortening the response time of the mobile applications, it suffers from many drawbacks. Since MAUI uses code annotations, it is unable to adapt the execution of code in different devices. Thus, the developer is forced to adapt an application to a specific device, which is considered a brute-force approach. Moreover, MAUI suffers from scalability, which means that each mobile that implements MAUI requires to be attached to one specific server acting as a surrogate.

Similarly, CloneCloud [12] encourages a dynamic approach at OS level, where a code profiler extrapolates pieces of bytecode of a given mobile component to a remote server. Unlike MAUI, CloneCloud offloads code at *thread level*. CloneCloud uses static analysis to partition code, which is an improvement over the annotation strategy proposed by MAUI. By using a static analyzer, code can be annotated dynamically. Thus, code to offload is adapted based on the type of device without modifying or changing any implementation of the application. However, code profiling is complicate as its execution is non-determistic. Thus, it is difficult to verify the runtime properties of the code, which can cause unnecessary code offloading or even offloading overhead. Moreover, many other parameters also influence when choosing a portion to code to offload, e.g. the serialization size, latency in the network, etc.

COMET [15] is another framework for code offloading, which follows a similar approach as CloneCloud. COMET strategy puts emphasis on how to offload rather than what and when. COMET's runtime system allows unmodified multi-threaded applications to use multiple machines. The system allows threads to migrate freely between machines depending on the workload. COMET is a realization built on top of the Dalvik Virtual Machine and leverages the underlying memory model of the runtime to implement distributed shared memory (DSM) with as few interactions between machines as possible. COMET makes use of VM-synchronization primitives. Multi-thread offloading accelerates even further the execution of applications in which code can be parallelized.

ThinkAir [13] framework is one which is targeted at increasing the power of smartphones using cloud computing. ThinkAir tries to address MAUI's lack of scalability by creating virtual machines (VMs) of a complete smartphone system on the cloud. Moreover, ThinkAir provides an efficient way to perform on-demand resource allocation, and exploits parallelism by dynamically creating, resuming, and destroying VMs in the cloud when needed. However, since the development of mobile application uses annotations, the developer must follow a brute-forced approach to adapt his/her application to a specific device. Moreover, resource allocation in the cloud seems to be static from the handset as the device must be aware of the infrastructure

with anticipation. Thus, the approach is neither flexible nor fault tolerant. The scalability claimed by ThinkAir is not multi-tenancy, the system creates multiple virtual machines based on Android-x86 within the same server for code parallelization.

Odessa [26] is a framework that focuses on improving the perception of augmented reality applications, in terms of accuracy and responsiveness. The framework relies on automatic parallel partitioning at data-flow level to improve the performance of the applications, so that multiple activities can be executed simultaneously. However, the framework does not consider dynamic allocation nor cloud provisioning on demand, which is a key point in a cloud environment.

History-based approaches are also proposed to determine what code to offload [27]. However, the weak point of history-based approaches is the large amount of time required to collect data, which is needed to produce accurate results. Moreover, these strategies are sensitive to changes, which means that when the device suffers drastic changes, e.g. more applications are installed, the history mechanisms need to gather new data to calibrate again. The size of the data collected in the mobile can also be counterproductive for the device as it steals storage space and processing power [32].

COSMOS [16] is a framework that provides code offloading as a service at method level using Android-x86. The framework introduces an extra layer in a traditional offloading architecture to solve the mismatch between how individual mobile devices demand computing resources and how cloud providers offer them. However, it is not clear how the offloading process is encapsulated as SOA. Moreover, the framework is compared with CloneCloud, which is an unfair comparison as CloneCloud mechanisms offload code at thread level. Other frameworks for computational offloading also are proposed [3], but they do not differ significantly from basic implementation or concept [2, 3, 8]. Other frameworks focus on different issues, such as stability [30] and D2D (Device-to-Device) cooperation [28] among others.

We claim that the instrumentation of apps alone is insufficient to adopt computational offloading in the design of mobile architectures that relies on cloud. Computational offloading on the wild is shown mostly to introduce more computational effort to the mobile rather than reduce processing load [29]. In this context, CDroid [29] is a framework that attempts to improve offloading in real scenarios. However, the framework focuses more on data offloading than computational offloading. As a result, we propose EMCA, which attempts to overcome the issues of computational offloading in practice. EMCA automates the process of infering the right matching between mobile and cloud considering multiple levels of granularity using big data [14, 31].

4 Towards an Evidence-Aware Mobile Cloud Architecture

While the need to offload or not for mobile applications is debatable [19], the effectiveness of code offloading implementation in practice shows to be mostly unfavorable for the device outside controlled environments. In fact, the utilization of code offloading in real scenarios shows to be mostly negative [20], which means

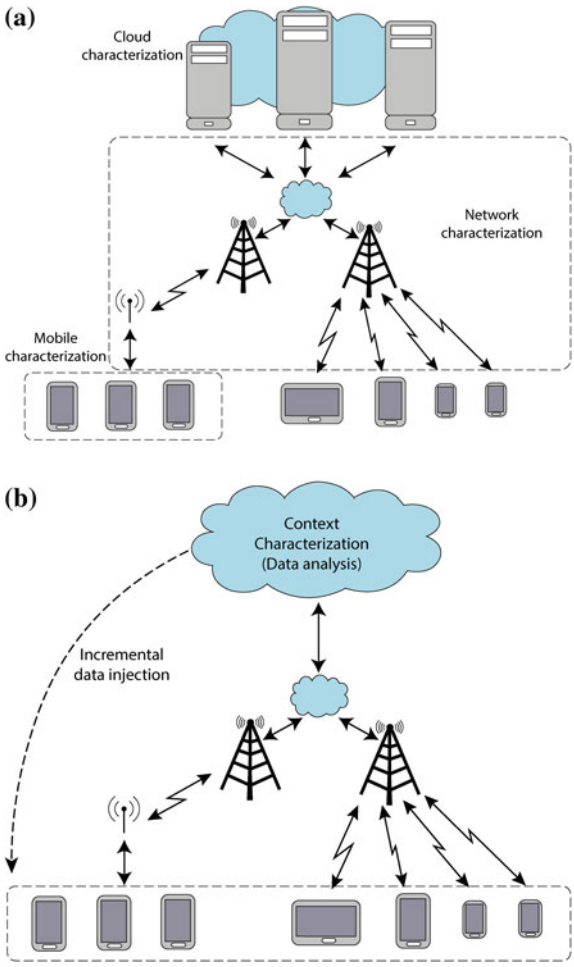


Fig. 3 Evidence-aware Mobile Cloud Architecture. **a** Characterization process of each component. **b** Diagnosis for each device posteriori to the characterization

that the device spends more energy on the offloading process compared to the actual energy that is saved. Consequently, the technique is far away from being adopted in the design of future mobile architectures. In section, we present our EMCA solution. EMCA relies on the smartphones to connect to cloud to characterize all the components of the architecture (Fig. 3a), e.g., network communication, cloud-based servers and type of devices, among others at different granularity levels, e.g., code, location, hardware specifications, etc. Once enough data is collected, then it is characterized in the cloud, such that the characterization can be used to create custom configurations for each particular device (Fig. 3b).

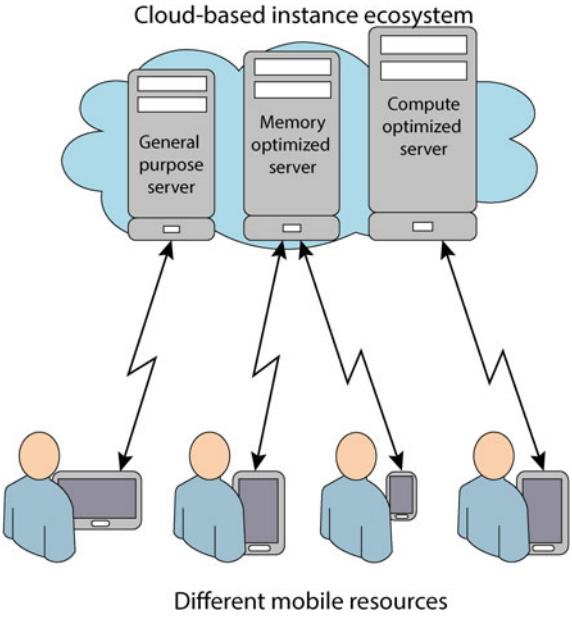


Fig. 4 Characterization of the offloading process that considers the smartphones diversity and the vast cloud ecosystem

4.1 Challenges and Technical Problems

Our goal is to highlight the challenges and technical obstacles of developing an EMCA. The issues are described as follows:

- ***Code partitioning approaches***—Code profiling is one of the most challenging problems in an offloading system, as the code has a non-deterministic behavior during runtime, which means that it is difficult to estimate the running cost of a piece of code considered for offloading. A portion of code becomes intensive based on multiple factors [20], such as user input that triggers the code, type of the device, execution environment, available memory and CPU, etc. Moreover, once code is selected as *OC*, it is also influenced by many other parameters of the system that come from multiple levels of fine-granularity, e.g. communication latency, data size transferred, etc. As a result, code offloading suffers from a sensitive tradeoff that is difficult to evaluate, and thus, code offloading can be productive or counterproductive for the device [33]. Most of the proposals in the field are unable to capture runtime properties of code, which makes them ineffective in real scenarios.
- ***Instrumentation complexity in the mobile applications***—The adaptation of code offloading mechanisms within the mobile development lifecycle depends on how easily the mechanisms are instrumented within the applications and how effective is the approach in releasing the device from intensive processing. However, implementation complexity does not necessarily correlate with effective runtime usage.

In fact, some of the drawbacks that make code offloading to fail are introduced at development stages, for example, in the case of manual code partitioning that relies on the expertise of the software developer, portions of code are annotated statically, which may cause unnecessary code offloading that drains energy [34]. Moreover, annotations can cause poor flexibility to execute the app in different mobile devices. Similarly, automated strategies are shown to be ineffective and require major low-level modifications in the core system of the mobile platform, which may lead to privacy and security issues.

- ***Dynamic configuration of the system***—Next generation mobile devices and the vast computational choices in the cloud ecosystem makes the offloading process a complex task as depicted in Fig. 4. Although the savings in energy that can be achieved by releasing the device from intensive processing, a computational offloading request requires to meet the requirements of user's satisfaction and experience, which is measured in terms of responsiveness of the app. Consequently, in the offloading decision, a smartphone has to consider not just potential savings in energy, but also it has to ensure that the acceleration in the response time of the request will not decrease. This is an evident issue as the computational capabilities of the latest smartphones are comparable with some servers running in the cloud, for instance, consider two devices, Samsung Galaxy S (i9000) and Samsung Galaxy S3 (i9300), and two Amazon instances, m1.xlarge and c3.2xlarge. In terms of mobile application performance, offloading intensive code from i9000 to m1.xlarge increases the responsiveness of a mobile application at comparable rates to an i9300. However, offloading from i9300 to m1.xlarge does not provide same benefit. Thus, to increase responsiveness is necessary to offload from i9300 to c3.2xlarge. It is important to note, however, that constantly increasing the capabilities of the back-end do not always speed up the execution of code exponentially, as in some cases, the execution of code depends on how the code is written, for instance, code is parallelizable for execution into multiple CPU cores (parallel offloading) or distribution into large scale GPUs (GPU offloading).
- ***Offloading as a service***—Typically, in a code offloading system, the code of a smartphone app must be located in both, the mobile and server as in a remote invocation, a mobile sends to the server not the intermediate code, but the data to reconstruct that intermediate representation so that it can be executed. As a result, an offloading system requires the surrogate to have similar execution environment as the mobile. To counter this problem, most of the offloading systems proposed to rely on the virtualization of the entire mobile platform in a server, e.g. Android-x86, .Net framework, etc., which tends to constrain the CPU resources and slows down performance. The reason is that a mobile platform is not developed for large-scale service provisioning. As a result, offloading architectures are designed to support one user at the time, in other words, one server for each mobile [13, 16, 39]. This restrains the features of the cloud for multi-tenancy and utility computing. Moreover while a cloud vendor provides the mechanisms to scale Service-Oriented Architectures (SOA) [4, 35, 40] on demand, e.g. Amazon autoscale, it does not provide the means to adapt such strategies to a computational offloading system as the requirements to support code offloading are different. The

requirements of a code offloading system are based on the perception that the user has towards the response time of the app [36]. The main insight is that a request should increase or maintain certain quality of responsiveness when the system handles heavy loads of computational requests. Thus, a code offloading request cannot be treated indifferently. The remote invocation of a method has to be monitored under different system's throughput to determine the limits of the system to not exceed the maximum number of invocations that can be handled simultaneously without losing quality of service. Furthermore, from a cloud point of view, allocation of resources cannot occur indiscriminately based on processing capabilities of the server as the use of computational resources are associated with a cost. Consequently, the need of policies for code offloading systems are necessary considering both, the mobile and the cloud.

- ***Utility model for code offloading***—A code offloaded task is accelerated differently based on the different underlying computational resources that can be acquired in the cloud [21]. While the cost of a server is charged by the cloud vendor based on time usage, e.g., an hour, it is unfeasible to create a bill for computational offloading following the same standard utility model. Since a task can have different levels of resource intensiveness, e.g., Chess, it requires different servers to deal with its specific processing requirements. As a result, a cloud deployment for code offloading comprises not one, but many servers that provisioning their computational resources to a mobile device. In this context, since a particular mobile application can use different cloud servers as surrogates in a single app session, then each offloading task that is offloaded needs to be charged based on the type of server that processed it. Naturally, this implies to change the current utility model of servers to one based on request-type, which introduces an extra level of complexity as it requires the execution of code to be segregated based on runtime properties, e.g., amount of acceleration required.
- ***Evidence usage within the mobile applications***—Since the characterization process is incremental based on data collected from the community of devices, evidence about the optimal configuration that is required to execute a mobile application needs to be transferred from the cloud periodically. Consequently, mechanisms to deliver and aggregate evidence need to be developed. Ideally, evidence should be delivered to the mobile device without introducing extra energetic overhead that harms its daily usage. Thus, evidence should be delivered by piggybacking data retrieval from other applications and services.

5 Discussion

In this section, we discuss about the opportunities and drawbacks of exploiting an EMCA.

1. ***Energy-aware offloading as a service for IoT (Internet of Things):***—
   It is well known that the main goal of MCC is to augment the processing capabilities and energetic resources of low-power devices, e.g., smartphones. To achieve this, applications installed in the devices are instrumented with offloading mechanisms, e.g., code offloading. However, despite of this instrumentation, applications are not aware about the productive or counterproductive effect that can be influenced in the mobile resources by outsourcing a task. For instance, how much the code should be accelerated?, how much energy can be saved? etc. In this chapter, we explore how to overcome the problem of determining the context required to offload a task by analysis in the cloud the runtime history of code execution from a community of devices. By relying on the massive computational resources of the cloud to process *big data*, we aim to exploit the knowledge of the crowd. However, many other sources of information collected from a community of devices can provide insight about how to configure the offloading process, e.g., sensor information, user's interaction, etc. To illustrate this, let's consider the following cases:
   Case 1: a smartphone that calculates and transmits its GPS coordinates every time the user uses an application. If the frequency of app usage is high, then the device will run out of energy quickly, e.g., facebook. If we assume that the end service in the cloud stores the data received, the data can be analyzed to build a prediction model in the cloud that suggests when the user changes his/her location. In this manner, the cloud service can be aware about the user's location and can configure the mobile app to recalculate and transmit GPS data when drastic changes of user's location are detected by the model. By implementing this approach, the device can save significant amounts of energy as the computational tasks of calculating and transmitting GPS data are not tied to app usage, but user's movement that is monitored by the cloud.
   Case 2: a low-power device, e.g., Arduino microcontroller, that monitors an environment via sensors, e.g., temperature. Since a client that connects to the microcontroller expects to obtain real time information, the microcontroller senses the environment regularly. Moreover, in order to provide scalability for multiple users, the environmental information is sent to the cloud, such that any user can access it from there. Naturally, this process requires considerable amount of energy of the device. However, by analyzing the collected data, it is able to equip the cloud service with the awareness to schedule the sensing process of the microcontroller based on opportunistic contexts, for instance, sensing data is likely to be replaced by other sensing data from a nearby device, sensing data can be predicted based on history data stored in the cloud, etc., in any situation, the main goal is to schedule from the cloud, the behaviour of the device, so that the device can be alleviated from unnecessary computational effort. Undoubtedly, it is expected that the change of behavior won't change the quality of service or experience of the user.
2. ***Tuning the fidelity of smartphone apps with mobile crowdsourcing***:—By characterizing the servers in the cloud, it is possible to identify multiple levels in which offloaded code is accelerated. Thus, we envisioned an approach to accel-

erate the response time of a mobile application dynamically. The ultimate goal of the approach is to enhance the QoE of the mobile apps in terms of *fidelity*, e.g., face recognition [41]. By improving the QoE, we aim to engage the user in order to increase application usage [37, 43].
Changing fidelity of mobile apps has been proved to be feasible by collecting data locally in the device [38]. However, this process is slow, because history data is required, and sensitive to changes, because the device is constantly upgrading and installing new apps. Thus, in order to overcome these problems, we envisioned fidelity tuning via data analytics from a community of devices.
Our idea is that apps are instrumented with mechanisms that capture their local execution at high level, e.g., method name, etc. This data is uploaded to the cloud for analysis. Based on the analysis, the cloud can perform individual diagnosis to each device and suggest optimal fidelity execution of each app installed in the device.

3. ***The effect of computational offloading in large scale provisioning scenarios***:—While the technique has been proved to be feasible with latest mobile technologies [14], still there are a lot of open issues regarding cloud deployment and provisioning in real scenarios. Previous work have proposed a one server per each smartphone architecture [19], which is unrealistic in practice if we consider the amount of smartphones nowadays and the provisioning cost of constantly running a server for a particular user.
Besides a few works that focus on scaling up (vertical scaling) a server to parallelize the code of computational requests [26], we have not found architectures that can scale in an horizontal fashion. This clearly can be seen as current frameworks do not take into consideration the utility computing features of the cloud, which is translated into server selection based on provisioning cost.
We are interested on analysis whether *it is possible to support large scale provisioning for computational offloading?* As a result, we want to study the capacity that cloud servers have to process multiple requests at once while maintaining requirements in code acceleration, which influences directly the response of a smartphone app. Moreover, we also want to analyze the effect of code acceleration in different cloud servers in order to foster surrogate selection based on utility computing, which can highlight new directions for the design of future mobile architectures supported by cloud computing, e.g., GPU offloading.
4. ***Context-aware hybrid computational offloading***:—Computational offloading is a promising technique to augment the computational capabilities of mobile devices. By connecting to remote servers, a mobile application can rely on code offloading to release the device from executing portions of code that requires heavy computational processing [42]. Yet, computational offloading is far away to be adopted as a mechanism within the mobile architectures, mainly due to drastic changes in communication latency to remote cloud can cause energy draining rather than energy saving for the device [14, 29]. Moreover, in the presence of high communication latency, the responsiveness of the mobile applications is degraded, which suggests that in order to avoid collateral effects, the benefits of

computational offloading can just be exploited in low latency proximity using rich nearby servers [28], which are also known as cloudlets.

Fortunately, 5G is arising as a promising solution to overcome the problem of high latency communication in cellular networks. 5G fosters the utilization of Device to Device (D2D) communication [30] to release the network from data traffic, and accelerate the transmission of data in end-to-end scenarios. By relying on D2D, and extrapolating features from remote cloud and cloutlets models, we envisioned a context-aware hybrid architecture for computational offloading. Our hybrid architecture introduces the concepts of network and cloud assistance, which can be utilized to coordinate the proximal devices in order to create a D2D infrastructure. Since the computational capabilities of next generation smarphones are comparable with some servers running in the cloud, we believe that multiple mobile devices can be merged together via D2D in order to create dynamic infrastructure in proximity that can be utilized by the devices themselves to share the load of processing heavy computational tasks. Naturally, this introduces new challenges mainly associated to social participation and collaboration.

Network assistance can be provided by cellular towers (Mobile Edge and Fog Computing [28]). The towers besides routing the communication between end-to-end points can be equipped with the logic to determine which devices are connected geographically close. When devices in proximity are detected, the tower can induce the devices to transmit data via D2D instead of using the cellular tower. The cellular towers can also be utilized to determine closer infrastructure (e.g., base stations), in which the device should be connected to reduce the communication latency, like in the cloudlet model. Similarly, cloud assistance can be utilized to group devices in a D2D cluster. Since devices are offloading to cloud-based servers (e.g., Amazon), the cloud can be equipped with the logic to determine which devices shared a common location. Cloud assistance introduces an extra level of complexity in the system than network assistance, due to a device is forced to send as part of the offloading process, the information about its location (e.g., GPS). However, cloud assistance alleviates completely the cellular network from computational offloading traffic, as all the process is managed entirely by the cloud.

6 Summary

Mobile and cloud computing convergence is shifting the way in which telecommunication architectures are designed and implemented. Several work have proposed different mobile offloading strategies to empower the smartphone apps with cloud based resources. Yet, the utilization of code offloading is debatable in practice as the approach has been demonstrated to be ineffective in increasing remaining battery life of mobile devices. The effectiveness of an offloading system is determined by its ability to infer opportunistically where the execution of code (local or remote) rep-

resents less computational effort to the mobile, such that by deciding *what, when, where and how to offload* correctly, the device obtains a benefit. Code offloading is productive when the device saves energy without degrading the normal response time of the apps, and counterproductive when the device wastes more energy executing a computational task remotely rather than executing it locally. Existing work offer partial solutions that ignore the majority of these considerations in the inference process. Thus, the approach suffers from many deficiencies, which are easily trackable in practice.

By characterizing the offloading process via crowdsensing, we explore the challenges and technical problems to overcome for developing an offloading architecture that learns to diagnose the optimal offloading process of a mobile application.

References

1. Flores, H.: Service-Oriented and Evidence-aware Mobile Cloud Computing. University of Tartu, Ph.D. thesis (2015)
2. Olteanu, A., Ţăpuş, N.: Offloading for Mobile Devices: A Survey, UPB Scientific Bulletin (2014)
3. Fernando, N., Loke, S.W., Rahayu, W.: Mobile cloud computing: a survey. Fut. Gen. computer systems, **29**, 1, 84 (2013)
4. Flores, H., Srirama, S.N.: Mobile cloud middleware. J. Syst. Soft. **92**, 82–94 (2014)
5. Flores, H., Srirama, S.N., Paniagua, C.: A generic middleware framework for handling process intensive hybrid cloud services from mobiles. In: Proceedings of the ACM International Conference on Advances in Mobile Computing and Multimedia (MoMM 2011), (Ho chi minh, Vietnam), Dec 5–7 (2011)
6. Mazzucco, M., Dumas, M.: Achieving performance and availability guarantees with spot instances. In: Proceedings of the IEEE International Conference on High Performance Computing and Communications (HPCC 2011), (Banff, Canada), September 2–4 (2011)
7. Han, B., Hui, P., Kumar, V.A., Marathe, M.V., Shao, J., Srinivasan, A.: Mobile data offloading through opportunistic communications and social participation, IEEE Trans. Mobile Comput. **11**, 5, 821 (2012)
8. Kaya, M., et al.: An adaptive mobile cloud computing framework using a call graph based model. J. Netw. Comput. Appl. **65**, 12–35 (2016)
9. Gu, X., Nahrstedt, K., Messer, A., Greenberg, I., Milojicic, D.: Adaptive offloading for pervasive computing. IEEE Perv. Comput. Mag. **3**(3), 66–74 (2004)
10. Kumar, K., Lu, Y.-H.: Cloud computing for mobile users: can offloading computation save energy? Comput. Mag. **43**(4), 51–56 (2010)
11. Cuervo, E., Balasubramanian, A., Cho, D.-k., Wolman, A.S., Saroiu, Chandra, R., Bahl, P. : Maui: making smartphones last longer with code offload. In: Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2010), (San Francisco, CA, USA), June 15–18 (2010)
12. Chun, B.-G., Ihm, S., Maniatis, P., Naik, M., Patti, A.: Clonecloud: elastic execution between mobile device and cloud. In: Proceedings of the ACM European Conference on Computer Systems (EuroSys 2011), (Salzburg, Austria), April 10–13 (2011)
13. Kosta, S., Aucinas, A., Hui, P., Mortier, R., Zhang, X.: Thinkair: dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM 2012), (Orlando, Florida, USA.), March 25–30 (2012)

14. Flores, H., Srirama, S.: Adaptive code offloading for mobile cloud applications: exploiting fuzzy sets and evidence-based learning. In: Proceedings of ACM MobiSys Workshop 2013, (Taipei, Taiwan), June 25–28 (2013)
15. Gordon, M.S., Jamshidi, D.A., S. Mahlke, Z. M. Mao, and X. Chen, comet: code offload by migrating execution transparently. In: Proceedings of USENIX Annual Technical Conference (ATC 2012) (Boston, MA, USA), June 13–15, (2012)
16. Shi, C., Habak, K., Pandurangan, P., Ammar, M., Naik, M., E. Zegura: Cosmos: computation offloading as a service for mobile devices, In: Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2014) (Philadelphia, PA, USA), August 11–14 (2014)
17. Flores, H., Hui, P., Tarkoma, S., Li, Y., Srirama, S., Buyya, R.: Mobile Code Offloading: From Concept to Practice and Beyond. IEEE Communications Magazine **53**(3), 80–88 (2015)
18. Verbelen, T., Simoens, P., De Turck, F., Dhoedt, B.: Cloudlets: Bringing the Cloud to the Mobile User, in Proceedings ACM MobiSys Workshop 2012, (Low Wood Bay, Lake District, United Kingdom), June 25–29 (2012)
19. Bahl, P., Han, R.Y., Li, L.E., Satyanarayanan, M.: Advancing the state of mobile cloud computing. In: Proceedings of ACM MobiSys Workshop 2012 (LowWood Bay, Lake District, United Kingdom), June 25–29 (2012)
20. Flores, H., Srirama, S.: Mobile code offloading: should it be a local decision or global inference? In: Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2013), (Taipei, Taiwan), June 25–28 (2013)
21. Flores, H., Sharma, R., Ferreira, D., Kostakos, V., Manner, J., Tarkoma, S., Hui, P., Li, Y., Manner, J.: Modeling mobile code acceleration in the cloud. In: Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS 2017), (Atlanta, GA, USA), June 5–8 (2017)
22. Skourletopoulos, G., Mavromoustakis, C.X., Mastorakis, G., Batalla, J.M., Pallis, E., Kormentzas, G.: Quantifying and evaluating the technical debt on mobile cloud-based service level. In: Proceedings of the IEEE International Conference on Communications (ICC 2016), (Kuala Lumpur, Malaysia), May 23–27 (2016)
23. Skourletopoulos, G., Mavromoustakis, C.X., Mastorakis, G., Batalla, J. M., Dobre, C., Panagiotakis, S., Pallis, E.: Big data and cloud computing: a survey of the state-of-the-art and research challenges. In: Advances in Mobile Cloud Computing and Big Data in the 5G Era, pp. 23–41 (2017)
24. Paniagua, C., et al.: Mobile Sensor Data Classification for Human Activity Recognition using MapReduce on Cloud. Procedia Comp. Sci. **10**, 585–592 (2012)
25. Satyanarayanan, M., Bahl, P., Caceres, R., Davies, N.: The case for VM-based cloudlets in mobile computing. IEEE Pervas. Comput. Mag. **8**, 4, 14–23 (2009). Evidence-aware Mobile Cloud Architectures 19
26. Ra, M.R., Sheth, A., Mummert, L., Pillai, P., Wetherall, D., Govindan, R.: Odessa: enabling interactive perception applications on mobile devices. In: Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2011), (Washington, DC, USA), June 28– July 1, 2011
27. Saarinen, A., Siekkinen, M., Xiao, Y., Nurminen, J.K., Kemppainen, M., Hui, P.: Smart-diet: offloading popular apps to save energy. ACM SIGCOMM Comput. Commun. Rev. **42**(4), 297–298 (2012)
28. Flores, H., Sharma, R., Ferreira, D., Kostakos, V., Manner, J., Tarkoma, S., Hui, P., Li, Y.: Social-aware hybrid mobile offloading. Pervas. Mobile Comput. J. **36**, 25–43 (2017)
29. Barbera, M.V., Kosta, S., Mei, A., Perta, V.C., Stefa, J.: Mobile offloading in the wild: findings and lessons learned through a real-life experiment with a new cloud-aware system. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM 2014), (Toronto, Canada), April 27–May 2 (2014)
30. Flores, H., Sharma, R., Ferreira, D., Kostakos, V., Manner, J., Tarkoma, S., Hui, P., Li, Y.: Social-aware device-to-device communication: a contribution for edge and fog computing? In: Proceedings of the ACM International Joint Conference on Pervasive And Ubiquitous Computing (UbiComp 2016): Adjunct, (Heidelberg, Germany), September 12–16 (2016)

31. Oliner, A.J., Iyer, A.P., Stoica, I., Lagerspetz, E., Tarkoma, S.: Carat: collaborative energy diagnosis for mobile devices. In: Proceedings of the ACM Conference on Embedded Networked Sensor (Systems 2013), (Rome, Italy), November 11–14 (2013)
32. Kchaou, H., Kechaou, Z., Alimi, A.M.: Towards an offloading framework based on big data analytics in mobile cloud computing environments. Procedia Comp. Sci. **53**, 292–297 (2016)
33. Chen, G., Kang, B.T., Kandemir, M., Vijaykrishnan, N., Irwin, M.J., Chandramouli, R.: Studying energy trade offs in offloading computation/compilation in java-enabled mobile devices. IEEE Trans. Parallel Dist. Syst. **15**(9), 795–809 (2004)
34. Miettinen, A., Nurminen, J.K.: Energy efficiency of mobile clients in cloud computing. In: Proceedings of the USENIXWorkshop on Hot Topics in Cloud Computing (HotCloud 2010), (Boston, MA, USA), June 22–25 (2010)
35. Flores, H., Srirama, S.N.: Dynamic Re-configuration of mobile cloud middleware based on traffic. In: Proceedings of the IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS 2012), (Las Vegas, Nevada, USA), October 8–11 (2012)
36. Flores, H., Su, X., Kostakos, V., Yi Ding, A., Nurmi, P., Tarkoma, S., Hui, P., Li, Y.: Largescale offloading in the internet of things. In: Proceedings of the IEEE Annual International Conference on Pervasive Computing and Communications (PerCom 2017): Adjunct, (Kona, Big Island, Hawaii, USA), March 13–17 (2017)
37. Balachandran, A., Aggarwal, V., Halepovic, E., Pang, J., Seshan, S., Venkataraman, S., Yan, H.: Modeling web quality-of-experience on cellular networks. In: Proceedings of the Annual ACM International Conference on Mobile Computing and Networking (MobiCom 2014), (Maui, Hawaii, USA), September 7–11 (2014)
38. Satyanarayanan, M., Narayanan, D.: Multi-fidelity algorithms for interactive mobile applications. Wireless Netw. J. **7**(6), 601–607 (2001)
39. Kristensen, M., Bouvin, N.O.: Scheduling and development support in the scavenger cyberforaging system. Pervas. Mobile Comput. J. **6**(6), 677–692 (2010)
40. Nawrocki, P., Reszelewski, W.: Resource usage optimization in mobile cloud computing. Comput, Commun (2016)
41. Silva, F.A., et al.: Mobile cloud face recognition based on smart cloud ranking. Computing. 1–25 (2016)
42. Schafer, D., et al.: Tasklets: better than best-effort computing. In: Proceedings of IEEE International Conference on Computer Communications and Networks (ICCCN 2016), (Waikoloa, Hawaii, USA), August 1–4, (2016)
43. Kwon, Y., et al.: Mantis: efficient predictions of execution time, energy usage, memory usage and network usage on smart mobile devices. IEEE Trans. Mobile Comput. **14**(10), 2059–2072 (2015)

Context-Awareness in Location Based Services in the Big Data Era

Patrizia Grifoni, Arianna D'Ulizia and Fernando Ferri

Abstract Integrating contextual information into the process of location-based service delivering is an emerging trend towards more advanced techniques aiming at personalization and intelligence of location-based services in the big data era. This chapter provides a systematic review of current context-aware location-based service systems using big data by analysing the methodological and practical choices that their developers made during the main phases of the context awareness process (i.e. context acquisition, context representation, and context reasoning and adaptation). Specifically, the chapter analyses ten location-based services, developed over the five years 2010–2014, by focusing on (1) context categories, data sources and level of automation of the context acquisition, (2) context models applied for context representation, and (3) adaptation strategies and reasoning methodologies used for context reasoning and adaptation. For each of these steps, a set of research questions and evaluation criteria are extracted that we use to evaluate and compare the surveyed context-aware location-based services. The results of this comparison are used to outline challenges and opportunities for future research in this research field.

1 Introduction

Big data represents a revolution for location-based applications [1]. These applications can produce and exchange massive amount of data in a very short time. The increasing explosion of data available on the web, together with the increasingly widespread use of telecommunication technologies, including wireless

P. Grifoni · A. D'Ulizia (✉) · F. Ferri
Consiglio Nazionale Delle Ricerche - IRPPS, Rome, Italy
e-mail: arianna.dulizia@irpps.cnr.it

P. Grifoni
e-mail: patrizia.grifoni@irpps.cnr.it

F. Ferri
e-mail: fernando.ferri@irpps.cnr.it

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_5

communications, Internet and mobile devices, has favoured the development of context-aware, ubiquitous computing methodologies for Location-Based Services (LBSs) provisioning. In the big data era, indeed, LBSs need to become selective, because only meaningful services from the massive amount of data have to be extracted in order to not overwhelm the computing resources of mobile devices. Therefore, integrating contextual information into the process of location-based service delivering is an emerging trend towards more advanced techniques aiming at personalization and intelligence of LBSs in the big data era.

Location-Based Services (LBSs), i.e. services that use the location of the user as the primary contextual information for delivering information to mobile users, have become increasingly popular due to the availability of powerful mobile devices equipped with positioning systems like GPS. There are several definitions for "*location-based services*": the first formal definition was given by Koeppel [2] in 2000, where they were defined as "*any service or application that extends spatial information processing, or GIS capabilities, to end users* via *the Internet and/or wireless network*". One most recent definition is given by Shiode et al. [3] and is the following: "*geographically-oriented data and information services to users across mobile telecommunication networks*". At the same time, LBSs are defined by Spiekermann [4] as "*services that integrate a mobile device's location or position with other information so as to provide added value to a user*". Finally, in 2010 Shek [5] defined LBS as "mobile computing applications that provide information and functionality to users based on their geographical location". From the above definitions, it is evident a shift of the environment, through which the services may be fruited, from networked environments to mobile-networked ones.

According to these definitions, dynamic navigation guidance, roadside assistance, mobile advertisements and traffic alerts can be considered examples of LBS, as they are services that use the location of the terminals to provide spatial information and GIS (geographic information system) functionalities to end users through the mobile, Internet, wireless or cloud networks [6, 7].

In LBSs, location plays a fundamental role as it determines the information and services the user may expect. Relying only on the location does not allow tailoring the answer to the user, since two persons asking for the same information in the same location receive the same answer despite their preferences may be different. Let us consider an example. We suppose two people, one vegetarian and one not, that search for restaurants near "*Piazza Navona*". Applying only location information, the LBS system returns the same list of restaurants located near "*Piazza Navona*" offering all kinds of cuisine (not only vegetarian cuisine).

Context awareness is a first step towards more advanced techniques aiming at personalization and intelligence of service provisioning in the big data era. Context awareness, indeed, is the ability to provide different services in different contexts, where context is more than location as it includes also personal (user activities, preferences, and needs), technical (e.g. device status), spatial, social, and physical (e.g. environmental status) information.

Therefore, concluding the example above, a context-aware LBS system can consider, for instance, also the information that the user is vegetarian, contained in

the personal context. Consequently, it restricts the results returned to the vegetarian person to restaurants located near "*Piazza Navona*" with a vegetarian menu.

Integrating contextual information about preferences, position and needs of the user, available resources and environmental features into the process of service delivering allows providing the user with more relevant services (among the massive amount of available services) that are better tailored to his/her needs. Therefore, context awareness is an essential feature that leads to a smart use [8] of the big amount of data that flows over the Internet in order to provide more and more personalized LBSs.

In this article, we are interested in surveying current context-aware LBSs by analysing the main phases of the context awareness process, that are the *context acquisition*, the *context representation*, and the *context reasoning and adaptation*. For each of these steps, we have analysed the methodological and practical choices that a LBS developer has to make during the design and implementation of a context-aware LBS system using big data. From this analysis, we have extracted a set of research questions and evaluation criteria that we have used to compare the surveyed context-aware LBSs. For this evaluation, ten LBS systems, developed over the five years 2010–2014, have been investigated.

The research contribution of this chapter is threefold. First, we introduce a new evaluation framework for context-aware LBSs, characterised by three orthogonal dimensions corresponding to the main phases of the context awareness process. Second, we use the evaluation framework to drive a comparative study of several current context-aware LBSs using big data. Finally, we outline challenges and opportunities for future research in this research field.

The structure of the chapter is organized as follows: we begin by providing some LBS categorizations from the literature and the evaluation criteria used to compare the context-awareness capabilities of current LBSs. Afterward, in Sect. 3 the main research issues of context acquisition and the corresponding evaluation criteria are described together with their application to evaluate the surveyed context-aware LBSs. Section 4 explores research issues and evaluation criteria for context representation and evaluates the context-aware LBSs with respect to that. Section 5 discusses research issues and evaluation criteria for context adaptation and describes their application to evaluate the surveyed context-aware LBSs. Based on this analysis, in Sect. 6 we investigate the open challenges of context-aware LBSs. Finally, Sect. 7 concludes the chapter.

2 Location-Based Services: Categorizations and Evaluation Criteria

In the literature, there are numerous classifications of LBSs. One first classification, proposed by Virrantaus et al. [9], distinguishes into pull and push services. In a pull service, the user makes explicitly a request to the service centre (as is the case of

dynamic navigation guidance and roadside assistance), while in push services the position of the user's device is utilized to estimate if s/he is a potential customer of the service and, if it is so, the information is automatically delivered to the user without his/her request (as is the case of mobile advertisements and traffic alerts).

A further classification of LBSs has been proposed by Schiller and Voisard [10] that distinguished between person-oriented and device-oriented. The former includes LBSs in which the position of a person is used to enhance the service and in which the user can control the service itself (e.g. navigation guidance), while the latter includes LBSs in which the position of a person or an object is tracked and the user is not necessary for controlling the service (e.g. car tracking).

Moreover, several classifications have been proposed that categorize LBSs according to the functionalities they provide. Reichenbacher [11] proposed the following five categories: orientation and localization, navigation, search, identification, and event check. Shek [5] introduced two further categories that are safety and emergency, and information services. Afterwards, Themistocleous et al. [12] re-arranged Reichenbacher and Shek's categories by proposing the following nine categories: identification, information, location, navigation and search, safety and emergency, access control, management, monitor, payments.

In this survey, we categorize and evaluate LBSs according to the methodological and practical choices that a LBS developer has to make during the design and implementation of a context-aware LBS system using big data. Malik et al. [13] consider a context-aware system composed of five main components, each one devoted to perform specific tasks of the context-awareness process: context acquisition, context representation, context storage, context interpretation, and context adaptation. Khattak et al. [14] list the following components of a context-aware system: context sensing, acquisition, representation, fusion, and reasoning. Perera et al. [15] describe four phases of the context-awareness process that are: context acquisition, context modeling, context reasoning, and context dissemination. A summary of these main phases of the context-awareness process proposed in the literature is provided in Table 1.

In this survey, we have taken into account these partitions in components of a context-aware system, for identifying the main steps of the context-awareness process that influence the design of a LBS system using big data. Specifically, we have considered the context-awareness process as composed of three major steps, which are the *context acquisition*, the *context representation*, and the *context reasoning and adaptation*. Respect to the existing partitions, we consider context sensing included in the context acquisition phase, the context storage included in the context representation phase and the context fusion and context dissemination included in the context reasoning and adaptation phase.

For each of the three identified steps we have analysed the main research questions that a LBS developer has to solve during the development of a context-aware LBS system using big data, which are summarized below (see Fig. 1 for a graphical summary):

Context acquisition: the context-awareness process starts with the sensing and gathering of contextual data. Several kinds of context data may be acquired,

Table 1 Main phases of the context-awareness process extracted from the literature

	Phases of the context-awareness process							
Malik et al. [13]		Context acquisition	Context representation	Context storage		Context interpretation	Context adaptation	
Khattak et al. [14]	Context sensing	Context acquisition	Context representation		Context fusion	Context reasoning		
Perera et al. [15]		Context acquisition	Context modeling			Context reasoning		Context dissemination
Our partition	Context acquisition		Context representation		Context reasoning and adaptation			

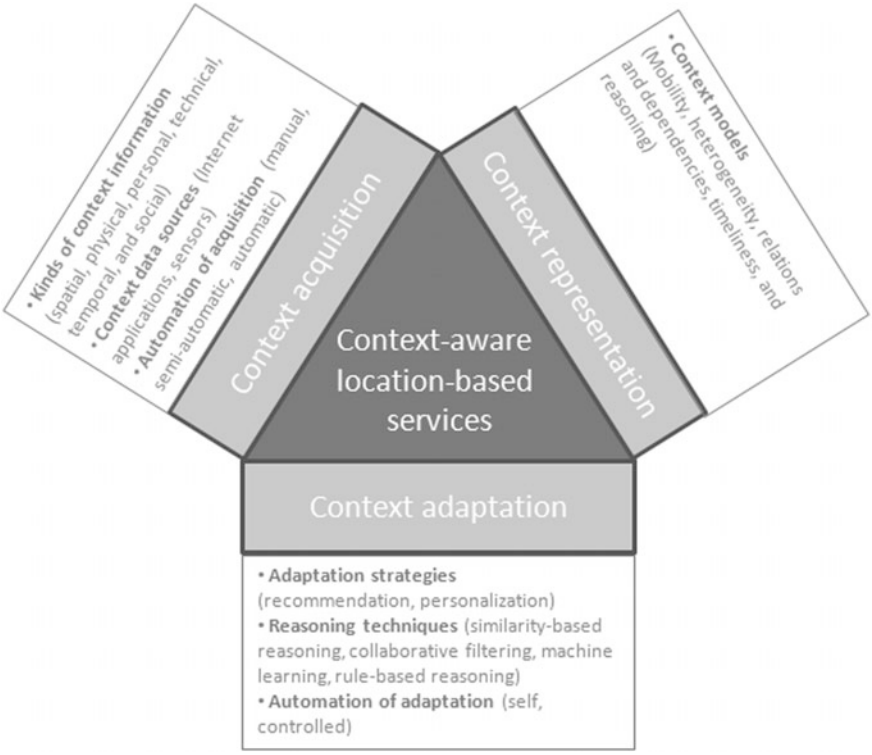


Fig. 1 Major steps of the context-awareness process along with the main evaluation criteria

depending on the specific application of the LBS system, ranging from spatial and temporal information to personal and social information. The acquisition can be usually done using various sources, such as sensors that are embedded into the mobile devices (mobile sensors) or present in the environment (static sensors), as well as monitoring or querying Web applications and services or asking explicitly to the user to input context information. Finally, the acquisition of context information from data sources can be characterised by various degrees of automation: manual, semi-automatic, and automatic. Therefore, the main issues that a LBS developer has to solve in this phase can be summarized in the following questions:

1. What kinds of contextual information have to be acquired in order to gather personalized results to the user request?
2. Through which knowledge sources?
3. Which level of automation should have the acquisition process?

Context representation: the gathered contextual data need to be represented through a data model that provides efficient structuring and retrieval. Therefore, the main issues that a LBS developer has to solve in this phase can be summarized in the following questions:

1. Which models have been applied for representing and managing location and contextual knowledge in order to properly use it in the discovering and adaptation processes?
2. How can these models be evaluated and compared?

Context reasoning and adaptation: the contextual data are used to discover, adapt and personalize services for the user. The context provides the basis for selecting the appropriate service among available services. The selection of the most appropriate services (mainly in the case numerous services of the same type are available) requires the use of some methodologies for reasoning on, filtering and ranking the available services, such as the collaborative filtering, similarity reasoning, etc. Therefore, the main issues that a LBS developer has to solve in this phase can be summarized in the following questions:

1. Which adaptation strategies have been applied to adapt query results to the captured contextual information?
2. Which reasoning and filtering techniques have been used to select the most appropriate services to be returned to the user?
3. Which level of automation should have the adaptation process?

All the context-aware LBS systems reviewed in this chapter have been analysed according to the aforementioned questions, which will be deeply discussed in the following sections.

3 Context Acquisition in LBSs in the Big Data Era

In this section, we analyse how LBS developers answered to the main research questions, introduced in the previous section, concerning the context acquisition. First of all, we analysed the categories of context information that may be acquired for adapting and personalizing LBSs (see Sect. 3.1). Second, we investigated the context data sources from which the various context categories can be acquired (see Sect. 3.2). Third, we reviewed the possible level of automation of the acquisition process (see Sect. 3.3). Finally, in Sect. 3.4 an overview of some LBS systems has been provided along with a discussion about how they answered to the three aforementioned research issues characterizing the phase of the context acquisition.

3.1 Context Categories

Despite the efforts made in the literature for providing a general definition of context, analogous efforts have been made by several researchers to categorize the contextual information needed to adapt and personalize the service request. One of the first categorisation was proposed by Schilit et al. [16] that proposed three

classes: *computing*, *user*, and *physical* context. Chen and Kotz [17] added the time category to this classification by arguing that context is a four-dimensional space composed of computing, *physical*, *time*, and *user* context. Zimmermann et al. [18] state that "elements for the description of this context information fall into five categories: *individually*, *activity*, *location*, *time*, and *relations*". Vieira et al. [19] divide user context into three categories: *physical*, *organizational* and *interaction*. Nieto et al. [20] categorized contextual information into two main classes, that are *static* and *situation* context, which are composed of *human* and *topographic* information (the static context), and *environmental*, *personal*, *location*, and *social* information, (the situation context). According to Hervás et al. [21], context can be categorized in *user*, *environment*, *devices*, and *services*. Finally, Köpper [22] categorized context information into five classes that are *personal*, *technical*, *spatial*, *social*, and *physical* context.

A summary of these main classifications is provided in Table 2. We have grouped over the same column the context categories that overlap or that are similar according to the examples of context attributes (that are provided in parentheses in Table 2 under the corresponding context category). For instance, the attribute "location" has been categorized over different context categories in the literature (see the second column of Table 2), which range from user context Schilit et al. [16], physical context Chen and Kotz [17], Vieira et al. [19], location context (Zimmermann et al. [18], Nieto et al. [20], and spatial context Köpper [22]. Analysing the overlapping of these categories, we have decided to group the location information over the spatial context category. Analogously, the environmental conditions have been categorised over physical context from the major number of literature classifications (see the third column of Table 2), therefore we have chosen physical context as main representative category for this kind of context information. Applying this grouping process to all context categories defined in the literature, we have identified six main context categories that are: *spatial*, *physical*, *personal*, *technical*, *temporal*, and *social*. Hereafter, we rely on this categorization when we refer to the context information. Specifically, each context category is characterized in the following way:

Spatial context refers to "the location and location-related information of the mobile user. It can be the precise geographical point (e.g. the latitude/longitude coordinates, the street address), or a personalized reference of the location (e.g. the home, the office, etc.)" [23].

Physical context refers to the physical status of the user and its surroundings, such as environmental status of the location (e.g. light, noise level, temperature, humidity, present objects, etc.) as well as the health status of the user (blood pressure, body temperature, etc.).

Personal context refers to all the information about the individual, contained typically in the user profile, such as age, gender, user's interest, attitudes, beliefs, etc., as well as the activities (tasks, roles) s/he is involved in, and people nearby.

Technical context refers to "the technical aspects related to computing capabilities and resources" [24] (e.g. network connectivity, bandwidth, memory, nearby resources, etc.).

Table 2 Main context information classifications extracted from the literature

Context information classification	Context categories					
	Spatial	Physical	Personal	Technical	Temporal	Social
Schilit et al. [16]	User (location)	Physical (lighting, noise levels, traffic conditions, temperature)	User (user profile, people nearby)	Computing (network connectivity, communication costs, communication bandwidth, nearby resources)		User (social situation)
Chen and Kotz [17]	Physical (location)	Physical (lighting, noise level, traffic condition, temperature, speed)	User (user profile, people nearby)	Computing (device connectivity, device capability)	Time (time of a day, week, month, season)	User (social situation)
Zimmermann et al. [18]	Location (name, coordinates)	Individually (environmental profile)	Individually (user profile, groups) activity (roles, tasks)	Individually (device profile)	Time (time zone, current time)	Relations (social relations, functional relations, compositional relations)
Vieira et al. [19]	Physical (location)	Physical (condition)	Organizational (user profile, group, roles, tasks)	Physical (device)		Interaction (synchronous or asynchronous interaction in a group)
Nieto et al. [20]	Topographic (coordinates) location (location name)	Environmental (temperature, light level)	Human (name, email, phone number) personal (status)			Social
Hervás et al. [21]	Environment (location)	Environment (lighting, noise level, traffic condition, temperature, speed)	User (user profile, status, tasks)	Devices (nearby resources) services (available services)		User (social situation)
Küpper [22]	Spatial (location name, location coordinates)	Physical (temperature, lighting, noise)	Personal (user profile)	Technical (connectivity, bandwidth)		Social (groups, relations)

Temporal context refers to the time references (e.g. time, date, and season) of events.

Social context refers to the social status of the user (e.g. social class, gender, or employment status), the social groups s/he belongs to (e.g. social network's groups) and the social roles s/he performs (e.g. parent, volunteer, employee, etc.).

3.2 Context Data Sources

Several different data sources have been used over the last decade to acquire context information, ranging from Web services, social media, static sensors, wearable devices, etc. In a recent survey, Zhang et al. [25] classify context data sources in three main categories that are: *Internet and Web services*, *static sensing infrastructures*, and *mobile devices and wearable sensors*. Starting from this classification, we have analysed the LBS literature for understanding what kinds of context data sources have been applied in LBS adaptation and personalization. This analysis resulted in a great use of both social media, such as social networks (and more specifically location based social networks), multimedia sharing sites, blogs, etc., and sensor networks (both static and mobile) as data sources for acquiring context data in LBSs. Therefore, we have evolved Zhang et al.'s classification by specifying two sub-classes of the category *Internet and Web services* (that we have re-named in *Internet applications and services*) that are *social media* and *Web-based information services*, and distinguishing between *Static sensors* and *Mobile and wearable sensors*, that have been grouped under the category *Sensors*. Specifically, each context data source is characterized in the following way:

Internet application and services

Web-based information services refer to services and applications that are available via the Internet for providing information. They can be used to acquire various types of context information, such as personal and temporal (through calendar, for example), spatial (through online maps), and physical (through traffic and weather info services, for example).

Social media refer to a kind of "Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content" [26], such as social networks, multimedia sharing sites, blogs, forums, etc. They can be used to extract information about personal preferences and interests, human activity, and social interactions.

Sensors

Static sensors, as defined by Dasgupta [27], "are sensors usually installed in fixed places in indoors and outdoors". They can be used to acquire various types of environmental parameters, such as temperature, humidity, pressure, pollution, light, etc., as well as human activity through surveillance cameras.

Mobile and wearable sensors are sensors attached to mobile devices or to the user's body. They can be used to acquire various parameters of a moving person,

Table 3 Context data sources for each context category

Context categories	Context data sources			
	Internet applications and services		Sensors	
	Web-based information services	Social media	Static sensors	Mobile and wearable sensors
Spatial	Maps	Media sharing site, location-based social networks	Wi-fi, surveillance cameras, cell tower	GPS, mobile cameras
Physical	Traffic and weather info services		Sensors for temperature, humidity, pressure, etc.	Health monitoring devices
Personal	Calendar	Social networks, media sharing site, forums, blogs, wiki		Mobile cameras, accelerometer
Technical	Yellow pages	Social networks	Cell tower	Bluetooth
Temporal	Calendar			GPS, Mobile cameras
Social		Social networks, forums, blogs, wiki		Mobile cameras

such as location (through GPS), acceleration, speed, etc., as well as healthcare status (blood pressure, body temperature, body weight etc.), and brain activity (EEG, EMG, etc.).

In Table 3, we show the proposed context data source categorization, along with some examples of context data sources that are used for acquiring each of the six context categories.

3.3 Automation of Context Acquisition

The acquisition of context information from knowledge sources can be characterised by various degrees of automation: manual (or explicit), semi-automatic, and automatic (or implicit) [28]. Manual knowledge acquisition means that the knowledge source requires user input for providing all necessary context information to the LBS. In automatic knowledge acquisition, user input is not required but the knowledge source is responsible of providing context information. Finally, semi-automatic knowledge acquisition requires the combined input of both the user and the knowledge source for providing context information to the LBS.

Generally, the spatial and temporal context is mainly acquired automatically through the GPS embedded in the user's device. Further automatic acquisition sources for spatial context are the surveillance cameras and the Wi-fi point access. Semi-automatic sources are the mobile cameras and the media sharing sites that

Table 4 Level of automation for each context category and data sources

Context categories	Level of automation		
	Manual	Semi-automatic	Automatic
Spatial		Mobile cameras, Media sharing site	GPS, Wi-fi, surveillance cameras
Physical			Sensors, health monitoring devices, traffic and weather info services
Personal	User profile	Mobile cameras	Social networks, media sharing sites, forums, blogs, wiki, accelerometer
Technical			Yellow pages, Social networks, Bluetooth
Temporal		Mobile cameras	GPS, calendar
Social		Mobile cameras	Social networks, forums, blogs, wiki

require the input of the user for recording and uploading multimedia data on the sharing sites. Physical context is usually acquired in an automatic way through sensors, health monitoring devices, or web information services. Personal context can be acquired manually, by asking the user to fill out a personal profile, semi-automatically by extracting information from videos recorded by the user through the mobile cameras, or automatically by deriving personal information from social media (user's personal website, social networks, media sharing sites, forums, etc.). Technical context is usually acquired in an automatic way through yellow pages, social networks, and Bluetooth. Social context can be acquired semi-automatically by extracting social information from videos recorded by the user through the mobile cameras, or automatically by deriving social information from social media in which the user is involved.

Table 4 summarizes the level of automation of the main context data sources according to the context categories.

Over the last years, there has been a shift from manual to automatic context acquisition. In an ideal scenario, context would be acquired automatically without the need for manual acquisition in order to reduce the user's workload. However, this scenario is not already feasible in the real world because some context information could not be sensed automatically and could be necessary to ask the user to manually provide it.

3.4 Context-Aware LBS Systems Using Big Data: Context Acquisition

This section takes an in-depth look at various context-aware LBS systems using big data, developed in the period 2010—2014 and listed in the first column of Table 5. The surveyed LBS systems have been analyzed according to the context categories, the context data sources, and the level of automation of context acquisition, defined in the previous sections. The results of this analysis are summarized in Table 5: the

second column provides the context categories used by each LBS system; the third column provides the context data sources used for acquiring context information; the fourth column provides the level of automation of the acquisition process.

The table shows that all the surveyed LBS systems acquire spatial context mainly from GPS embedded into the user device. This is quite obvious due to the nature of LBSs. Most of the systems (60%) use personal context for personalizing LBSs both acquiring it from social media and analyzing GPS data history. Technical context has been effectively employed by 50% of the surveyed LBS systems. 30% of the surveyed LBS systems use physical context acquired from web information services, social media and sensors. Finally, few systems (20%) apply temporal context acquired from GPS and social context acquired mainly from social media. All the surveyed LBS systems acquire context in an automatic way. All the data captured by the surveyed LBS systems has typical '4 V' features of big data, namely big volume, variety, high updating velocity, and low value.

iWISE [29] is a LBS cloud computing system that extracts location knowledge from Internet text, pictures, videos, and other related multimedia and uses this knowledge for supporting social awareness in the LBS discovery process. Spatial context is extracted from multimedia material searched on the Web by the user. A Web crawler is used to capture search multimedia content that is later analyzed by a text and picture information processing modules that extract geological locations. The spatial information is used to perform social awareness by recognizing and extracting individual behaviors and community interaction characteristics in real-time which are used for improving personalization and intelligence for LBSs.

Social Telescope [30] is a LBS that automatically compiles, indexes and ranks locations, based on user interactions with locations in mobile social networks, specifically by using geo-tweets made by Twitter users. A crawler records all public user geo-tweets and converts them into 4-tuples of the form <*user, location, time, text*>. Afterwards, locations are converted into semantic place names using the location-based social network Foursquare [31] and a ranking process is performed that ranks services according to the social popularity resulted from the crawled geo-tweets.

Biancalana et al. [32] propose a social recommender system that provides recommendations for location-based queries. They extract both physical context information (specifically, available service profiles and environmental conditions) from popular data sources on the Web (e.g. Yellow Pages, weather and traffic report services, Google Maps, Yelp, Zagat, etc.), and personal context information (preferences and interests of the user) from folksonomies, forums, blogs, and social networking sites.

SHERLOCK [33] is a system that provides LBSs based on the use of mobile agents, ontologies and semantic techniques for personalizing the service request. It acquires personal and technical context information from Web information

providers and from the other devices, connected in a peer-to-peer (P2P) network. A specific agent (called Alfred) is in charge of performing this acquisition task. Moreover, spatio-temporal context information is acquired from GPS and P2P network devices.

BOTTARI [34] is an application for personalized location-based recommendations rely on the opinions of the social media (specifically Twitter). It acquires two main context information: technical context (e.g. service descriptions) from Point-of-Interest's (PoI) websites and location-based social networks (e.g. Yelp, PoiFriend, Yahoo! Local, TrueLocal), and social context from Twitter. Tweets are crawled by the semantic media crawler and opinion miner that extract users' opinions on the PoI and rate them in positive, negative and neutral.

Zheng et al. [35] propose a collaborative recommendation system for location-based queries. This system extracts spatial and personal context information from GPS history data. Specifically, personal context (e.g. activity of the user) is extracted from the user-generated text comments that the user can add to the GPS data.

D'Ulizia et al. [36] provide a context-aware discovery system for delivering personalized LBSs. The personalization is performed according to the following five context categories: personal context (e.g. user profile) is acquired manually, since the user has to fill in a form with his/her personal information; spatial and temporal context are acquired from GPS; physical context (e.g. weather, traffic) is acquired from web information services; technical context (e.g. service descriptions) is acquired from web information providers and location-based social networks.

GeoSPLIS (Geographic Semantic Personalized Location Information System) [37] is a context-aware LBS that uses contextualized preferences of users regarding PoIs in order to personalize service delivering. To achieve that, the system collects personal context (e.g. user profile) and social context (e.g. groups and relationships) from the social network Google +, spatial context (e.g. location) and temporal context (e.g. time, day) from GPS, and the physical context (e.g. weather) from a web information provider.

KnockAround [38] is a P2P-based LBS application, which provides pull-type LBSs. It enriches the spatial context (e.g. the user's location), which is acquired from static and mobile sensors (GPS, Cell tower, and Wi-fi), with the technical context (e.g. service description) coming from text comments provided from surrounding users visiting the same location.

Bao et al. [39] propose a context-aware and preference-aware location recommender system. To achieve that, the system infers personal context (e.g. user preferences and expertise) from the location histories extracted from GPS data, and technical context (e.g. service descriptions) from the opinions of other people extracted from location-based social networks (e.g. Foursquare).

Table 5 The surveyed LBS systems and their context acquisition features

Context-aware LBS systems	Context categories						Context data sources				Level of automation
	Spatial	Physical	Personal	Technical	Temporal	Social	Internet applications and services		Sensors		
							Web-based information services	Social media	Static sensors	Mobile and wearable sensors	
iWISE [29]	√					√	Internet multimedia info			GPS	automatic
Social Telescope [30]	√							Mobile social networks			automatic
Biancalana et al. [32]	√	√	√				Yellow pages, weather and traffic report services, online maps	Folksonomies, social-based local search sites (e.g. Yelp, Foursquare), forums, blogs, Facebook pages		GPS	automatic
SHERLOCK [33]	√		√	√	√		Web information providers			GPS, Network of P2P devices	automatic
BOTTARI [34]	√		√	√		√	PoIs' websites	Location-based social networks (e.g. Yelp, PoiFriend, Yahoo! Local,		GPS	semi-automatic

(continued)

Table 5 (continued)

Context-aware LBS systems	Context categories						Context data sources				Level of automation
	Spatial	Physical	Personal	Technical	Temporal	Social	Internet applications and services		Sensors		
							Web-based information services	Social media	Static sensors	Mobile and wearable sensors	
								TrueLocal), Twitter			
Zheng et al. [35]	√		√							GPS	automatic
D'Ulizia et al. [36]	√	√	√	√	√		Weather and traffic info service, Web user profile, Web information providers	Location-based social networks		GPS	semi-automatic
GeoSPLIS [37]	√	√	√		√	√	PoIs' websites, Weather info service	Google+		GPS	automatic
KnockAround [38]	√			√					Cell tower, Wi-fi	GPS, Network of P2P devices	automatic
Bao et al. [39]	√		√	√				Location-based social networks		GPS	automatic

4 Context Representation in LBSs in the Big Data Era

The gathered contextual data need to be represented through a context model that provides efficient structuring and retrieval of the huge amount of this gathered data. In this chapter, we use the term "context model" to refer to the generic underlying data structures and available operations that can be performed on them.

Several classifications of models for contextual knowledge representation have been proposed in the literature, according to the scheme of data structure that has been used to exchange contextual information. Chen and Kotz [17] identify the following four context model categories: *key-value*, *tagged encoding*, *object-oriented*, and *logic-based* models. Strang and Linnhoff-Popien [40] classified context models in the following six categories: *key-value*, *markup scheme*, *graphical*, *object oriented*, *logic based*, and *ontology based* models. Bettini et al. [41] classify context modeling approaches into *key-value*, *markup-based*, *object-role-based*, *spatial*, and *ontology-based* models.

A further classification more specific for big data models is given by Pop and Cristea [42] that considers the following categories: *structured data, text file data, semi-structured data, key-value pair data, XML data,* and *RDF (Resource Description Framework) data*. This classification is referred to big data models and does not focus specifically on contextual data.

A summary of these main classifications is provided in Table 6. In the first three rows, we have grouped over the same column the context model categories that overlap or that are similar according to the examples of context models provided in the papers. Analysing the overlapping of these categories, we have observed that the classification proposed by Strang and Linnhoff-Popien is the most inclusive one. Therefore, hereafter we rely on that classification.

Comparing the classification of big data models provided by Pop and Cristea [42] with the classification of context data models provided by Strang and Linnhoff-Popien [40], we can observe that the key-value and markup scheme (e.g. XML) categories are present in both classifications. The main difference between these two classifications relies on the emphasis that is given to the structuring of data. In big data models, indeed, the manner in which information is structured profoundly influences the efficiency of the big data processing, exchange and analysis. Therefore, the classification of big data models places a greater importance on the level of data structuring.

To have a picture of the kinds of context and big data models applied in current LBS systems, we have checked which kinds of models, among those categorised by Strang and Linnhoff-Popien and Pop and Cristea, have been applied in the surveyed LBS systems. Table 7 provides the results of this analysis. Specifically, we observed that key-value and ontology-based models have been used as context data representation models in the analysed LBS systems. Therefore, our analysis will focus only on the following two classes: *key-value* and *ontology-based* models. Specifically, each class is characterized in the following way:

Table 6 Main classifications of context models and big data models extracted from the literature

Context model classification	Context model categories					
Chen and Kotz [17]	Key-value	Tagged encoding	Object oriented	Logic-based		
Strang and Linnhoff-Popien [40]	Key-value	Markup scheme	Object oriented	Logic based	Ontology-based	Graphical
Bettini et al. [41]	Key-value	Markup-based			Ontology-based	Object-role-based spatial
Big data model classification	Big data model categories					
Pop and Cristea [42]	Key-value pair data	XML data	RDF data	Structured data	Text file data	Semi-structured data

Table 7 The surveyed LBS systems and their data model representations

Context-aware LBS systems	Context representation		Big data representation			
	Key-value	Ontology	Key-value pair	Semi-structured	RDF data	Structured data
iWISE [29]		√	√			
Social Telescope [30]	√			√		
Biancalana et al. [32]	√			√		
SHERLOCK [33]		√			√	
BOTTARI [34]		√			√	
Zheng et al. [35]	√					√
D'Ulizia et al. [36]	√					√
GeoSPLIS [37]		√			√	
KnockAround [38]	√		√			
Bao et al. [39]	√		√			

key-value models refer to context models that use key-value pairs that identify the attributes and their values describing the context [43].

ontology-based models refer to models that use ontologies to represent concepts and relations between concepts. They represent a uniform way for specifying the model's core concepts as well as sub-concepts and facts, thus enabling contextual knowledge sharing and reuse [44]. Ontologies provide a generic and, at the same time, formal way to "capture and specify the domain knowledge with its semantics" [45] and, therefore, they turn out to be appropriate for handling contextual knowledge. Various modeling web languages have been developed to express ontologies, most of them based on XML [43] syntax. On 2004, the World Wide Web Consortium (W3C) included two modeling web languages as recommended semantic Web technology standards: Resource Description Framework (RDF) [46] and Web Ontology Language (OWL) [47]. Both RDF and OWL provide a standard for metadata about resources on the Web; however, OWL is an extension of RDF, built to cope with the limitations of RDF, such as, for instance, the ability to define classes in terms of other classes. Therefore, most of recent ontology-based LBS systems use OWL as language to formalize context knowledge, as described in the following ontology-based LBS systems.

Considering big data representation models, we observed that *key-value pair, structured, semi-structured,* and *RDF data* models were used. Key-value pairs model is the same described above. The other three classes can be described as follows:

Structured data models have records of columns and each column has a value the meaning of which is consistent from record to record.

Semi-structured data models refer to models where there is no separation between the data and the schema, and the amount of structure used depends on the purpose.

RDF data models use triples that follow a subject–predicate–object structure. The subject denotes the resource, and the predicate denotes traits or aspects of the resource, and expresses a relationship between the subject and the object.

The first two columns of Table 7 show the surveyed LBS systems and their context representations. The table shows that most of the surveyed context-aware LBS systems (60%) use key-value model for representing context information. Ontologies have been effectively employed by 40% of the surveyed LBS systems.

The last four columns of Table 7 show the surveyed LBS systems and their big data representations. Most of the surveyed context-aware LBS systems use key-value model (30%) and RDF data model (30%) for representing context information.

The large use of the key-value model can be justified by its main simplicity to implement and manage context information. However, this model has the main drawback that is the limited capabilities in describing and managing complex context information. Ontology-based model outperforms key-value model in the capability to specify very complex context information, providing capabilities for reasoning, knowledge sharing and reusing. On the other hand, ontology-based model requires expensive efforts in defining the ontology.

In order to provide an evaluation of the surveyed LBS systems, we have extracted a set of requirements, both from those proposed by Bettini et al. [41], which can be applied for assessing the context models, and from those proposed by Agrawal et al. [48], which can be applied for assessing the fulfillment of the big data challenge. Specifically, the requirements of Bettini et al. [41] are the following:

1. mobility: the context model is able to adapt to the mobile environment;
2. heterogeneity: the context model is able to express and manage different types of contextual information coming from multiple data sources;
3. relationships and dependencies: the context model is able to capture various relationships, in particular dependency, between different context information;
4. timeliness: the context model is able to capture context histories (past and future states);
5. reasoning: the context model is able to support context reasoning techniques in order to derive new context facts from existing ones.

The requirements that we have extracted from of Agrawal et al. [48] are the following:

6. scalability: the data model is able to scale up the system's use of the data by allowing it to handle an increasing variety of data sources;
7. privacy: the data model is able to prevent the inappropriate use of personal data, which can particularly arise from linking of data from multiple sources;

8. human collaboration: the data model is able to support input from multiple distributed users, and their collaboration.

The fulfilment of these requirements ensures good performance of the context model also from the point of view of the big data challenge.

In the remainder of this section, we take an in-depth look at the contextual knowledge representation adopted by the surveyed LBS systems. Moreover, we conclude this section with an evaluation of these systems according to all the requirements introduced above.

4.1 Context-Aware LBS Systems Using Big Data: Context Representation

This section takes an in-depth look at the context representation used by various context-aware LBS systems using big data, developed in the period 2010–2014.

Social Telescope [30] uses a knowledge repository composed of 4-tuples of the form *<user, location, time, text>*. The values of the tuples are extracted from all public user geo-tweets crawled from Twitter. An indexer maintains indexes corresponding to the location, tags and user names, and updates the indexes each time a new tuple is added in the repository.

Biancalana et al. [32] rely on a local database that is populated with information extracted from Web data sources (Yellow Pages, Google Maps, Yelp, and Zagat that provide business listings, phone numbers, and addresses). All this information is stored in the local database along with plain tags, semantic tags, and sub-categories. Plain tags are extracted by a keyphrase extraction module that retrieves meaningful keyphrases from documents, while semantic tags and sub-categories are extracted by the source-specific extractor. In Fig. 2 an example of tuple stored in the local database is shown.

Zheng et al. [35] model the context information through multi-dimensional arrays, called tensors. Specifically, users, locations, and activities are extracted from GPS history data and they are stored in user-location-activity 3D tensors. Each entry of the tensor denotes the frequency of a user visiting a location and doing an activity there. Figure 3 shows an example of 3D tensor, where the bi-dimensional array on the left represents the frequency of visiting Forbidden City, Bird's Nest, and Zhongguancun for Tourism, Exhibition and Shopping for the user named Vincent.

Name	Category	Sub-category	City	Address	Phone	...	Plain Tags	Semantic Tags
Eataly	Restaurant	Italian	Rome	Piazzale XII Ottobre, 1492	06 9027 9201	...	{wine, pasta, pizza, meat}	{Take away: yes, Meal served: Dinner, ...}

Fig. 2 An example of tuple populating the local database of Biancalana et al. Figure adapted from [32]

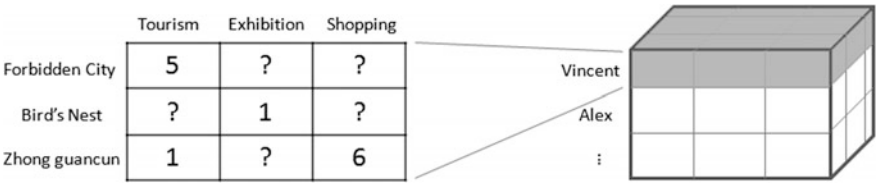


Fig. 3 An example of user-location-activity tensor of Zheng et al. Figure adapted from [35]

D'Ulizia et al. [36] applies a key-value model to represent context information. Specifically, they define three kinds of profiles (i.e. user, context, and service profiles), each one represented as a set of typed attributes continuously monitored and updated by the system. Figure 4 shows an example of user, context, and service profiles used by D'Ulizia et al.: the first column contains the name and type of the attributes and the last column contains the current value.

KnockAround [38] uses a database of context information that is distributed over people's smartphones and is dynamically updated by the users themselves. A specific module, called Database populator, is in charge of continuously keeping track of the user's location and, if the current location is not yet registered in the database, requesting to the user to provide information about it. This information is stored in a database containing 11 attributes that are: *id, latitude, longitude, name of place, address, keyword, comment, date, shareable* (boolean), *event* (boolean), and *bloom filter*.

An instantiation of the user profile's schema

User profile	User name	Mario Rossi
Typed attributes	Residence: string	Residence: Florence
	Profession: string	Profession: Researcher
	Age: integer	Age: 35
	Language: string	Language: {English, French}
	Sport: string	Sport: {football, basketball}
	Cuisine: string	Cuisine: {vegetarian, Japanese}
	Hobby: string	Hobby: {gardening, how-to}

An instantiation of the context profile's schema

Context profile	User name	Mario Rossi
Typed attributes	User's location: spatial coordinates	User's location: (39.65749, –105.10458)
	Current_time: time	Current_time: 18:00
	Weather: string	Weather: sunny
	Traffic: string	Traffic: low

An instantiation of the service profile's schema

Service profile	Service name	Museum
Typed attributes	Location: spatial coordinates	Location: (39.65677, –105.10444)
	Ticket price: real	Ticket price: 7,5
	Style: string	Style: {modernism, futurism, surrealism}
	Opening day: string	Closing day: Thursday
	Opening hours: time	Opening hours: {9:00–18:00}
	Tour time: int	Tour time: 3
	Guide language: string	Guide language: {English, French, Spanish}

Fig. 4 The user, context, and service profiles used by D'Ulizia et al. Figure adapted from [36]

Bao et al. [39] use the following five key data structures in the form of tuples of repositories and weighted hierarchy: *user repository*, which contains user profile information, such as ID, name, age, gender, etc.; *check-in repository*, which contains the locations visited by the user and commented in a location-based social networks; *venues*, which are the locations associated with a pair of coordinates on a map and a set of categories; *user location history*/matrix, in which each entry denotes the number of visits of a user to a location; *category hierarchy*, which is a weighted hierarchy where nodes occurring on a deeper layer denote the categories of a finer granularity.

iWISE [29] relies on a location knowledge ontology that aggregates and indexes all context information acquired from Internet text and pictures, as shown in Fig. 5. This ontology is based on a multi-level location index that indexes the relationships between locations. Each location contains the location geological information obtained by processing Internet textual information (title, abstract, URL, etc.), and pictures information (picture, URL, etc.).

SHERLOCK [33] uses an OWL ontology to represent and manage, in a distributed way, the context knowledge. The system starts with a basic ontology containing the user's common knowledge (device capabilities, username, etc.) and basic terms. An agent, called Ontology Manager, has in charge of keeping the ontologies shared by other surrounding devices and integrating them into its own local ontology. Figure 6 shows the ontology defined in SHERLOCK for a location-based transportation service.

BOTTARI [34] represents the context information through an ontology, shown in Fig. 7, which uses two W3C vocabularies that are the SIOC vocabulary [49] for defining *UserAccount* and *Post* and the WGS-84 vocabulary [50] for *SpatialThing*. These two standard ontologies have been extended for representing the other

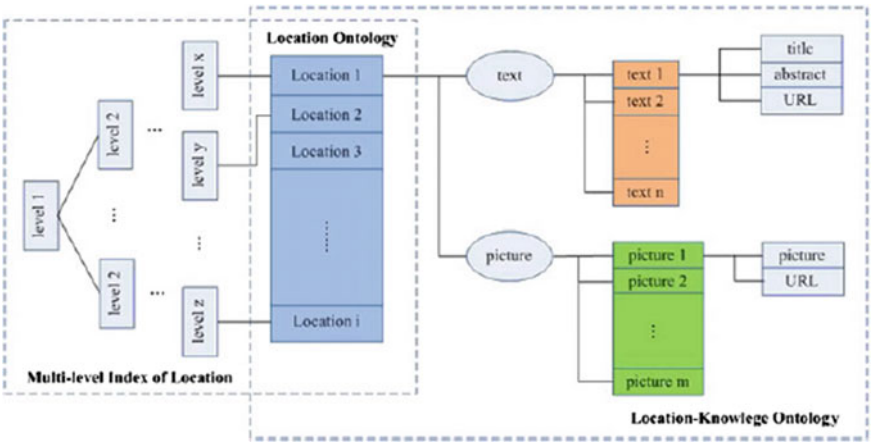


Fig. 5 The location knowledge ontology used by iWISE. Reprinted with kind permission from Springer Science + Business Media: [29]

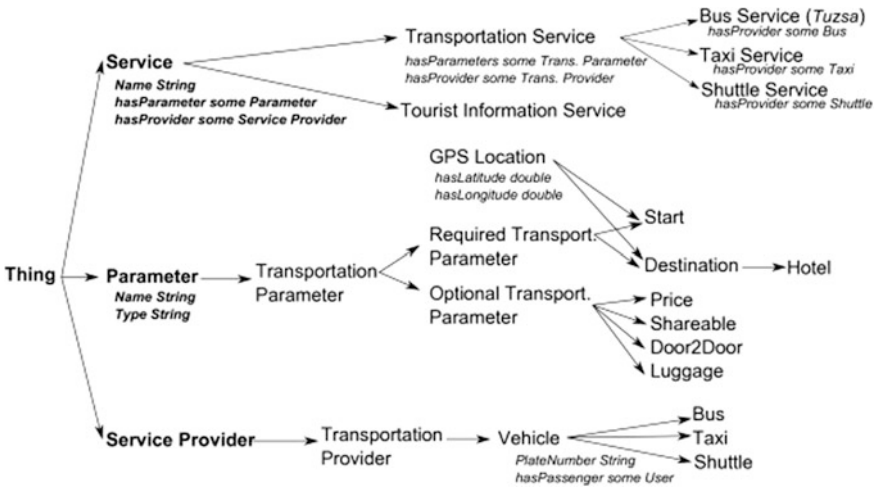


Fig. 6 The ontology used in SHERLOCK [33] for defining a location-based transportation service. Reprinted from [33], with permission from Elsevier

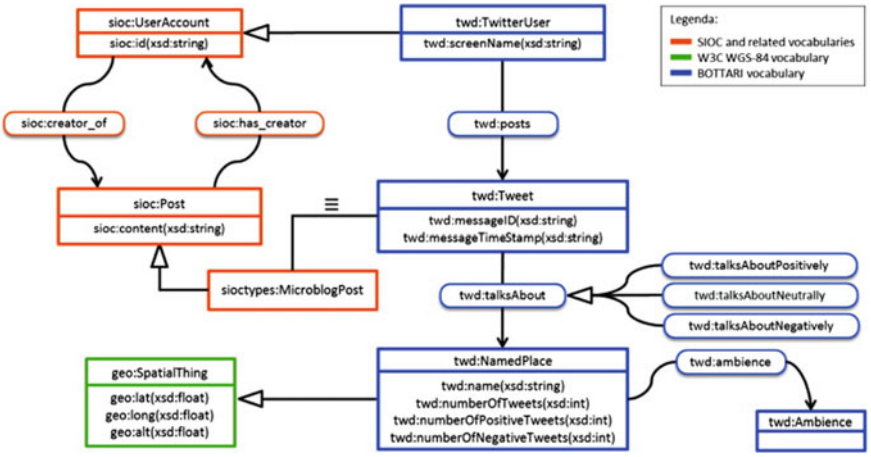


Fig. 7 The BOTTARI ontology. Reprinted from [34], with permission from Elsevier

technical, social and spatial context features (i.e. *TwitterUser*, *Tweet*, *NamedPlace*, *Ambience*).

GeoSPLIS [38] adopts the *schema.org* ontology (http://schema.org/) to represent people and place profiles and incorporates dynamically its RDF Schema version. This ontology allows representing physical and digital entities (persons, places, movies etc.), and also the connections between them. All data are stored in RDF format using Sesame [51], an architecture for storing and querying RDF data.

Table 8 Evaluation of the context models of the surveyed LBS systems (ontology-based are highlighted in dark-grey, key-value are highlighted in light-grey)

Context-aware LBS systems	Context model requirements							
	Mobility	Heterogeneity	Relationships and dependencies	Timeliness	Reasoning	Scalability	Privacy	Human collaboration
iWISE [29]	+	+	+	+	–	+	–	+
Social Telescope [30]	–	–	–	+	–	+	+	+
Biancalana et al. [32]	+	+	–	–	–	+	–	+
SHERLOCK [33]	+	+	+	+	+	+	–	+
BOTTARI [34]	+	+	–	+	+	+	–	+
Zheng et al. [35]	+	–	–	+	–	+	+	+
D'Ulizia et al. [36]	+	–	–	+	–	–	–	–
GeoSPLIS [37]	–	+	+	+	+	–	+	+
KnockAround [38]	+	–	–	+	–	–	+	–
Bao et al. [39]	–	–	+	–	–	+	–	+

4.2 Evaluation of Context Representation

This evaluation consisted in the analysis of the fulfillment of the requirements introduced in Sect. 4 by the context models used in the surveyed LBS systems. Table 8 summarizes the results of this evaluation. The plus sign indicates that the context model fulfills the corresponding requirement, otherwise the minus sign is used.

From the analysis of these results, we can observe that ontology-based models (highlighted in dark-grey) meet most of the requirements. Specifically, all models meet heterogeneity and human collaboration requirements and captures timeliness. The majority of them (75%) meets also the mobility and scalability requirements, captures relationships and dependencies of context information and supports context reasoning. These results are consistent with the strengths of ontology-based models described by Bettini et al. [41] that are the fulfillment of heterogeneity, relationship and reasoning requirements.

On the contrary, key-value models (highlighted in light-grey) do not meet many of the requirements. The most fulfilled requirements are the mobility, timeliness, scalability and human collaboration that are met by 66% of key-value models. Half of key-value models meets privacy requirement. Only one model (17%) meets heterogeneity requirement and captures relationships and dependencies of context information. Finally, no key-value model supports reasoning. These results are consistent with some of the limitations of key-value models described by Bettini et al. [41], which are the limited capabilities both in capturing relationships, dependencies, timeliness and quality of context information, and in supporting reasoning on context.

Therefore, ontology-based models outperform key-value models in all the evaluated requirements.

5 Context Reasoning and Adaptation in LBSs in the Big Data Era

After to be acquired and modelled, the contextual data are used to adapt services for the user. Context adaptation, as defined by Klein et al. [52], is a system's capability of gathering information about the domain, evaluating this information and changing its observable behaviour according to the current situation. In context-aware LBSs, context adaptation means that the system is able to gather contextual information and, according to that, to discover the appropriate service among the available ones.

In this section, we analyse how LBS developers answered to the main research questions, introduced in Sect. 2, concerning the context adaptation. First of all, we analysed the adaptation strategies that can be applied for adapting LBSs (see Sect. 5.1) to the current context. Second, we investigated the reasoning and filtering

methodologies that can be used to select the most appropriate services to be returned to the user (see Sect. 5.2). Third, we reviewed the possible degree of automation of the adaptation process (see Sect. 5.3). Finally, in Sect. 5.4 a discussion about how the surveyed LBS systems answered to the three research issues characterizing the phase of the context acquisition.

5.1 Adaptation Strategies

Adaptation is a fundamental property for the design of flexible LBS systems. Generally, various forms of adaptation can be applied according to the kinds of contextual information that is considered for adapting the list of services related to the user request.

Benazzouz [53] identifies three classes of adaptation strategies: personalization, recommendation, and reconfiguration. Personalization consists in the process of tailoring system's functionalities and behaviour to respond to the contextual information (mainly user preferences). Recommendation makes use of user past behaviour and preferences (user history) and community opinions to tailor system's functionalities and behaviour. Finally, reconfiguration considers only the technical context (the technical aspects related to computing capabilities and resources, see Sect. 3.1) and consists in replacing a system's component that is no longer able to deliver a service with similar services.

In this survey, we do not take into account reconfiguration, as it is not applied for context-awareness by the surveyed LBS systems due to its limitation to adapt to technical context only. Therefore, we consider personalization and recommendation. Of course, these two adaptation strategies are not exclusive and there are some LBS systems which apply both together.

Personalization and recommendation are also addressed by Mokbel et al. [54] as two main aspects of LBS 2.0, i.e. the new generation of LBSs, where users can generate location-based content to be shared through location-based interaction with other users. According to Mokbel et al., personalization consists in allowing users to express their preferences which are taken into account when answering queries. Recommendation consists in extracting community opinions and user behaviour in order to identify the list of services to be returned to the user.

In the following sub-sections we give some more details about these two main adaptation strategies applied in LBS systems.

5.1.1 Personalization

Personalization consists in the process of tailoring system's functionalities and behaviour to respond to the contextual information (mainly user preferences). Zimmermann et al. [55] used the term personalization to refer to the tailoring of

products, services, or content to user needs, goals, knowledge, interests, or other characteristics.

In the provisioning of LBSs, personalization means to adapt the delivered services to the preferences of the user. Therefore, the LBS system will elaborate different answers to the same query depending on which user is querying the LBS. In other words, the system delivers a set of services that is custom-tailored to the individual needs.

To achieve that, the early approach that has been proposed in the literature is the preference query processing that aims at finding the best answer according to a certain preference method (such as top-k [56], skylines [57], k-dominance [58], etc.). In this approach the best answer results from the preference method that considers on the same level both spatial and non-spatial attributes. A further approach is the k-nearest-neighbour (KNN) query processing [59] that shifts the concept of "best" answer to the concept of "closest" answer by using distance-based measures. A middle way between the two previous approaches is the k-best-neighbour (KBN) query processing, in which both the user preferences and context are taken into account for delivering the "best" services to the user.

5.1.2 Recommendation

Recommendation makes use of user past behaviour and preferences (user history) and community opinions to tailor system's functionalities and behaviour.

To achieve that, the main techniques used to enhance recommendation are collaborative filtering and content-based filtering. Collaborative filtering [60] is based on the opinions that people provide about available services. Opinions are expressed through ratings that are then analysed to find similarities between users and to predict possible services that user might like. Content-based filtering recommends services by analyzing service properties that are similar in individual's past queries. It requires the availability of service information and the monitoring of user behavior.

Therefore, the main difference between collaborative filtering and content-based filtering is that the former only uses ratings data to make predictions and recommendations, while the latter uses features of users and services.

Several surveyed context-aware LBSs (that are Social Telescope, Biancalana et al., BOTTARI, Zheng et al., KnockAround, Bao et al.) apply collaborative filtering to make recommendation, and no LBS applies content-based filtering. This is mainly due to the enhancement of the filtering process that is produced involving people in the process of service rating. This is also confirmed by the increasing availability of social networking systems and web portals that ask to rate and provide opinions on services.

5.2 Reasoning Methodologies

LBS systems can be categorized according to the technique for reasoning and filtering the available services and selecting the most appropriate ones (in case multiple services of the same type match with the user request) to be returned to the user.

In the LBS literature, the following four kinds of service filtering techniques can be distinguished: techniques based on similarity-based reasoning, techniques based on collaborative filtering, techniques based on machine learning, techniques based on rule-based reasoning. A brief description of these categories is given in the following sub-sections.

5.2.1 Similarity-Based Reasoning

Similarity-based reasoning is a powerful tool to choose and classify available services according to their relevance to a given query and to the contextual information. It can be used for several aims in the context adaptation phase, such as to compare users' profiles and preferences in order to recommend similar services, to match the description of a request with available services, and to compare the current context with already known contexts.

In a recent survey, Guessoum et al. [61] explore the pervasive computing environments where the similarity-based reasoning has been applied in the literature and among them they include the service discovery and service recommendation. Moreover, they list also various kinds of similarity measures used in these environments that are summarised as follows: Pearson coefficient of correlation [62], Cosine method [63], Euclidean distance [64], and feature-based semantic similarity measures [65, 66].

Considering the surveyed context-aware LBSs, three of them (that are iWISE, D'Ulizia et al., and Bao et al.) apply similarity-based reasoning and the following kinds of similarity measures, respectively: social distance and user motion similarity [29], semantic and typed structural similarity [37], and user similarity [31].

Similarity-based techniques have high flexibility, but limited precision and recall.

5.2.2 Collaborative Filtering

Collaborative filtering selects the services to recommend according to how other users, identified to be similar, ranked the services. This approach assumes that users, who had had common interests in the past, tend to have similar tastes in the future [67]. Therefore, users are asked to provide ratings of the services as their feedback, which is used to find other users who have provided similar feedback.

Collaborative filtering techniques can be classified in memory-based and model-based [60]. The former finds users that are similar to the current user and computes the prediction by aggregating the ratings of these users for the same service asked by the current user. The latter provides service recommendation by first developing a model of user ratings, which is generally based on a probabilistic approach that allows computing the expected value of a user prediction, given his/her ratings on other items.

The main advantage of memory-based collaborative filtering is the good accuracy of the predictions compared to model-based techniques. On the contrary, it is time-consuming as it uses the whole database to make a prediction.

Model-based collaborative filtering is faster compared to memory-based because it queries a model instead of the whole dataset of the user ratings. This advantage turns out to be a disadvantage, because using a restricted set of data can make the prediction accuracy worse.

5.2.3 Machine Learning

Reasoning techniques based on machine learning uses data about previous user preferences and activities to first learn a predictive model, which is then used to predict the future services to recommend to the user.

To this aim, several machine learning techniques have been used in the LBS literature, which can be roughly classified in supervised learning techniques (e.g. [29]) and unsupervised learning techniques (e.g. [32]).

Supervised learning requires the training data to be pre-classified (or labelled). Therefore, each training example is associated with a unique label representing the class (i.e. possible services to recommend) in which the item belongs. This means that the classes have to be defined according to the labelling of the training data. Supervised learning techniques include Bayesian networks, Hidden Markov Models, logistic regression, support vector machine, etc. Supervised learning approaches require good training data which is often not easy to obtain.

On the contrary, unsupervised learning methods do not require pre-classification of the training examples. These techniques include artificial neural networks, suffix trees, clustering techniques, etc. Unsupervised learning techniques have the advantage of uncovering unanticipated services. However, this advantage turns out to be a disadvantage, because without any pre-classification these techniques may found a classification of possible services to recommend that is not relevant.

5.2.4 Rule-Based Reasoning

Rule-based reasoning makes use of a domain knowledge and heuristics to define causal relationships between context information and available services through a set of rules and a set of activation conditions for these rules. The rules are generally specified through if–then statements that define the possible contextual

configurations (for instance "if the temperature is between 23 °C and 35 °C and the day is Sunday then provide me a list of seaside resorts that are closer than 100 km from here). They are specified by the system's developer before the operation of the system.

Techniques based on rule-based reasoning are characterized by high precision and recall, but low flexibility.

Rule-based reasoning is generally applied by ontology-based LBS systems as a powerful tool for representing additional attributes that cannot naturally be inferred using traditional ontological models [68]. Considering the surveyed context-aware LBSs, SHERLOCK and GeoSPLIS use rule-based reasoning over their ontological model. Specifically, SHERLOCK uses of a reasoner based on Description Logics (DL) [69] that enables the system to infer information about the objects that a user device discovers and select the most appropriate service providers according to the collected contextual information. GeoSPLIS uses RuleML [70] and Jess [71] compatible rules to model user preferences and available services.

5.3 *Automation of Context Adaptation*

The adaptation of the system's response to the context information can be characterised by various degrees of automation. Schou [72] identifies two kinds of adaptation according to the degree of automation:

1. *self-adaptation*, if the system adapts without any interaction with the user, and
2. *controlled adaptation*, if the user makes decisions and the system automates the change of behaviour.

The main inconvenience of self-adaptation relies in the fact that the user might get a feeling of losing control over the system. On the contrary, controlled adaptation could require too much interaction of the user with the system [73]. These two degree of automation represent the extreme cases. Many context adaptation approaches have been developed as a middle way between them, for instance, by requiring user permission before applying adaptation. We refer to this case as semi-controlled adaptation.

Two terms that are often used to refer to self-adaptation are *adaptability* and *adaptivity*. Adaptability refers to self-adaptation which is based on knowledge (concerning the user, the environment, the context of use, etc.) available to (or, acquired by) the system prior to the initiation of interaction [74]. Adaptivity refers to self-adaptation which is based on knowledge (concerning the user, the environment, the context of use, etc.) that is acquired and/or maintained by the system during interactive session [74].

5.4 Context-Aware LBS Systems Using Big Data: Context Reasoning and Adaptation

In this section, the surveyed LBS systems have been analyzed according to the adaptation strategies, the reasoning techniques, and the degree of automation of context adaptation, defined in the previous sections. The results of this analysis are summarized in Table 9: the second column provides the adaptation strategies applied by each LBS system, the third column provides the reasoning technique, and the fourth column provides the degree of automation of the adaptation process.

The table shows that four surveyed LBS systems (40%) apply personalization as adaptation strategy, while two systems (20%) apply recommendation, and four systems apply both together (40%). Considering reasoning techniques, half of the systems (50%) use collaborative filtering, among which 30% use model-based collaborative filtering and 20% memory-based collaborative filtering. It is interesting to note that all the LBS systems that apply collaborative filtering rely on recommendation strategy. Similarity-based reasoning is applied by 30% of the surveyed LSB systems (all relying on a personalization strategy). Finally, few systems (20%) apply machine learning and rule-based reasoning (all relying on a personalization strategy). Regarding the degree of automation, the majority of the surveyed LBS systems (70%) apply self-adaptation, while the remaining 30% uses a controlled adaptation.

iWISE [29] provides a three layer structure for performing the LBS service discovery. The first layer, called IaaS (Infrastructure as a Service) provides real-time precise positioning technology for collecting location related data. The second layer, called PaaS (Platform as a Service) is responsible for location resource aggregation and management. It extracts location knowledge from Internet multimedia information to associate with corresponding locations. Finally, the third layer, called SaaS (Software as a Service), analyses and process all captured locations and user data for performing location-based social awareness. Specifically, it implements four kinds of algorithms for social awareness: (i) semantic awareness for locations, which includes the computation of social distance between locations; (ii) location-based user relationship awareness, which mines social network relationships and user habits by comparing similarity between user motions; (iii) user mobility awareness that analyses periodic behaviors in user movements and constructs a dynamic Bayesian network for user motion detection and prediction; (iv) location-based social characteristic awareness, which uncovers behavioral patterns from user activities. This social awareness is used to find locations with higher relevance with users' current locations and self-adapting information to deliver to users.

Social Telescope [30] applies recommendation as adaptation strategy and collaborative filtering as reasoning technique. The context aware service discovery, indeed, acts in the following way. The system first computes the set of services based on the matching with the user request. Next, it ranks these services according to the social popularity resulted from the geo-tweets crawled from Twitter. The

Table 9 The surveyed LBS systems and their context adaptation features

Context-aware LBS systems	Adaptation strategy		Reasoning technique				Degree of automation	
	Personalization	Recommendation	Similarity reasoning	Collaborative filtering	Machine learning	Rule-based reasoning	Self-adaptation	Controlled adaptation
iWISE [29]	√		Social distance User motion similarity		Dynamic bayesian networks		√	
Social Telescope [30]		√		Memory-based			√	
Biancalana et al. [32]	√	√			Artificial neural networks			√
SHERLOCK [33]	√					Description logic		√
BOTTARI [34]	√	√		Model-based			√	
Zheng et al. [35]	√	√		Model-based			√	
D'Ulizia et al. [36]	√		- Semantic similarity - Structural similarity				√	
GeoSPLIS [37]	√					Jess rule engine		√
KnockAround [38]		√		Model-based			√	
Bao et al. [39]	√	√	- User similarity	Memory-based			√	

popularity is computed by giving weights to users proportionally to their expertise (that is a function of the number of using that service).

Biancalana et al. [32] apply recommendation as adaptation and an unsupervised machine learning technique based on artificial neural networks as method for filtering the available PoIs. Specifically, the current contextual features and the features of the PoI extracted and stored in the local database are given in input to the context-aware recommendation engine that is based on a feed-forward multi-layer neural network with one hidden layer. This artificial neural network maps the input vector to one of the five classes (from 0 = non interesting to 4 = very interesting) representing how close the PoI is to the user current context. The highly ranked results are put on top of the returned list of PoIs.

SHERLOCK [33] uses mobile agents and description logic (DL) reasoning to represent and manage the knowledge and to guide the user in the process of selecting the LBS that best fits his/her needs. The service discovery process is composed of the following two steps: (i) request generation, which searches in the local ontology for services that can be interesting for the user and helps him/her to generate a request using ontology-guided mechanisms and DL reasoner, and (ii) request processing, which connects to other devices and third-party information providers to retrieve further information interesting for the user request.

BOTTARI [34] provides personalized recommendations of PoIs based on weighted opinions of the social media community. It is based on an ontology-based information integration platform, called LarKC [75], that uses three plug-ins for computing the recommendations: the Sor plug-in orders the available PoIs by distance from the location of the user; the Suns plug-in orders the PoIs by the estimated probability that the user like them considering his/her preferences; the Sld plug-in orders the PoIs by the number of tweets that talk positively in a fixed period. Afterwards, a query evaluator computes the global answer from the three lists outputted by the plug-ins.

Zheng et al. [35] propose a collaborative filtering algorithm to provide personalized recommendations to users. The algorithm uses a ranking-based collective tensor (i.e. a multi-dimensional array) and matrix factorization model to provide personalized recommendations. It formulates the recommendation as a ranking problem and tries to optimize the ranking performance by exploiting information about user-user similarity, location features, activity-activity correlations and user-location visiting preferences.

D'Ulizia et al. [36] address the problem of providing LBSs personalized on the base of the user profile and context. They use a similarity assessment method that combines two types of similarity models, namely semantic and structural typed [76] [77]. The *semantic similarity* model evaluates conceptual similarities between available service names and the requested service name. The *structural typed similarity* model aims at assessing the similarity of the attributes and types of the user, context and service profiles in order to make the user request more precise and selective.

GeoSPLIS [37] provides proactive, personalized and contextualized information to each user by using a rule-based reasoning approach to select and rank the

available PoIs. The user provides their preferences concerning POIs by authoring rules through a web editor. These rules are then translated into RuleML [70] and then into the Jess rule engine so that to be machine understandable. The Jess rule engine takes as input data about user context and nearby PoIs and evaluates user's rules and places' rules using the input data. If the rule is fulfilled than the corresponding place is considered relevant for the user and it is visualized on Google Maps.

KnockAround [38] provides a P2P pull-type LBS system that exploits the location information already uncovered by the surrounding people to recommend PoIs to a user. A specific module, called information retriever, is in charge of searching through the local database, populated by the people during their day-to-day visits, and trying to partially match the user request against the place-name, keywords or comments in the database entries, which fall within the acceptable range of distance. The search results are sorted in ascending order of distance.

Bao et al. [39] propose a location-based preference-aware recommender system that simultaneously considers current user location, user preferences and social opinions for making recommendations of a service. The selection and rating of the possible services is performed by the online recommendation component, which applies both similarity reasoning and collaborative filtering to this aim. This component first selects the candidate local experts and services in the user specified spatial range that best fit user preferences. Secondly, it computes similarity scores between the user and the selected local experts. These similarity scores are inputted to a collaborative filtering algorithm that infers the rating that the user would give to the candidate services. The services with high predict ratings are recommended to the user.

6 Open Challenges of Context-Aware LBSs

In the previous sections, we have analysed how several recent context-aware LBS systems have answered to the main issues occurring during context acquisition, context representation, and context reasoning and adaptation. However, many open challenges still remain to be solved. In this section, we briefly discuss the open challenges related to the three main phases of the context-awareness process, and resulted from both the literature and the analysis of the surveyed LBS systems.

6.1 Challenges in Context Acquisition

The privacy is one of the main open challenges that need to be considered when capturing and using contextual knowledge in context-aware LBSs. It is important that LBS users can control their personal information and that the LBS system asks

for authorization to collect them. Many endeavours have been made in fostering the privacy in context-aware LBSs [78, 79]. Further work is necessary to integrate effectively privacy management within the service discovery process to enhance LBS users' privacy.

Another open challenge in context acquisition is the presence of missing values in the contextual data acquired by sensors due to possible inefficiencies in sensor hardware and unstable network communication. This fact causes a loss in the accuracy of personalized service discovery. To solve this problem, acquired contextual data need to be cleaned by filling missing values, removing outliers, validating context via multiple sources, and many more [80].

Further challenges may be deducted from the results of our analysis. As discussed in Sect. 3.4 (see Table 5), many existing LBS systems do not yet acquire all kinds of context categories and very few systems use temporal and social context to recommend personalised services to users. Moreover, the use of sensors is mainly restricted to GPS and networks of P2P devices, despite of the availability of further types of static, wearable and mobile sensors (e.g. accelerometer, environmental sensors, healthcare sensors, etc.). Therefore, an open challenge is the exploration of new techniques to capture easily all kinds of contextual data (spatial, physical, personal, technical, temporal, and social) from different types of sensors (static, mobile, and wearable), as well as Web-based information services, in order to improve context-awareness process in LBS systems.

6.2 Challenges in Context Representation

The main challenge of context representation is the lack of a standard representation for context. Despite the definition of various context representation models (see Sect. 4), current LBS systems rely on specific representations of contextual data. This lack in standardization restricts the portability, sharing and re-use of contextual data across different systems. Therefore, future work on context representation should address the definition of a standardized representation as well as a mapping from existing context representation models to this standard.

Moreover, results of the evaluation of context models used by the surveyed LBS systems, provided in Sect. 4.3, have shown that there is a lack of LBS systems that meet all the requirements of mobility, heterogeneity, timeliness, relationships and dependencies, and reasoning (as proposed by Bettini et al. [41]). Therefore, future research should investigate how to define a standardized context representation able to fulfill all these requirements.

6.3 Challenges in Context Reasoning and Adaptation

Context reasoning and adaptation has some open challenges concerning the personalization and the recommendation strategies.

Context-aware LBS systems relying on personalization have to face three major challenges in the next years that are the privacy, accuracy of personalized results, and accessibility.

The contextual information that drives personalization is normally based upon user's personal information and that gives rise to the privacy challenge. It is necessary, indeed, that the user gives his/her consensus with respect to what personal information to share with the surrounding services in order to encompass the service provisioning based on personal information. In this direction, as said before, further efforts are necessary to integrate effectively privacy management within the service discovery process.

The more available is the contextual information, the more reliable will be the personalization of services to be returned to the user. Hence, there is a clear trade-off between the accuracy of the personalization and the privacy of users' data [81]. Many endeavours have been made in enhancing the accuracy of personalization preserving user's privacy [81].

Accessibility is a further open challenge that is strictly connected with the availability of user-friendly interfaces for LBS systems. Improving accessibility can be realized, indeed, by designing user interfaces in such a way to ensure an easy interaction with the system and to enable the request and delivering of service information through customized multimedia and multimodal channels [82–85]. Future work in this field should explore how to adapt the interfaces of LBSs according to the user dynamic interactive behavior [86, 87].

Therefore, future work on personalized context-aware LBSs should investigate new adaptivity solutions characterized by more accessibility, accuracy and privacy preservation.

Context-aware LBS systems relying on recommendation face two major challenges that are the cold start problem and the data sparseness.

The cold start problem occurs when new users or new services are entered in the system and, therefore, the system does not have enough information (e.g., ratings, browsing history, etc.) to provide the new user with accurate recommendations or to reliably recommend the new service to any user [88]. As having new users or new services is a common situation for location-based recommender systems, the cold start problem is an interesting research challenge that recent works are trying to solve [88, 89].

Data sparseness arises from the fact that users generally rate only a limited number of services and, therefore, the user-service matrix, containing the ratings of services for each user, is generally very sparse. This fact influences the accuracy of recommendations because the system does not allow accurately predicting the user

interest in a service using such a sparse matrix. Current research is investigating how to enhance the accuracy of recommendations by using additional information, such as user generated content and social relationships [90–92].

7 Conclusion

In this survey, we discussed the problem of context-awareness in location-based services and gave an overview of the existing context-aware LBS systems. In particular, we discussed the main steps of the context-awareness process for LBS systems, i.e. context acquisition, context representation, and context reasoning and adaptation, and we reviewed and evaluated some existing LBS systems according to these three perspectives. Finally, we have presented some open challenges and future directions of this research field.

From this overview, the following conclusions can be drawn:

- all the surveyed LBS systems acquires spatial context mainly from GPS embedded into the user device. Most of the systems (60%) use personal context for personalizing LBSs both acquiring it from social media and analyzing GPS data history. Technical context has been effectively employed by 50% of the surveyed LBS systems. 30% of the surveyed LBS systems use physical context acquired from web information services, social media and sensors. Finally, few systems (20%) apply temporal context acquired from GPS and social context acquired mainly from social media. All the surveyed LBS systems acquire context in an automatic way.
- most of the surveyed context-aware LBS systems (60%) use key-value model for representing context information. Ontologies have been effectively employed by 40% of the surveyed LBS systems. This is due mainly to the main simplicity to implement and manage context information and the time-consuming and costly creation of ontologies. However, ontology-based models outperform key-value models in all the evaluated requirements (mobility, heterogeneity, relationships and dependencies, timeliness, and reasoning). Therefore, the simplicity of key-value models win over the better performance of ontology-based models.
- concerning the context reasoning and adaptation, most of the surveyed LBS systems (80%) apply either personalization or personalization and recommendation both together, and collaborative filtering is the most used reasoning technique applied by half of the systems. Moreover, the majority of the surveyed LBS systems (70%) apply self-adaptation.

References

1. Xi, W., Han, J., Li, K., Jiang, Z., Ding, H.: Location inferring in internet of things and big data. In: Buyya, R., Calheiros, R.N., Dastjerdi, A.V. (eds.) Big Data: Principles and Paradigms, pp. 309–335. Morgan Kaufmann (2016)
2. Koeppel, I.: What are Location Services? From a GIS Perspective. ESRI white paper (2000)
3. Shiode, N., Li, C., Batty, M., Longley, P., Maguire, D.: The impact and penetration of location-based services. In: Karimi, H.A., Hammad, A. (eds.), Telegeoinformatics: Location-Based Computing and Services, pp. 349–366. CRC Press (2004)
4. Spiekermann, S.: General aspects of location-based services. In: Schiller, J., Voisard, A. (eds). Location-Based Services. Morgan Kaufman (2004)
5. Shek, S.: Next Generation Location Based Services for Mobile Devices, pp. 1–66. Computer Science Corporation, Leading Edge Forum (2010)
6. Hadjioannou, V., Mavromoustakis, C.X., Papanikolaou, K., Mastorakis, G., Goleva, R., Dobre, C., Batalla, J.M.: On the comparison of location based software solutions used for tracking purposes in ambient assisted living applications. In: 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 5–11. IEEE, Sept 2016
7. Skourletopoulos, G., Mavromoustakis, C.X., Mastorakis, G., Batalla, J.M., Sahalos, J.N.: An evaluation of cloud-based mobile services with limited capacity: a linear approach. Soft Comput. 1–8 (2016)
8. Bourdena, A., Mavromoustakis, C.X., Mastorakis, G., Rodrigues, J.J.P.C., Dobre, C.: Using socio-spatial context in mobile cloud process offloading for energy conservation in wireless devices. IEEE Trans. Cloud Comput. **1** (2015)
9. Virrantaus, K., Markkula, J., Garmash, A., Terziyan, Y.V.: Developing GIS-supported location-based services, 423–432. In: Proceedings of WGIS'2001–First International Workshop on Web Geographical Information Systems, Kyoto, Japan (2001)
10. Schiller, J., Voisard, A.: Location Based Services. Morgan Kaufmann, San Francisco, CA (2004)
11. Reichenbacher, T.: Mobile Cartography: Adaptive Visualisation of Geographic Information on Mobile Devices (PhD thesis) (2004). https://www.tumb1.biblio.tu-muenchen.de/pub1/diss/bv/2004/reichenbacher.pdf
12. Themistocleous, M., Azab, N.A., Kamal, M.M., Ali, M., Morabito, V.: Location-based services for public policy making: The direct and indirect way to e-participation. Inf. Syst. Manag. **29**(4), 269–283 (2012)
13. Malik, N., Mahmud, U., Javed, Y.: Future challenges in context-aware computing. In: proceedings of the IADIS International Conference WWW/Internet, pp. 306–310 (2007)
14. Khattak, A.M., Akbar, N., Aazam, M., Ali, T., Khan, A.M., Jeon, S., Lee, S.: Context representation and fusion: advancements and opportunities. Sensors **14**(6), 9628–9668 (2014)
15. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D.: Context aware computing for the internet of things: a survey. Commun. Surv. Tutor. IEEE **16**(1), 414–454 (2014)
16. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: Proceedings of the First Workshop Mobile Computing Systems and Applications (WMCSA '94), pp. 85–90 (1994)
17. Chen, G., Kotz, D.: A survey of context-aware mobile computing research. Technical Report TR2000–381, Department of Computer Science, Dartmouth College (2000)
18. Zimmermann, A., Lorenz, A., Oppermann, R.: An operational definition of context. In: Proceedings of the 6th International and Interdisciplinary Conference on Modeling and using Context (CONTEXT07), pp. 558–571. Springer Press (2007)
19. Vieira, V., Tedesco, P., Salgado, A.C.: Towards an Ontology for Context Representation in Groupware. Proceedings of the International Workshop on Groupware (CRIWG'05); Porto de Galinhas, Brazil, pp. 367–375, 25–29 Sept 2005

20. Nieto, I., Bota, J.A., Gómez-Skarmeta, A.F.: Information and hybrid architecture model of the OCP contextual information management system. J. Univ. Comput. Sci. **12**, 357–366 (2006)
21. Hervás, R., Bravo, J., Fontecha, J.: A context model based on ontological languages: a proposal for information visualization. J. Univ. Comput. Sci. **16**, 1539–1555 (2010)
22. Köpper, A.: Location Based Service, Fundamental and Operation. England, Chichester (2005)
23. Tu, Y.: xTolk-A Context-Aware Mobile Application on the Nokia N95 8 GB Smartphone (Doctoral dissertation, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark) (2008)
24. Bellavista, P., Corradi, A., Fanelli, M., Foschini, L.: A survey of context data distribution for mobile ubiquitous systems. ACM Comput. Surv. (CSUR) **44**(4), 24 (2012)
25. Zhang, D., Yu, Z., Guo, B., Wang, Z.: Exploiting personal and community context in mobile social networks. In: Mobile Social Networking, pp. 109–138. Springer, New York (2014)
26. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. Bus. Horiz. **53**(1), 59–68 (2010)
27. Dasgupta, R., Chattopadhyay, D., Pal, A., Chakravarty, T. (2014). A comprehensive seven layer sensor model: cyber-physical system. In: Sensing Technology: Current Status and Future Trends I, pp. 57–81. Springer International Publishing
28. Schwinger, W., Grün, C., Pröll, B., Retschitzegger, W., Schauerhuber, A.: Context-awareness in mobile tourism guides–a comprehensive survey. Rapport Technique. Johannes Kepler University Linz (2005)
29. Guo, C., Liu, J., Fang, Y., Wan, Y., Cui, J.: iWISE: A location-based service cloud computing system with content aggregation and social awareness. In: Principle and Application Progress in Location-Based Services, pp. 139–157. Springer International Publishing (2014)
30. Shankar, P., Huang, Y.W., Castro, P., Nath, B., Iftode, L.: Crowds replace experts: Building better location-based services using mobile social network interactions. In: 2012 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 20–29. IEEE (2012)
31. Foursquare. https://www.it.foursquare.com/
32. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G. : Social tagging for personalized location-based services. In: Proceedings of the 2nd International Workshop on Social Recommender Systems (2011)
33. Yus, R., Mena, E., Ilarri, S., Illarramendi, A.: SHERLOCK: semantic management of location-based services in wireless environments. Pervasive Mob. Comput. **15**, 87–99 (2014)
34. Balduini, M., Celino, I., Dell'Aglio, D., Della Valle, E., Huang, Y., Lee, T., Tresp, V.: BOTTARI: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. Web Semant. Sci. Serv. Agents World Wide Web **16**, 33–41 (2012)
35. Zheng, V.W., Zheng, Y., Xie, X., Yang, Q.: Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. Artif. Intell. **184**, 17–37 (2012)
36. D'Ulizia, A., Ferri, F., Grifoni, P.: A similarity assessment method for discovering and adapting business services. Int. J. Comput. Sci. Eng. **5**(2), 97–109 (2010)
37. Viktoratos, I., Tsadiras, A., Bassiliades, N.: Providing a context-aware location based web service through semantics and user-defined rules. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics, p. 9. ACM (2014)
38. Rizia, R., Tanviruzzaman, M., Ahamed, S.I.: KnockAround: location based service via social knowledge. In: 2012 IEEE 36th Annual Computer Software and Applications Conference (COMPSAC), pp. 623–631. IEEE (2012)
39. Bao, J., Zheng, Y., Mokbel, M.F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In: Proceedings of the 20th International Conference on Advances in GIS, pp. 199–208. ACM (2012)
40. Strang, T., Linnhoff-Popien, C.: A context modeling survey. In Workshop Proceedings, Sept 2004

41. Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., Riboni, D.: A survey of context modelling and reasoning techniques. Pervasive Mob. Comput. **6**(2), 161–180 (2010)
42. Pop, F., Cristea, V.: The art of scheduling for big data science. In: Li, K.-C., Jiang, H., Yang, L.T., Cuzzocrea, A. (eds.) Big Data: Algorithms, Analytics, and Applications, pp. 105–120. Chapman & Hall/CRC Big Data Series (2015). ISBN 978-1482240559
43. Bray, T., Paoli, J., Sperberg-McQueen, C.M.: Extensible Markup Language (XML). Present World Wide Web J. 27–66 (1997)
44. Schmohl, R., Baumgarten, U.: The contextual map-a context model for detecting affinity between contexts. In: Mobile Wireless Middleware, Operating Systems, and Applications, pp. 171–184. Springer, Berlin (2009)
45. Ye, J., Coyle, L., Dobson, S., Nixon, P.: Ontology-based models in pervasive computing systems. Knowl. Eng. Rev. **22**(04), 315–347 (2007)
46. Brickley, D., Guha, R.: Resource Description Framework (RDF) Schema Specification (2000). https://www.w3.org/TR/RDF-schema
47. OWL Web Ontology Language. https://www.w3.org/TR/owl-ref/
48. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Jagadish, H.V.: Challenges and Opportunities with Big Data. A Community White Paper Developed by Leading Researchers Across the United States. Computing Research Association, Washington (2012)
49. Berrueta, D., et al.: SIOC core ontology specification. W3C Member Submission (2007)
50. http://www.w3.org/2003/01/geo/
51. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: an architecture for storing and querying RDF data and schema information. In: Lieberman, H., Fensel, D., Hendler, J., Wahlster, W. (eds.) Semantics for the WWW. MIT Press (2001)
52. Klein, C., Schmid, R., Leuxner, C., Sitou, W., Spanfelner, B.: A survey of context adaptation in autonomic computing. In: Fourth International Conference on Autonomic and Autonomous Systems, 2008. ICAS 2008, pp. 106–111. IEEE (2008)
53. Benazzouz, Y.: Context discovery for autonomic service adaptation in intelligent space. Adv. Next Gener. Serv. Serv. Archit. **14**, 281 (2011)
54. Mokbel, M., Bao, J., Eldawy, A., Levandoski, J., Sarwat, M.: Personalization, socialization, and recommendations in location-based services 2.0. In: PersDB 2011 Workshop, 2 Sept 2011, Seattle, Washington, USA (2011)
55. Zimmermann, A., Specht, M., Lorenz, A.: Personalization and context management. User Model User Adapt. **15**(3–4), 275–302 (2005)
56. Chaudhuri, S., Gravano, L.: Evaluating Top-K selection queries. In: VLDB (1999)
57. Borzsonyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: ICDE (2001)
58. Chan, C.-Y., Jagadish, H., Tan, K.-L., Tung, A.K., Zhang, Z.: Finding k-Dominant skylines in high dimensional space. In: SIGMOD (2006)
59. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. In: SIGMOD (1995)
60. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 4 (2009)
61. Guessoum, D., Miraoui, M., Tadj, C.: Survey of semantic similarity measures in pervasive computing. Int. J. Smart Sens. Intell. Syst. **8**(1), 125–158 (2015)
62. Chen, A.: Context-Aware collaborative filtering system: predicting the user's preference in the ubiquitous computing environment. Location- and Context-Awareness, pp. 244–253. Springer, Berlin (2005)
63. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
64. Liu, L., Lecue, F., Mehandjiev, N., Xu, L.: Using context similarity for service recommendation. In: 2010 IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 277–284. IEEE (2010)

65. García-Crespo, A., Chamizo, J., Rivera, I., Mencke, M., Colomo-Palacios, R., Gómez-Berbís, J.M.: SPETA: social pervasive e-tourism advisor. Telemat. Inf. **26**(3), 306–315 (2009)
66. Grifoni, P., D'Ulizia, A., Ferri, F.: A semantic-based approach for context-aware service discovery. Int. J. Inf. Syst. Serv. Sect. (IJISSS) **6**(4), 1–26 (2014)
67. Anand, S.S., Mobasher, B.: Intelligent Techniques for Web Personalization, pp. 1–36. Springer, Berlin (2005)
68. Wang, X.H., Zhang, D.Q., Gu, T., Pung, H.K.: Ontology based context modeling and reasoning using OWL. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004, pp. 18–22. IEEE (2004)
69. Baader, F., Horrocks, I., Sattler, U.: Description logics as ontology languages for the semantic web. In: Mechanizing Mathematical Reasoning, in: Lecture Notes in Computer Science, vol. 2605, pp. 228–248. Springer (2005)
70. Boley, H., Paschke, A., Shafiq, O.: RuleML 1.0: the overarching specification of web rules. Lect. Notes Comput. Sci. **6403**(4), 162–178 (2010)
71. Friedman-Hill, E.: Jess, the Rule Engine for the Java Platform (2008)
72. Schou, S.: Context-based service adaptation platform: improving the user experience towards mobile location services. In: International Conference on Information Networking, 2008. ICOIN 2008, pp. 1–5. IEEE, Jan 2008
73. Reichenbacher, T.: Mobile Cartography—Adaptive Visualisation of Geographic Information on Mobile Devices. Verlag Dr. Hut, München (2004)
74. Stephanidis, C., Paramythis, A., Akoumianakis, D., Sfyrakis, M.: Self-adapting web-based systems: towards universal accessibility. In: 4th Workshop on User Interface For All, Stockholm, Sweden (1998)
75. Cheptsov, A., et al.: Large knowledge collider. A service-oriented platform for large-scale semantic reasoning. In: Proceedings of WIMS 2011 (2011)
76. D'Ulizia, A., Ferri F., Formica, A., Grifoni, P.: Approximating geographical queries. J. Comput. Sci. Technol. **24**(6), 1109–1124, Nov 2009
77. D'Ulizia, A., Ferri, F., Grifoni, P., Rafanelli, M.: Relaxing constraints on GeoPQL operators for improving query answering. In: 17th International Conference on Database and Expert Systems Applications (DEXA'06), Lecture Notes in Computer Science 4080, pp 728–737. Springer, (2006)
78. Shin, K.G., Ju, X., Chen, Z., Hu, X.: Privacy protection for users of location-based services. Wirel. Commun. IEEE **19**(1), 30–39 (2012)
79. Pan, J., Zuo, Z., Xu, Z., Jin, Q.: Privacy protection for LBS in mobile environments: progresses, issues and challenges. Int. J. Secur. Its Appl. **9**(1), 249–258 (2015)
80. Zelenik, D., Bielikova, M.: Reducing the sparsity of contextual information for recommender systems. In: Proceedings of the sixth ACM conference on Recommender systems, pp. 341–344. ACM (2012)
81. Berkovsky, S., Eytani, Y., Kuflik, T., Ricci, F.: Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In: Proceedings of the 2007 ACM Conference on Recommender Systems, pp. 9–16. ACM, Oct 2007
82. Grifoni, P., Ferri, F., Caschera, M.C., D'Ulizia, A., Mazzei, M.: MIS: Multimodal interaction services in a cloud perspective. J. Next Gener. Inf. Technol. **5**(4), 1 (2014)
83. D'Ulizia, A.: Exploring multimodal input fusion strategies. The Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility, pp. 34–57 (2009)
84. D'Andrea, A., D'Ulizia, A., Ferri, F., Grifoni, P.: EMAG: An Extended Multimodal Attribute Grammar for Behavioural Features, Digital Scholarship in the Humanities, vol. 32(2), pp. 251–275. Oxford University Press (2017)
85. D'Ulizia, A., Ferri, F., Grifoni, P.: Moving GeoPQL: a Pictorial Language towards Spatio-Temporal Queries. GeoInformatica **16**(2), 357–389 (2012)
86. Feng, J., Liu, Y.: Intelligent context-aware and adaptive interface for mobile LBS. Comput. Intell.Neurosci. (2015)

87. Grifoni, P., D'Ulizia, A., Ferri, F.: Computational methods and grammars in language evolution: a survey. Artif. Intell. Rev. **45**(3), 369–403 (2016)
88. Braunhofer, M.: Hybrid solution of the cold-start problem in context-aware recommender systems. In: User Modeling, Adaptation, and Personalization, pp. 484–489. Springer International Publishing (2014)
89. Lika, B., Kolomvatsos, K., Hadjiefthymiades, S.: Facing the cold start problem in recommender systems. Expert Syst. Appl. **41**(4), 2065–2073 (2014)
90. Huang, L.W., Chen, G.S., Liu, Y.C., Li, D.Y.: Enhancing recommender systems by incorporating social information. J. Zhejiang Univ. Sci. C **14**(9), 711–721 (2013)
91. Than, C., Han, S.: Improving recommender systems by incorporating similarity, trust and reputation. J. Internet Serv. Inf. Secur. (JISIS) **4**(1), 64–76 (2014)
92. Ferri, F., Grifoni, P., Caschera, M.C., D'Ulizia, A., Praticò, C.: KRC: KnowInG crowdsourcing platform supporting creativity and innovation. Adv. Inf. Sci. Serv. Sci. **5** (16), 1–15 (2013)

Mobile Big Data in Vehicular Networks: The Road to Internet of Vehicles

Ali Kamouch, Abdelaali Chaoub and Zouhair Guennoun

Abstract In the emerging 5G communication systems, the need for advanced data handling technologies will be more crucial than ever. In addition, handling data generated by Internet of Things (IoT) is a promising challenge for both scientists and business. Internet of Vehicles (IoV) is a key member of the Internet of Things (IoT) family to improve road safety and improve driving experience. In this vision, collecting and processing Big data generated by vehicles is a real challenge in the specific context of vehicles. Big data means that data cannot be handled by conventional information systems. The volume of data in the Big data era is such that it cannot be loaded into a single machine. It also implies that most traditional methods of data mining and data analysis developed for centralized architectures will not be applicable. In this context, this chapter discuss the interaction of Internet of vehicles and Big data technologies. First, we present the evolution of IoV and its features. Second, we discuss the data life cycle and big data challenges in vehicular context. Finally, IoV Big data and data model are discussed.

1 Introduction

Recent years have seen the emergence of the concept of smart cities as a promising type of urban development aimed at improving the quality of services and reducing their costs. Smart cities make use of new information and communication technologies to meet the increasing social, economic and environmental needs of modern

A. Kamouch · Z. Guennoun
Laboratory of Electronic and Communication, Mohammadia School of Engineers,
Mohammed V-Agdal University, Rabat, Morocco
e-mail: kamouch.ali@gmail.com

Z. Guennoun
e-mail: zouhair@emi.ac.ma

A. Chaoub (✉)
Communications Systems Department, National Institute
of Posts and Telecommunications, Rabat, Morocco
e-mail: chaoub@inpt.ac.ma

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_6

cities. In smart cities, the object would have the capability to connect with each other through heterogeneous networks, which is also termed as Internet of thing (IoT). IoT is a system of interconnected objects equipped with the capabilities of communicating and transferring data through wired or wireless networks without requiring man-to-computer interactions. According to a study by the European Commission, IoT is a "pervasive innovative technology building on the universal connectivity of things and people" [1]. IoT is an emerging technology that promises to radically change our vision of the Internet by providing the objects of everyday life with communication and computing abilities. Allowing these devices to generate, exchange and process data with minimal human intervention. IoT aims to give another dimension to the conventional Internet towards a hyperconnected world and smart cities by expanding adoption of IP-based communications together with ubiquitous connectivity through interoperability between different existing communication systems.

Internet of Vehicle (IoV) is a member of the IoT family in which smart vehicles will play the role of nodes. Within the Context of the IoV, the vehicles will be equipped with several sensors to retrieve speed, position and engines conditions. Personal devices will also be used for collection of data related to the use of applications. In this chapter we will decorticate the important elements to manage IoV Big Data. We first present notions relating to IoT. Then, we will discuss evolution of vehicular networks from vehicular Adhoc to the internet of vehicles. Finally, we will detail the architecture and the management models of IoV Big data.

2 Internet of Things Communication Models

IoT allows to link intelligent objects with different protocol stack, security features and access technologies. Therefore, it would be useful to see how these objects can connect with each other and with services hosted in the cloud. RFC 7452 defines four communication models used in IoT [2]. In this part we describe these models and discuss their key characteristics.

2.1 Device-to-Device Pattern

Device-to-Device pattern describes a direct communication between two devices coming from different manufacturers (Fig. 1). Various aspects of the protocol stack must be designed to ensure this communication: physical medium technology, supported IP version, IP address assignment mechanism, communication architecture model (peer-to-peer or client-server), Service discovery mechanism, transport protocol and application layer. To avoid redundant development efforts, an open approach must be adopted to meet the security and privacy requirements. It includes security threats definition, services to be deployed to deal with the defined threats, the layer

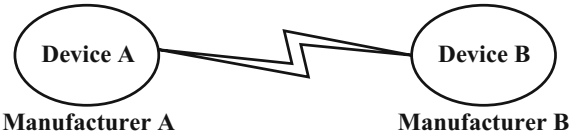


Fig. 1 Device-to-Device communication pattern [2]

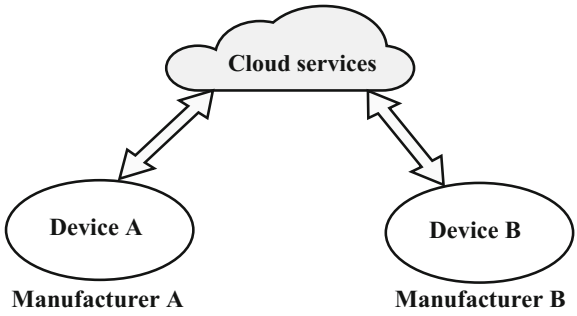


Fig. 2 Device-to-Cloud communication pattern [2]

on which these services are deployed, and implications that design decisions will have on the privacy preservation.

2.2 Device-to-Cloud Communication Pattern

Device-to-Cloud model allows a device to connect to a cloud service of the same manufacturer using wired or wireless traditional access technologies or even cellular networks. It also allows to connect to the smart objects of the same cloud provider without any concern of interoperability, since the entire communication takes place at the level of service provider. However, the use of standardized protocols and mechanisms greatly reduces the cost of designing, implementing and verifying smart objects. To integrate an object from a third party, the cloud interface must be available and different standards can be used, such as, Constrained application (CoAP), Datagram Transport Layer Security (DTLS), UDP, IP, etc., as shown in Fig. 2. Moreover, dependence on a single service provider can make hardware unusable elsewhere. The device manufacturers must make available the source code of the object and/or allow the installation of a new IoT operating system on their devices.

2.3 Device-to-Gateway Communication Pattern

In the Device-to-Gateway model, smart objects connect to cloud services through a gateway that serves as an intermediary between the objects and the service provider. The gateway made it possible to disregard the access technology used in devices.

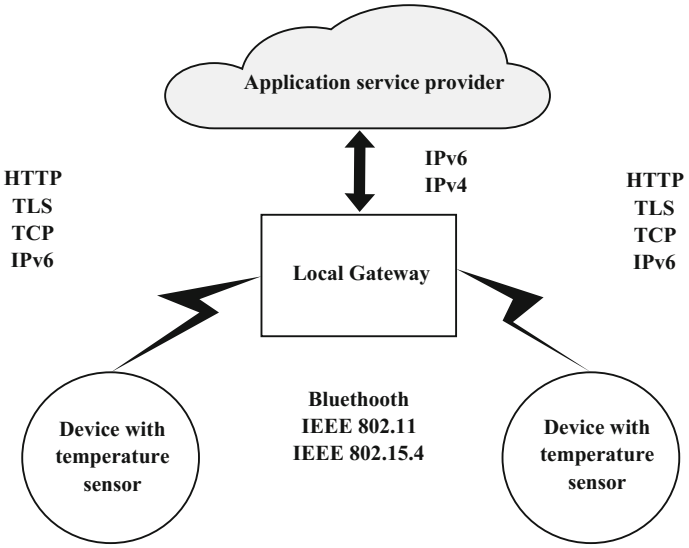


Fig. 3 Device-to-Gateway communication pattern [2]

It also provides security features and translation capabilities for data and protocols. This method of abstraction allows the integration of objects that use less-widely-available access technologies, such as Low Rate Wireless Personal Area Networks (LR WPAN) IEEE 802.15.4, or specific authentication mechanisms. However, this can increase the complexity and cost of installations for end users. To reduce this complexity, efforts are made by intelligent object manufacturers to develop generic gateways using generic internet protocols (Fig. 3).

2.4 Back-End Data Sharing Pattern

The Back-End Data Sharing Pattern (Fig. 4) is an extension of device-to-cloud communication model. It allows a smart object to be connected to multiple vendors at the same time. This also, makes end users independent from the cloud service provider. It also offers the ability to process data in combination with other information from sources connected to third-party providers. To this end, a federated approach is needed for authentication and authorization technologies (like OAuth 2.0). it is also important to develop cloud applications programmer interfaces (APIs) are needed to achieve interoperability of smart device data hosted in the cloud. However, despite the use of standard protocols, this model often leads to data silos and therefore can not completely overcome closed systems [3].

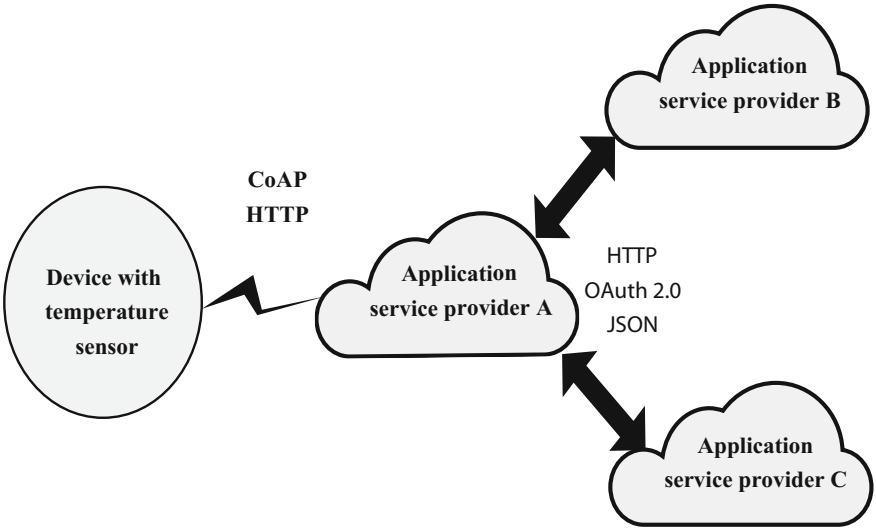


Fig. 4 Back-End data sharing pattern [2]

3 From Vehicular Networks to Internet of Vehicles

3.1 Vehicular Adhoc Networks Evolution

3.1.1 Evolution of the Automotive Market and Its Impact

The number of vehicles has exploded in recent years to reach a billion vehicles worldwide [4]. This number is expected to increase in the coming years to reach 25 billions by 2030 (Fig. 5). This leads to longer congestion and longer waiting periods in addition to a higher number of traffic accidents affecting the populations quality of life. Beyond the direct costs associated with additional wasted time on roads and mobility reduction, congestion imposes other costs such as unreliability cost, vehicle operating cost and emission cost. According to a study by US Department of Transportation, the annual cost of congestion in urban roads is estimated to $85 billions in 2009 (Fig. 6).

3.1.2 Vehicular Adhoc Networks

The development of vehicular networks can provide solutions for problems of congestion and inefficient traffic. Mainly by allowing vehicles to exchange alerts relating to potential road hazards. vehicular Adhoc networks (VANETs) have been introduced to enable vehicles to exchange information without the need for massive infrastructure deployment along roads. WIFI (802.11x) based technologies were the

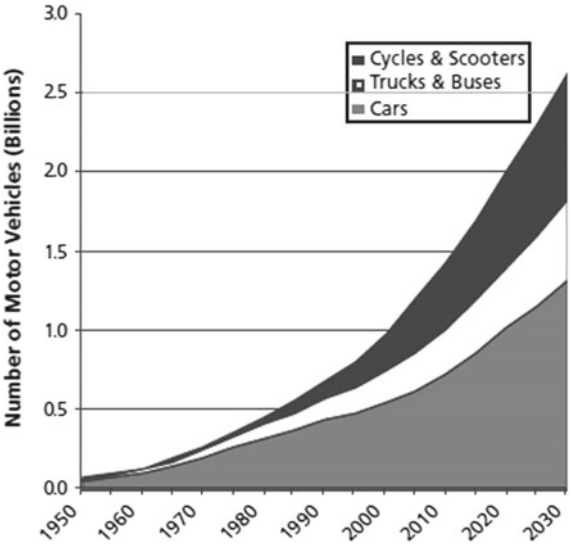


Fig. 5 Evolution of the global vehicles population [4]

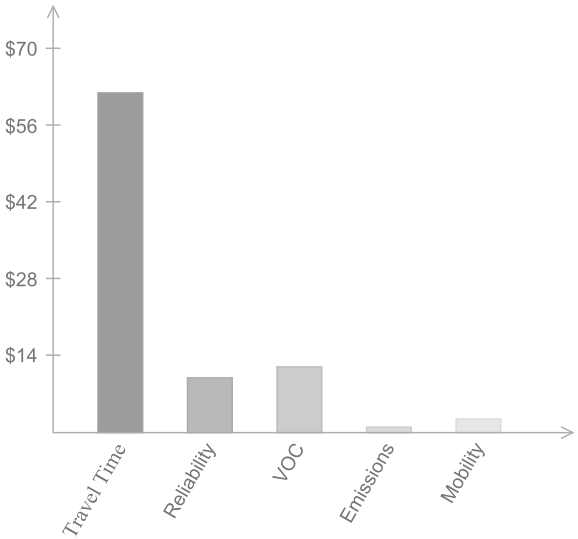


Fig. 6 Annual congestion cost in USA (in $ billion) [5]

most commonly used to support car-to-car communications in an Adhoc fashion. This technology quickly proved to be unsuitable for the very dynamic context of vehicular networks. Currently, Dedicated Short-range Communication (DSRC) is adopted as the standard VANET system. Development of vehicular networks was conducted under the general theme of Intelligent Transport Systems (ITS). Many efforts have been made by academia and industry to achieve a mature standard for deploying ITS services.

Two wireless devices are used in a DSRC system: On Board Units (OBU) and Road Side Units RSU). Vehicle-to-Roadside (V2R) is ensured by OBU and RSU while direct vehicle-to-vehicle (V2V) is achieved via OBU. The direct communication between vehicles reduces latency which allow a fast propagation of security warning. In addition, the GeoNetworking based on the vehicle position allows an efficient dissemination of the information along the road [6].

DSRC applications use a frequency band of 75 MHz located around the frequency 5.9 GHz. The frequency allocation in DSRC is shown in Fig. 7. DSRC frequency band consists of a guard band of 5 MHz and seven channels of 10 MHz each. Channel 178 is a control channel (CCH) used for transmission of short high priority messages and management data. The other six service channels (SCH) are used for the transmission of other data [7]. Three familly of standards are involved in DSRC/WAVE standard. Namely: IEEE 1609 standards which define the communication services. The standard SAE J2735 DSRC which implement the application level needed to exchange messages. In addition the IEEE 802.11p, which is a modification of IEEE 802.11 define the wireless access for WAVE/DSRC to support ITS applications [8]. DSRC/WAVE protocol stack is shown in Fig. 8. WAVE Physical layer is defined by the standard IEEE 802.11p, while IEEE 1609.4 multi-channel and IEEE 802.11p define the upper and lower MAC layers respectively. The logical layer control (LLC) is defined by the standard IEEE 802.2.

Despite the potential interest of the VANET networks in improving the driving experience and preventing road casualties, this technology had little commercial penetration. The government and manufacturers remains skeptical about the effectiveness of the investments to be put in place for the deployment of this technology. In fact, only basic versions of VANETs have been deployed in developed countries such as USA and Japan [9]. This underscores the importance of designing a more reliable and market-oriented vehicular network. In addition to providing communication capability to vehicles, IoV must be able to have a great market penetration due to the opportunities assessed for this upcoming technology.

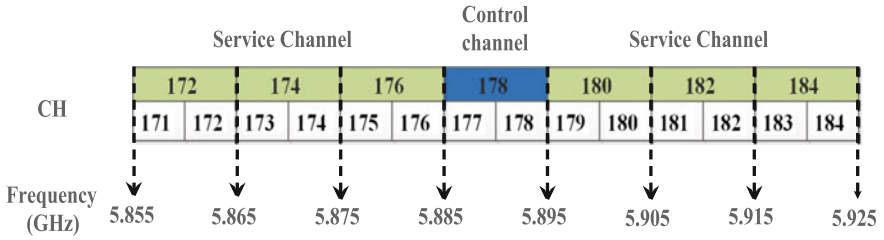


Fig. 7 DSRC frequency allocation

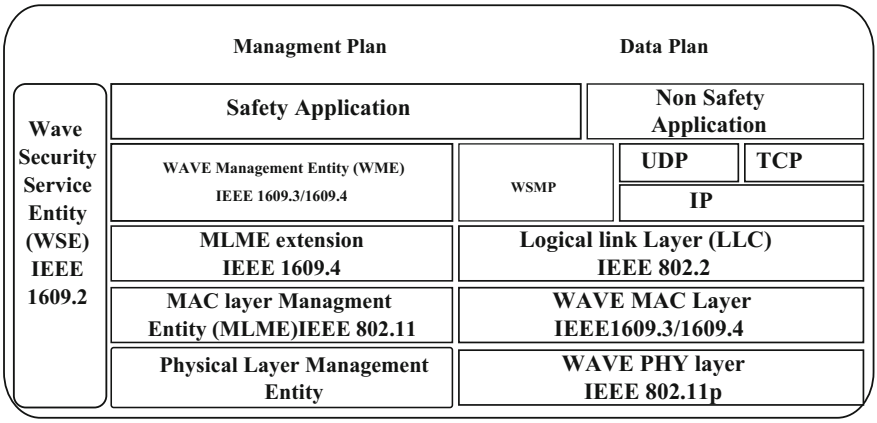


Fig. 8 DSRC protocol stack

3.2 Internet of Vehicles

3.2.1 Motivation

IoV is the application of IoT in the field of transportation. It aims to provide vehicles with communication capabilities without the need of a man-to-computer interaction. It also enable gathering and processing information on vehicles and surrounds. This information is collected from connected "Things" such as radio frequency identification (RFID), Global Positioning System (GPS), sensors and any other connected device. This emerging technology, part of the 5th Generation, will allow a real-time exchange of information on urban transport through internet technology. The processing of such information will contribute in reducing crashes, congestion costs and greenhouse gas emissions and improving driving safety. From a network perspective, IoV can be conceived as a dynamic system allowing Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Vehicle-to-Cloud (V2C) and Vehicle-to-Human (V2H communications. The processing of retrieved big data will provide an efficient supervision of Vehicle nodes and provide services and applications from the public internet network and cloud platforms.

Various motivations have stimulated the search for a reliable communication system to improve road safety and exploit the commercialization opportunities expected for this segment. VANETs are encountering difficulties to penetrate in the market despite the projected opportunities. In addition, the number of road casualties has continued to increase according to the reports of the World Health Organization [10]

Among the major limitations of VANETs networks we can cite:

- Disruptions in vehicular communications and intermittent communication losses make it impossible for the VANET networks to provide the reliability required to support ITS applications without interruption [11].

- Internet connectivity is a prerequisite for many commercial applications. However, this can only be ensured in the ubiquitous RSU connection. Unfortunately, such full coverage is not practical or very expensive in terms of deployment and maintenance [12].
- Dependency on network users is a major concern that makes the services of VANETS networks unreliable.
- The need to include vehicular networks in the ongoing development of IoT which represents enormous opportunities. Estimated annual revenues range from $210 billion to $374 billions.

3.2.2 IoV Architecture Model

IoT is a theme that evolves from ITS VANET to meet the growing needs in smart vehicles that can benefit from the recent developments in information technology. To this end, IoV must be able to interface with several types of wireless access networks to allow smart vehicles to be connected anytime and anywhere. In addition to communications defined for VANET networks, the IoV architecture model supports other types such as Vehicle-to-Infrastructure (V2I), Vehicle-to-Device (V2D) and Vehicle-to-Sensors (V2S) (Fig. 9). In [13], Kaiwartya et al. propose a layered architecture for IoV. This architecture consists of five layers to allow an IoV object to access services of artificial intelligence through heterogeneous access networks.

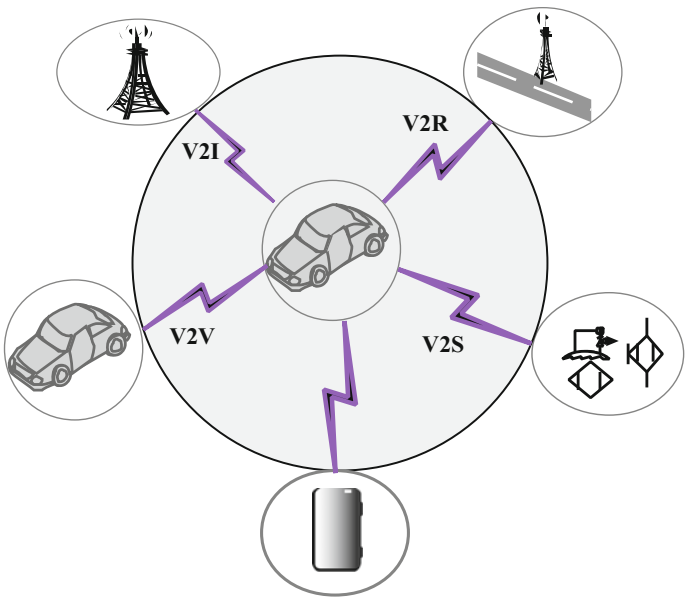


Fig. 9 IoV communication types

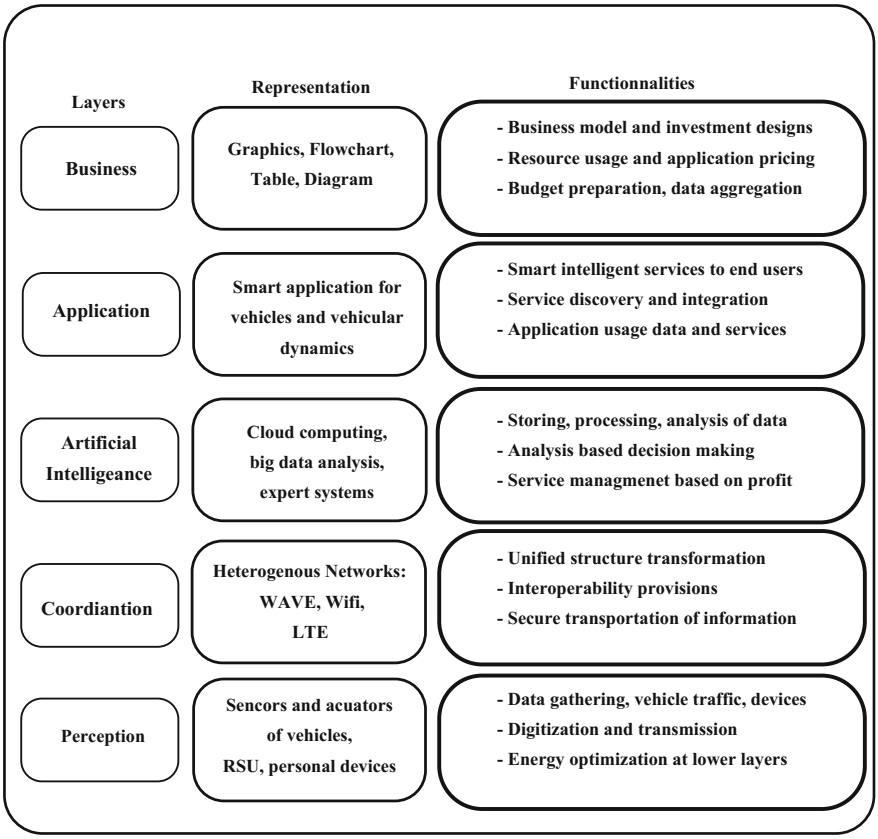


Fig. 10 IoV layered architecture [13]

This architecture also contains application layer and business layer which represent the operational management module of IoV. IoV architecture is depicted in Fig. 10.

In this model, the perception layer is responsible for collecting data from both vehicle itself and the people on board. Information relating to the vehicle is collected using sensors and actuators connected to the vehicle. This includes vehicle speed, direction, engine condition, traffic density, etc. The perception layer also collects infotainment information from the personal devices used by people on board. This layer is also responsible for the conversion of data for the coordination layer. To do this, the collection and transformation of data must be done in an efficient way in terms of energy and cost. These data are passed to the coordination layer which process information from different sources and format them in a structure that can be processed in a heterogeneous network context. This poses a challenge to establish a unified process of cooperation between various supported access networks that can be WAVE, Wi-Fi, 4G/LTE or stallite networks.

The third layer in this model is the artificial intelligence layer. It is the core of IoV and is responsible for storing and processing data. The essential components of this layer are: virtual could computing (VCC), Bid data analysis (BDA) and expert system.

The fourth layer provides smart applications of driving safety as well as those of infotainment based on the web. In addition to the applications already included in the ITS VANET standard, this layer provides commercial applications and smart services based on the analysis performed at the artificial intelligence layer. The application layer also allows efficient discovery of available services and supplies application use information to the business layer. Based on this information, the business layer is used to assist decision-making in the development of business models in the light of provided statistics.

4 Big Data and Internet of Vehicles

4.1 IoV Big Data Requirements and Challenges

Vehicles equipped with IoV technology would generate enormous amounts of data through different embedded sensors and actuators. This information is very useful for ensuring safe driving, managing infrastructure and fighting pollution caused by vehicles. For this, it is very primordial that this information is collected, stored and processed in an efficient way so that it can be exploited at the various levels of decision-making processes. A global architecture is essentially composed of three elements: users, connection and Cloud. An IoV data processing architecture must ensure a deep understanding of users and their devices, an uninterrupted connection connection based on ubiquitous coverage provided by heterogeneous networks and finally powerful tools to analyze the data and find the various common patterns. To manage the Big Data generated by the IoV, a Cloud Platform as a Service (PaaS) environment is the most appropriate since it allows the mutualization of systems and offers a great elasticity and capacity to adapt automatically to the demand.

4.2 IoV Could Computing Architecture

The cloud allows remote resources and services to be used instead of local resources. It has the advantage of having more standardized assignments such as broad network access, resource pooling, rapid elasticity and customized on demand services [14]. Cloud computing providers offer their services according to three models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) (Fig. 11). In a SaaS model, the user runs applications installed on the cloud infrastructure. Applications can be accessed via a web browser or a locally installed in-

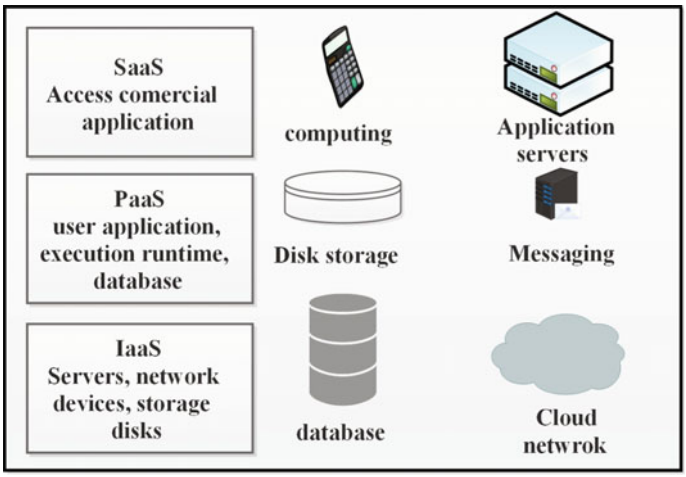


Fig. 11 Cloud computing models

terface. The user cannot manage the underlying infrastructure nor even the attributes of the applications except possibly, for a limited configuration of specific parameters. The PaaS model offers the user the ability to deploy consumer-created or acquired applications. It provide the user with libraries, services and tools for the creation and development of its applications. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage. In IaaS, a cloud user can run an arbitrary software using provided resources such as processing, storage, networks, and other fundamental computing resources. In this model, user does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications.

4.3 IoV Big Data Architecture

The model to analyze IoV Big Data is depicted in Fig. 12. Different sensors will be installed in each vehicle to measure speed, vehicle position, detect congestion, engines conditions and on-board systems, etc. In addition to these sensors, personal device aboard vehicles will be part of the collection of data to be transferred to the Cloud. The raw Big Data received will then be processed according to predefined rules in order to structure the information and prepare it for the next processing level. For example, positions calculated from several neighboring GPS sensors will be processed according to a predetermined algorithm to calculate the most likelihood position. Structured data will then be passed to a Big Data analysis tool and results transferred to be exploited by concerned entity. The amount of data generated by vehicles within the framework of IoV can reach magnitude of petabyte scale. To

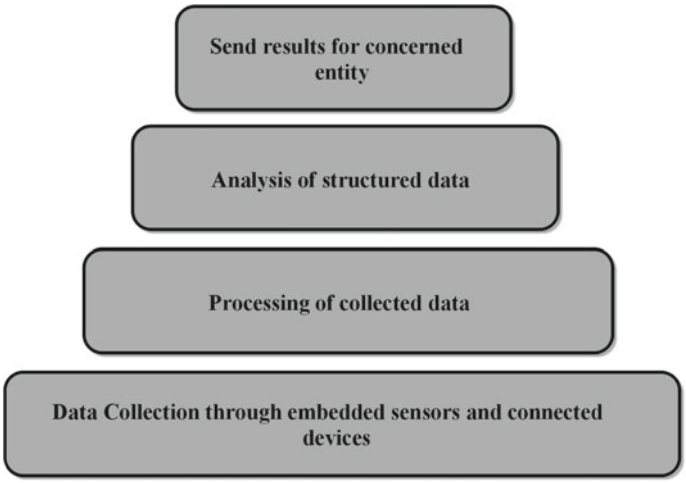


Fig. 12 IoV data analysis model

process and disseminate this amount of information, systems with adequate processing and storage capacity will be required. In [13], an architecture to support the Big Data is proposed. This architecture highlights the indispensable role that the Cloud would play in such a system (Fig. 13). IoV-oriented Cloud services will need to be deployed to provide Co-operation as a Service (CaaS), Storage as a Service (STaaS), Gateway as a Service (GaaS), Computing as a Service (COaaS), Network as a Service (NaaS), Data as a Service (DaaS). The platform will also offer ITS intelligent application servers for the collection and processing of Big Data. It will also ensure an end-to-end delivery of services to users.

4.4 IoV Big Data Limitations

The Internet of vehicles is a promising technology for the automotive and telecommunications industries. However, many limitations will have to be addressed before this technology can be fully operational. First, the V2V and V2I accesses must be bridged and merged. This raises the need to standardize the coexistence of several access technologies such as LTE and WAVE. The second limitation is related to the extremely large amount of IOV data that is beyond the limits of ordinary platforms. The operation of such a system can not ensured by traditional telecom operators or automobile manufacturers and a virtual operation must be set up. Finally, position determination using GPS can not meet the needs in terms of accuracy and security; There is a need for a more precise and secure navigation system.

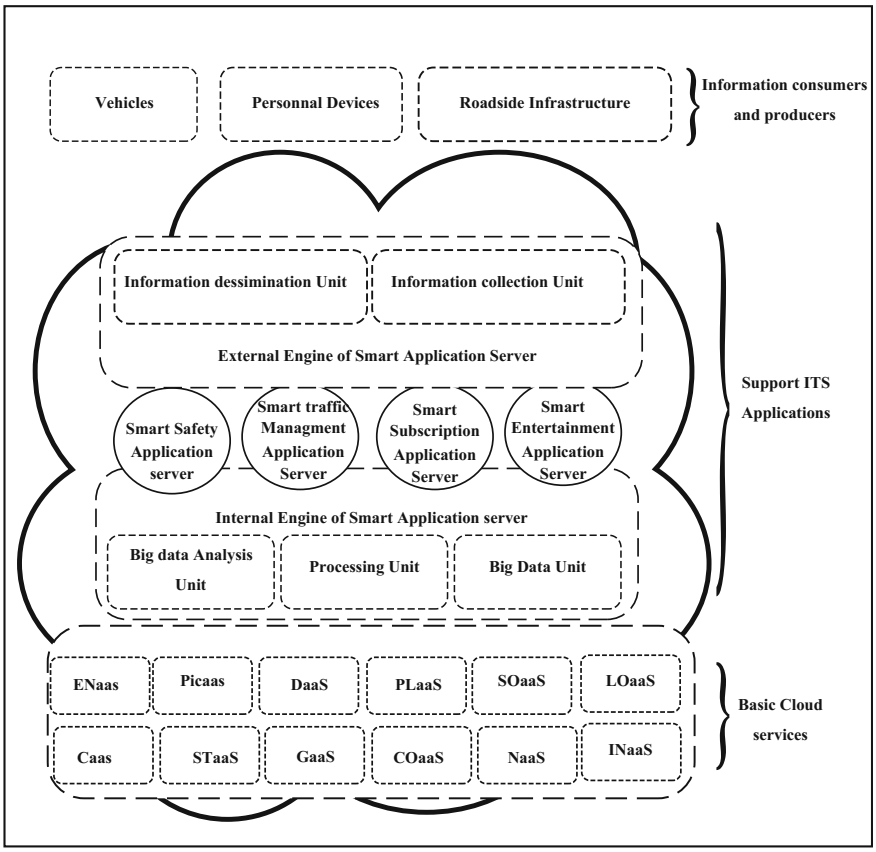


Fig. 13 IoV big data architecture

5 Conclusion

The operation of Big Data to improve and optimize costs of services offered by modern cities, will be one of the major keys to global development in the years to come. This is especially important in the context of the Internet of Vehicles given the amount and speed with which this kind of information is generated. In this chapter we have tried to clarify various issues that frame the development of technologies that can benefit from the use of Vehicular Big Data to improve the driving experience and allow decision-makers to have in their hands effective tools for infrastructure management. We have presented the evolution of vehicular networks from Adhoc with very limited area and unguaranteed service to networks with heterogeneous and universal access allowing ubiquitous connectivity. We then demonstrated the critical role that cloud platforms will play in collection, processing and operation of data gathered from embedded sensors as well as personal devices in a smart vehicle. We

finally detailed the architecture of IoV Big Data and highlighted the Cloud elements that will have to be deployed for an efficient IoV Big Data handling.

References

1. IDC Italia S.R.L., TXT e-solutions S.P.A.: Definition of a Research and Innovation Policy Leveraging Cloud Computing and IoT Combination. A study prepared for the European Commission–DG Communications Networks–Content and Technology. https://ec.europa.eu/digital-single-market/en/news/definition-research-and-innovation-policy-leveraging-cloud-computing-and-iot-combination. Accessed 26 Jan 2017
2. Tschofenig, H., Arkko, J., Thaler, D., McPherson, D.: Architectural considerations in smart object networking. Tech. no. RFC 7452. Internet Architecture Board, Mar 2015. https://www.rfc-editor.org/rfc/rfc7452.txt
3. Rose, K., Eldridge, S., Chapin, L.: The internet of things: an overview—understanding the issues and challenges of a more connected world. The Internet Society (ISOC), Oct 2015. https://digitalwatch.giplatform.org/resources/internet-things-iot-overview-understanding-issues-and-challenges-more-connected-world
4. Sperling, D., Gordonüth, D.: Two Billion Cars: Driving Toward Sustainability. Oxford University Press, July 2010
5. Study Prepared by HDR for the Office of Economic and Strategic Analysis, U.S. Department of Transportation, Assessing the Full Costs of Congestion on Surface Transportation Systems and Reducing Them through Pricing, February 2009. https://www.transportation.gov/office-policy/transportation-policy/assessing-full-costs-congestion-surface-transportation-systems. Accessed 26 Jan 2017
6. Menouar, H., Roscher, K.: In: C. Campolo, A. Molinaro, R. Scopigno (eds.) Forwarding in VANETs: GeoNetworking, Vehicular ad hoc Networks, vol. 74, pp. 221–251. Springer International Publishing (2015)
7. Li, Y.J.: An overview of the DSRC/WAVE technology. In: Zhang, X., Qiao, D. (eds.) Quality, Reliability, Security and Robustness in Heterogeneous Networks, vol. 74. pp. 544–558. Springer, Berlin (2012)
8. SAE International: DSRC Implementation Guide: a guide to users of SAE J2735 message sets over DSRC, Feb 2010
9. Saini, M., Alelaiwi, A., Saddik, A.E.: How close are we to realizing a pragmatic VANET solution? A Meta-Survey. ACM Comput. Surv. (CSUR), vol. 48. No 2, Nov 2015
10. World Health Organization, Global status report on road safety 2015. https://www.who.int. Accessed 26 Jan 2017
11. Hasan, S.F., Ding, X., Siddique, N.H., Chakraborty, S.: Measuring disruption in vehicular communications. IEEE Trans. Veh. Technol. **60**(1), 148–159 (2011)
12. Bi, S., Chen, C., Du, R., Guan, X.: Proper Handover between VANET and cellular network improves internet access. IEEE Veh. Technol. Conf. (VTC Fall) (2014)
13. Kaiwartya, O., Abdullah, A.H., Cao, Y., Altameem, A., Prasad, M.: Internet of vehicles: motivation layered architecture network model challenges and future aspects. IEEE Accessd **4**, 5356–5373 (2016)
14. Mell, T., Grance, P.: The NIST Definition of Cloud Computing, (Technical report). National Institute of Standards and Technology: U.S. Department of Commerce, NIST Special Publication 800-145, Oct 2011. https://www.nist.gov/publications. Accessed 26 Jan 2017

Part III
Data Management for Mobile Big Data

Mobile Distributed Complex Event Processing—Ubi Sumus? Quo Vadimus?

Fabrice Starks, Vera Goebel, Stein Kristiansen and Thomas Plagemann

Abstract One important class of applications for the Internet of Things is related to the need to gain timely and continuous situational awareness, like smart cities, automated traffic control, or emergency and rescue operations. Events happening in the real-world need to be detected in real-time based on sensor data and other data sources. Complex Event Processing (CEP) is a technology to detect complex (or composite) events in data streams and has been successfully applied in high volume and high velocity applications like stock market analysis. However, these application domains faced only the challenge of high performance, while the Internet of Things and Mobile Big Data introduce a new set of challenges caused by mobility. This chapter aims to explain these challenges and give an overview on how they are solved respectively how far state-of-the-art research has advanced to be useful to solve Mobile Big Data problems. At the infrastructure level the main challenge is to trade performance against resource consumption; and operator placement is the most dominant mechanism to address these problems. At the application and consumer level, mobile queries pose a new set of challenges for CEP. These are related to continuously changing positions of consumers and data sources, and the need to adapt the query processing to these changes. Finally, proper methods and tools for systematical testing and reproducible performance evaluation for mobile distributed CEP are needed but not yet available.

F. Starks · V. Goebel · S. Kristiansen · T. Plagemann (✉)
University of Oslo, Oslo, Norway
e-mail: plageman@ifi.uio.no

F. Starks
e-mail: fabriceb@ifi.uio.no

V. Goebel
e-mail: goebel@ifi.uio.no

S. Kristiansen
e-mail: steikr@ifi.uio.no

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_7

1 Introduction and Motivation

The Internet of Things means pervasive deployment of stationary and mobile sensors, which produce high velocity and high volume data in the form of data streams. A data stream is conceptually an infinite sequence of data tuples comprising typically the digital values of a signal measured by a sensor in the real world and a timestamp denoting when a sensor has generated the value. Real-time analysis of data streams is in Mobile Big Data important for two reasons: (1) the sheer amount of data can make it infeasible to store all data on secondary store and index it before the analysis and (2) many application domains, like smart city, automated traffic control, environmental monitoring, or emergency and rescue operations aim to maintain continuous situational awareness and if certain events happen to react to as fast as possible to them.

One promising technology to achieve situational awareness and detect events of interest in real-time is Complex Event Processing (CEP). The core idea of CEP is to regard the tuples in data streams that are generated by sources like sensors as primitive (also called atomic) events and to extract new knowledge out of the primitive events and represent it as composite events. CEP systems have become rather popular due to the powerful event paradigm and the fact that consumers can describe the composite events they are interested in the form of declarative statements or queries. Originally, the need for real-time processing of data streams, for example in stock trading, triggered the development of CEP systems and a lot of emphasis has been put onto efficiency and scalability of these systems. Naturally, CEP systems have evolved from centralized solutions to distributed solutions to be able to process larger amounts of data in real-time. Most of the DCEP research results and systems target high performance systems with stable infrastructures.

One important challenge for DCEP in Mobile Big Data and the Internet of Things is the fact that one cannot always rely on a stable and high performance infrastructure. Mobility implies the use of wireless networking technologies with potential bandwidth limitations, dependency on battery lifetime in mobile devices, and a dynamic network topology. These challenges are especially severe if infrastructure is not available, e.g., in disaster areas, and multi-hop wireless networks are established for communication. Thus, to use CEP for Mobile Big Data these infrastructure challenges have to be addressed. On the other hand properly designed DCEP can be well suited to address these challenges. For example, source filtering and data aggregation as close as possible to the data sources saves scarce resources, like bandwidth and energy of mobile devices. Furthermore, data aggregation at the network edge has the potential to improve privacy protection. The most important mechanism to address these challenges in DCEP is *operator placement* to determine which data processing tasks should be performed on which node.

Mobile consumers and/or mobile data sources introduce another important challenge for DCEP. In such scenarios, so-called *mobile queries* are traditionally processed in spatio-temporal databases to support for example location aware services, e.g., to provide a car driver continuously updated information about congestions in

the range of 1 Km of the drivers current position. Handling properly such spatio-temporal data in CEP systems is a rather new, but important research topic.

Finally, we need to point out that there has been so far no systematic attempt for methods and approaches to evaluate the performance of mobile DCEP in such a way that evaluation results are (easily) reproducible by peer researchers.

It is the aim of this chapter to enable the reader to understand the potential of DCEP for Mobile Big Data and the particular challenges that are introduced by Mobile Big Data. Based on a survey of the state-of-the-art in DCEP we analyze to which extent DCEP is ready for such mobile environments. Finally, we provide the reader with an insight into the main unsolved technical issues and future research directions in the area. Several papers have captured the state-of-the-art in the area of Data Stream Managements Systems and CEP, but to the best of our knowledge there are no surveys on DCEP and especially not on DCEP in the mobile context.

The reminder of this book chapter is structured as follows. In Sect. 2 we provide some background information on CEP and DCEP, followed by an analysis of the main challenges for DCEP in Mobile Big Data. Section 4 presents the operator placement problem and classifies existing solutions and Sect. 5 focuses on the challenges introduced by mobile consumer, mobile data sources, and mobile queries to handle spatio-temporal data; and Sect. 6 discusses the needs for proper testing and performance evaluation methods and approaches. The conclusions in Sect. 7 summarize the current status of mobile DCEP research and yet unsolved challenges.

2 Complex Event Processing Background

Traditionally, database systems have been used to manage large amounts of data, typically by materializing it on secondary storage, e.g., storing it on disks, indexing the data, and providing a declarative Application Programming Interface (API) like SQL for asynchronous data processing on demand. However, the emergence of new applications for sensor networks, Internet traffic analysis, financial tickers, online auctions and analysis of transactional logs from web usage and telephone records introduced in the beginning of this century the need for new software solutions to be able to analyze data streams in real-time [1]. These new software solutions, called Data Stream Management Systems (DSMS), introduced the concept of data streams. A *data stream* is basically a continuous, ordered sequence of data tuples. Conceptually, data streams are similar to classical database tables. Furthermore, the concept of classical database queries has been adopted to run continuous queries over data streams to return continuously new results as new data tuples arrive. The query languages for DSMS, called Continuous Query Language (CQL), are very similar to SQL. The main difference between CQL and SQL is the need to use windows over the data stream for processing. Blocking operators, like aggregations or joins, introduce this need because they can only be used with the entire data set to produce a resultd. It is in most cases not feasible to wait until the data stream finishes, which means the entire data set is available. Therefore, windows are used to process subsets

of the data, which in turn is a number of sequential data tuples from the data stream. The size of a window is either defined by time or by number of data tuples that should be processed from a data stream at a time (per query result). Once the set of samples in a window is processed, the result for this window is returned and the window is forwarded over the data stream and processed again with the new sample(s). DSMS are capable of querying several streaming sources at once, and additionally joining and correlating them in real-time. Large queries can be split into smaller queries and easily processed in a distributed manner, since a query usually results in another stream that can be sent to another query for further analysis. Examples of DSMS include SQLstream [2], STREAM [3], AURORA [4], StreamGlobe [5] and Esper [6].

Esper is also a good example how new achievements in data stream processing lead to a new class of systems, CEP systems, with even stronger abstractions and stronger stream processing capabilities. These innovations are based on the concept of events. An event can intuitively be defined as *something that happend* and is either an atomic event or a composite event (also called complex event). In probability theory, an atomic event (also called elementary event or simple event) is a subset of the sample space that only contains a single outcome. In computer science, an atomic event is often understood as an event that can be detected by a system within a minimum time period and cannot be divided into other events. The authors understand an atomic event as a single sample from a sensor measuring a signal in the real world, or it is a transformation of an atomic event. For example, a sample from a sensor measuring temperature in degrees of Celsius is an atomic event, as well as a later transformation of this sample into a corresponding value in degrees of Fahrenheit. A composite event is the result of processing a set of events that are combined with operators, like statistical, logical, temporal, or spatial operators. The basic idea is that application programmers define the event they are interested in and the CEP system is analyzing in real-time the incoming event stream(s) and informs the application as soon as it detected the event of interest. Examples of existing CEP systems are SQLstream [2], StreamInsight [7], EVAM [8], or Esper [6].

DSMS [9] and CEP [10] have common goals, but the systems differ in many aspects: architecture, data models, rule languages, and processing mechanisms [11]. Furthermore, DSMS and CEP have their roots in different research communities: DSMS have their roots in the data base systems community, whereas CEP has evolved from Publish/Subscribe systems [12].

The main difference between DSMS and CEP is according to [11] that data items are considered as streams of data versus notifications of events. This means that DSMS handle the Information Flow Processing problem as processing streams of data, which originate from different sources in order to produce new data streams as output. DSMS deal with transient data that is continuously updated executing continuous (standing) queries over the stream items.

In contrast, CEP considers data items as notifications of events. Events are happening in the physical world, which have to be filtered and combined to understand what is happening in terms of higher-level events. The focus of CEP is to detect occurrences of particular patterns of (low-level) events that represent the higher-level

events. The occurrence of higher-level events has to be notified to consumers that have subscribed to these events, typically by registering a continuous query to the system that describes the patterns of events. This relationship between the CEP system and consumers is inherited from the simpler form in Publish/Subscribe systems. Traditional Publish/Subscribe systems consider each event separately from the others, and support topic or content filtering to determine whether a notification should be send to a subscriber. CEP systems have much more expressive subscription languages, e.g., CQL, to describe composite event patterns. Typically, mathematical, logical, temporal, and spatial relationships can be used to describe these composite event patterns. If A and B are two different events (for example a tuple in a data stream has the value A respectively B), the following composite event patterns could be described:

- $A \wedge B$: the logical $\wedge$ operator can be combined with a time window during which A and B must happen.
- $A \vee B$: the logical $\vee$ can be combined with a time window during which either A or B must happen.
- $A \rightarrow B$: the temporal operator $\rightarrow$ defines that A must happen before B. This operator can also be combined with a time window.
- $A <> B$: a spatial location operator which defines that the location where A happens and the location where B happens overlap.

Another important difference between DSMS and CEP is the fact that CEP is stateful and DSMS stateless. DSMS use windows to enable the use of blocking operators. A window determines one particular sequence of data tuples. Once all tuples in a window are processed, the result is forwarded as output in DSMS. Therefore, DSMS are not able to detect specific sequences of events in an event stream. To be able to detect specific event sequences in CEP, they are typically built on a state machine and use so called *selection* and *consumption policies* to determine which events to consider for processing. Events that are part of a given event sequence trigger a state transition in this state machine until the final state is reached and the event pattern is detected. Selection policies determine which incoming events are used during processing and consumption policies determine what to do with events that have been processed, e.g., whether to evict an event from the memory or to re-use it in the next processing iteration.

The step from centralized CEP to distributed CEP (DCEP) has two main reasons: (1) parallelizing CEP engines to scale the performance of CEP and beeing able to process more data in real-time, and (2) the fact that recent CEP application domains like environmental monitoring or smart cities comprise a large number of distributed information sources, e.g., sensors and information sinks, human consumers or control systems [11]. Etzion et al. [13] structure channel based event distribution into event producers, event channels, and consumers. Event channels can be an intermediary service that is often called broker. As such, a DCEP engine can be seen as a set of event brokers that are connected in an overlay network, also called *event processing network* [11].

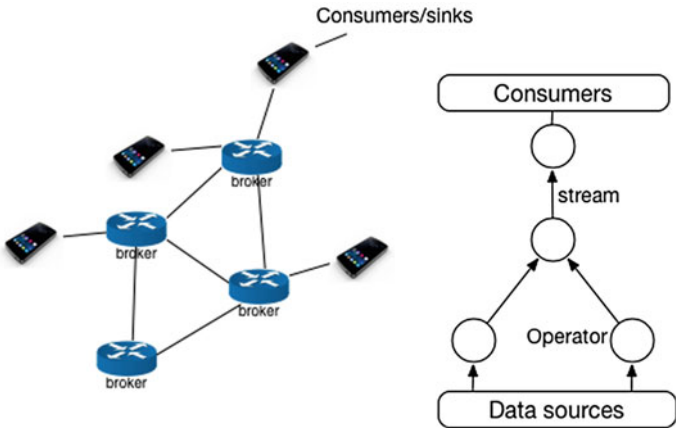


Fig. 1 System model and CEP operator tree (according to [14])

Koldehofe et al. [14] model the operation of a DCEP system by an operator tree (see Fig. 1). The operator tree is a directed graph with three different types of nodes: operators, sources, and consumers; and the links are event streams between operators, sources, and consumers. Each operator is hosted by some broker and implements a correlation function which defines the mapping of input events of the operator to outgoing event stream. Operator placement is the task to assign each operator to a broker in the event processing network (or operator network). The event processing network implements specialized routing and forwarding and aims at high scalability for high performance CEP. Therefore, a lot of DCEP research has aimed to optimize bandwidth utilization and end-to-end latency, which are usually ignored in DSMS [11]. For mobile DCEP many more optimization parameters are important, like energy consumption or security constraints, and are discussed in detail in Sect. 4.

3 Requirements for Mobile DCEP

It is well known that the main challenges of Big Data are caused by the volume, velocity, variety, and veracity of the data. Mobility in Mobile Big Data adds another dimension to this problem domain. In order to understand the challenges mobility introduces for DCEP we consider mobility from two viewpoints: (1) the impact of mobility on the computing infrastructure and (2) from the applications respectively consumers point of view. We assume, without loss of generality, that the goal of CEP applications is to provide timely situational awareness to consumers.

Mobile devices, like sensors, smart phones, tablets, laptops, and other computing devices obviously require the use of wireless networking technology and need to be battery driven. There are two basic classes of networking approaches that are

used to connect mobile devices, which are often called *infrastructure-based* and *infrastructure-less*. In infrastructure-based approaches only the edge of the network to which the mobile devices connect is wireless, typically a cellular network (e.g., 3G, 4G, and the future 5G), or a WiFi network. These wireless edge networks are connected with the Internet by a wired network infrastructure. In infrastructure-less networks, computing devices form with their wireless networking interfaces in promiscuous mode a multi-hop wireless network like a Mobile Ad-Hoc Networks (MANET), Wireless Sensor Networks (WSN), or Vehicle Area Networks (VANET). Obviously, there are many combinations of these two classes of networking possible, but these two are sufficient to identify the challenges caused by mobile devices. The fundamental mechanism in DCEP to address these challenges is operator placement. Section 4 gives an explanation and definition of the operator placement problem, as well as a classification of state-of-the-art solutions of operator placement for mobile DCEP.

Before discussing these infrastructure-related challenges we first aim to give the reader an intuitive understanding of the issues caused by consumer and application needs in mobile settings due to the spatio-temporal nature of data that needs to be handled. Spatio-temporal means for example that objects have a location, i.e., the spatial property of an object, and that moving objects change their location over time, which in turn is a spatio-temporal aspect of moving objects. Moving objects can have the role of data sources, e.g., a car that continuously reports its location and other sensor data collected by the car; or moving objects can have the role of consumers. Consider for example a service that is using data from road and parking lot sensors to give the moving consumer in real-time information on free parking lots in the vicinity of the consumers. To provide the consumer with this information only data from sensors in the vicinity of the user need to be analyzed. The query to produce the information for the service is continuously running while the location of the mobile user is changing. Due to the change in user location, the set of sensors that are in the vicinity of the user is changing. Such a *mobile query* requires to continuously adapt the set of sensors to be used. Additionally, many users typically use such a service at the same time; some might be at very distant locations and some closer to each other. In the latter case, the sets of relevant sensors for the users that are currently close to each other overlap. Researchers face the problem of how to avoid that the common subset of sensor data is transferred and processed multiple times, because redundancy reduction means resource savings, in terms of bandwidth, computational capacity, or energy. The fact that these users have different mobility patterns makes this kind of redundancy reduction harder since the common subset of relevant sensors is continuously changing. In Sect. 5, we explain in more detail the issues DCEP needs to address to support mobile queries, and to properly and efficiently handle spatio-temporal data.

The need of efficient data handling and careful resource consumption is directly implied by the use of mobile devices. Mobile devices are battery driven and one important research goal in mobile wireless networks in general is to use the limited amount of energy in the battery as good as possible. Recharging of batteries or changing of batteries (like in wireless sensors) is in in the best case cumbersome and

in the case of WSN potentially very expensive. Therefore, energy efficiency is the ultimate goal in WSN, rather important in MANETs since it directly relates to the lifetime of the MANET, and of less importance in VANETs since the engine of a car can continuously charge the battery. In case of infrastructure-based networks the device owner is confronted with the consequences of battery lifetime and the need for recharging.

The fact that transmit and receive operations of mobile devices are substantially contributing to their energy consumption implies directly to design solutions that carefully handle networking resources, both in terms of bytes per second transmitted (i.e., bandwidth consumption) and packets per second. Another reason to consider bandwidth consumption is the fact that wireless networks are based on a shared medium with a limited amount of bandwidth. Wireless networks can also be affected by noise, high rates of packet collisions and unstable connectivity due to (too) long distances between sender and receiver, which in turn can result in higher packet loss rate and lower bandwidth.

This situation results in a rather large set of conflicting requirements for design, implementation, and deployment of mobile DCEP.

- Low event delivery delay is important to enable situational awareness for the consumer and to initiate immediately certain actions to react to detected events of interest. The potentially large amounts of data that need to be handled increase this challenge.
- Complete and consistent results are needed to achieve correct situational awareness. This means for example that a DCEP system needs to guarantee a high event delivery ratio with a high Quality of Information.
- Efficient resource consumption in terms of computational costs, network utilization, and even monetary costs (if it is necessary to buy resources) is important for several reasons: to achieve low cost services for consumers, saving energy consumed for computational and networking tasks, and being able to handle as good as possible high volume and high velocity data.
- Enable scalable solutions to handle the ever increasing amount of data sources, different consumer interests and volume of data. On the architectural level, distributed and parallel processing needs to be supported. On the application level flexible concepts and corresponding support for Quality of Information need to be supported such that in case of too high system load the Quality of Information can be degraded to a certain level and still acceptable results can be produced, e.g., through load shedding.
- Reliability and fault-tolerance is important especially if the situation awareness is be used for crucial tasks, like traffic control or industrial control systems. Infrastructure components can be prone to hardware and software failures, packets can be lost in wireless networks due to noise, mobile devices might be turned off due to empty batteries, connections might be lost, or even networks might be partitioned.

Operator placement is a rather powerful mechanism in DCEP and can be used to address several of the above-mentioned requirements.

4 Operator Placement

The classical approach for data mining is to send all data to a central server and to process it on the server. However, in Internet of Things applications that establish situational awareness typically only a particular subset of the data is of interest for the application. Sending irrelevant data to the server is obviously a waste of resources. A simple, but efficient approach to reduce resource consumption and enable scalable mobile DCEP systems is to filter events as close as possible to their data source, in the best case, directly at the source, i.e., source filtering. Source filtering is the first step to minimize the consumption of shared resources by eliminating irrelevant events at their sources. The next step is to perform the aggregation and matching of the events in the vicinities of their sources [15]. Processing events near their sources, referred to as *in-network processing*, filters out events that are not of interest and eliminates duplicates early, which in turn reduces system bandwidth and energy consumption, which is especially important in wireless networks. In-network processing takes advantage of increasingly powerful fixed and mobile devices in wireless edge networks. However, the heterogeneity, resource limitations, privacy, security and other challenges related to these edge networks makes in-network processing intricate to implement.

The basic idea behind in-network processing in CEP is that queries are transformed into an operator tree, and that the operators in the tree structure can be processed independently on event brokers. The operators are assigned to brokers in such a way that the performance goals of the system are achieved. Once placed on the brokers, the operators are processed in a CEP overlay called *operator network* [16].

An operator network is a class of overlay networks used for data stream and event stream in-network processing. Operators assigned to physical hosts form an overlay network, which process data from distributed data sources. Results from the operator network data processing are delivered to user applications which are hosted to physical host(s) called sinks.

In large scale operator networks, the physical hosts to which operators are assigned have a significant and direct impact on the performance of the entire system [17]. The operator placement mechanism is responsible for building and maintaining the operator network through operator placement and adaptation, and has an important role in the optimization of the system performance.

Due to its importance in DCEP, operator placement is the most investigated mechanism in DCEP related research and its description is correspondingly prominent in this book chapter. In the following, we first give a more in-depth, but informal explanation of operator placement before we formulate the operator placement problem as multi-dimensional optimization problem in Sect. 4.2. The formal problem definition represents also the foundation for a classification of existing operator placement research. The structure of existing operator placement mechanisms, i.e., centralized and decentralized operator placement is explained in Sect. 4.3 and operator placement adaptation is explained and classified in Sect. 4.4.

4.1 General Idea

An operator placement mechanism is concerned with how to optimally assign a set of operators to brokers. The goal of an operator placement mechanism is to build an operator network, which optimizes resource consumption and achieves the performance targets of the system. As an example, in Mobile DCEP systems, the operator network would need to optimize the consumption of shared and scarce system resources such as bandwidth while ensuring the performance in terms of low latency.

To achieve its goal, the placement mechanism is provided with the following information:

- an operator tree,
- a set of physical hosts with stream processing capability, i.e., a set of brokers,
- resource availability and demand profiles, and
- a set of constraints.

The operator tree is an internal system representation of a CEP query ready to be assigned to brokers. Figure 2 shows a simple operator network to process $(A \wedge B) \vee (C \wedge D)$. Some operators in the operator tree are intuitively pre-assigned to specific brokers. The leaves of the operator tree are typically placed on their respective data sources. In Fig. 2, the filters for atomic events A, B , C and D are pre-

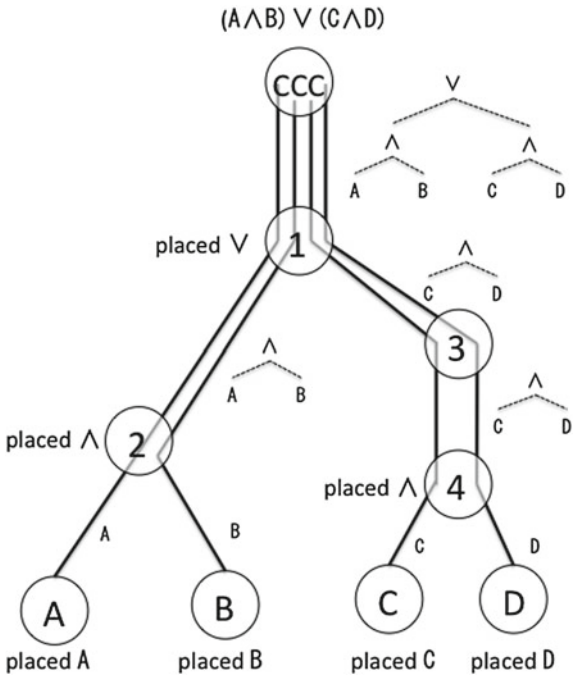


Fig. 2 Operator network in a MANET for an Emergency and Rescue Mission: The circles represent physical hosts. the sink (ccc), the event brokers (1, 2, 3, 4) and data sources labeled with their corresponding pinned operators (A, B, C, D). The unpinned operators are placed besides their processors/ event brokers. Events traversing the operator network edges are results from sub-trees in the operator tree. The edges in the operator tree are labeled with corresponding sub-trees

assigned to the sources of the data streams they are supposed to filter, i.e., atomic events that match A, B, C, or D. As an example, the temperature sampling operator should only be placed on nodes with temperature sensors. The output of the root in the operator tree is forwarded directly to the node hosting the CEP application. In Fig. 2 the root operator ∨ placed on Broker 1 forwards its output to the application node: the Command and Control Center (CCC) in this case. Other operators can be bound to specific brokers for monetary, privacy or security reasons. We refer to the operators with predefined placement assignment as *pinned operators*. The remaining operators are referred to as *unpinned* and are assigned to brokers by the placement algorithm.

It is typical to differentiate between physical hosts in the operator network based on their role, i.e., data source nodes, brokers, and sink(s).

The data source nodes generate atomic events for the DCEP system and therefore are pre-assigned the leaves of the operator tree. In Fig. 2, the nodes A, B, C and D are data sources for the corresponding leaves in the operator tree. Data source nodes can also process other operators in the operator tree. The brokers are those which are eligible to process unpinned operators (nodes 1, 2, 3, 4 in Fig. 2) and sink(s) (node CCC in Fig. 2) are nodes which have a direct connection to CEP application(s) and are responsible for submitting queries to the DCEP system. A sink is a typical location to place the root of an operator tree. Notice that it is possible for a single node to process several operators of an operator tree.

The output of an operator placement mechanism is an operator placement scheme which is a blueprint for an operator network. An operator placement algorithm can build the operator network in a centralized or decentralized manner. Furthermore, most operator placement algorithms implement an adaptation strategy in order to maintain the desired performance as the system and its environment change. The underlying problem of assigning a set of operators to a set of processing nodes has been found to be NP-Complete. However, heuristics based algorithms can be used to find placement solutions in large scale scenarios [18].

In essence, the operator placement problem is an optimization problem. It aims to find query processing scheme which yields an optimal system performance and resource consumption within certain system or application constraints. While the performance of CEP applications is a priority, the operator placement mechanism needs to find an optimal resource consumption scheme in order to ensure the scalability of the system. More so, in some systems, the consumption of system resources such as energy, has a direct impact on how long the system remains operational.

More so, in some environments, the consumption of system resource determines how long it can remain operational.

The optimal placement assignment scheme is found within the predefined constraints provided to the placement mechanism [19, 20]. An example of an application-defined constraint is the maximum allowed end-to-end latency. Such a constraint defines the solution space for the placement mechanism and the latter typically use an objective function to find the optimal placement assignment solution.

Finding the optimal operator placement assignment for DCEP system involves two main activities. The first activity is concerned with defining the main optimiza-

tion metrics for a system and formulating a constrained or unconstrained optimization function. The second activity is concerned with creating an algorithm that effectively solves the optimization function and finds an optimal placement for an operator tree.

In the next section, we formally define the operator placement problem and explore the main optimization goals addressed in existing research along with examples for illustration. Afterwards, we investigate existing placement algorithm design characteristics and adaptation approaches.

4.2 Problem Formulation

The operator placement problem is an optimization problem similar to the task assignment problem. Given a set of operators and nodes on which they can be processed, the optimal assignment that yields the best system performance should be determined. Existing operator placement algorithms try to solve either a constrained or unconstrained placement optimization problem. Constrained placement algorithms consider the optimization constraints as a means to ensure some QoS for the application-perceived performance [20]. However, it is also possible to apply resource consumption-related constraints to the placement optimization problem. This ensures an efficient usage of the system's shared resources in order to achieve high scalability and longer lifespan (when applicable). Other placement algorithms solve an unconstrained optimization problem with just an objective function to optimize.

An objective function, which captures relevant system performance and resource consumption metrics, is used to determine the optimal solution. The objective function typically defines critical resources and performance metrics to optimize.

For example, given N processing nodes available for processing O operators, the cost of processing an operator o on a node n is: $C_{(o,n)}$ for $o = 1, \ldots, O$ and $n = 1, \ldots, N$. If we consider $P_{(o,n)}$ as the assignment of operator o to node n, the objective function is defined as follows:

$$min \sum_{o=1}^{O} \sum_{n=1}^{N} C_{on} P_{on} \quad : \quad P_{on} \in \{0, 1\} \tag{1}$$

where $P_{on} = 1$ when operator o is placed on node n, and $P_{on} = 0$ otherwise.
In this particular case, the objective is to minimize the overall cost of processing all operators from an operator tree. Other examples of objectives are end-to-end-latency, energy consumption, etc.

The optimization objective and constraints are used to model the targeted system performance. Consequently, they reflect aspects of the challenges faced by the system and its overall performance goals.

In particular, the constraints are used to define boundaries for allowed resource consumption and application performance schemes. They determine the placement

assignment solution space from which the optimal solution is to be selected. As an example, for real time data stream systems, timeliness is a pre-requisite to function appropriately. End-to-end latency constraints can be applied on the optimization problem in order to ensure a maximum end-to-end delay.

Using information about system resource availability and application demands for such resources, constraints for the placement assignment problem can be defined to ensure a certain degree of application performance while containing the consumption of system resources within acceptable levels for the scalability of the system. It is also possible to define constraints that enforce policies related to privacy, security, etc. For example, Cipriano et al. [21] consider security as a deployment constraint, which requires that only physical nodes that hold a certain certificate can serve as brokers.

Objective functions can be used with or without constraints. The definition of the objective function is the first step in the process towards creating an efficient and effective operator network, because the parameters in the objective function reflect the critical resources or performance metrics that should be optimized. The objective is a quantitative measure of the performance targets of the system that needs to be maximized or minimized. Obviously, different systems have different performance targets, which are determined by either the application performance requirements or the scarcity of certain system resources. For example, the performance goal of the system might be to minimize end-to-end latency in cases with real-time applications such as CEP. In other cases, the main goal might be to minimize the consumption of scarce resource in order to ensure the scalability of the system.

In the following subsections, we present how the most important parameters, i.e., energy consumption and network usage, are included in objective functions and considered in constrained placement, before we use the parameters to classify existing placement solutions.

4.2.1 Energy Consumption

The first group of research works focuses on the issue of energy scarcity in WSN. Energy consumption optimization stands out as a critical part of the design, deployment and operation of WSNs [22, 23]. Data transmission has been found to be the biggest energy consumer, therefore, most research papers on operator placement in WSNs focus on minimizing data transmission over the network with in-network processing [23–28]. As such the cost function to minimize is defined as:

$$\sum_{a \in A} d_a C_{a_t}^{a_h} \tag{2}$$

where A is the set of all links in the operator network and d_a is the data rate on link $a \quad \forall a \in A$. $C_{a_t}^{a_h}$ is the communication cost on the link a where a_h and a_t are the ingress and egress operators of the link. Given an operator network link between two operators, the placement mechanism should place the two operators such that

the communication cost is minimized, especially if the data rate between the two operators is high. Please take note that the operator network link might comprise at the physical network level several nodes and the links between them.

One example where the problem of operator placement for energy optimization is addressed is the research work of Bonfils et al. [24]. Their optimization goal is to minimize the amount of data transferred over the network in order to minimize network energy consumption. For example, a user wants to be notified when two related events are detected in two distinct regions of the network within a predefined time window. This is expressed using a correlation operator, which consumes the related events from the two regions. In this scenario, the data sources reside in the two regions and the sink consumes events produced by the correlation operator. A correlation operator is very selective, which means that it produces a significantly lower amount of data compared to its data input. As such, its placement is crucial for the amount of data transmitted in the network. Ideally, the correlation operator should be pushed close to the data sources in order to eliminate duplicates as soon as possible. Figure 3 [24] shows two cases with to different placements of the correlation operator that each minimize data transmission, depending on the data rate from the data sources. In Fig. 3a, one of the regions is generating a significantly large, amount of data compared to the other region. Therefore, it makes sense to place the correlation operator close to the data source in the high data rate region, to minimize the overall network data transmission. In Fig. 3b, both regions are producing approximately the same amount of data, therefore, the path length between the data sources is considered instead. Thus, the placement of the correlation operator depends on: (1) the data rate of both the operator and the data sources, and (2) the path length between the data sources, the correlation operator and the sink.

Consequently, the placement optimization problem in [24] captures the data rate and path length between an operator and each one of its children as optimization goals. They consider a sensor network as a directed graph where vertices represent

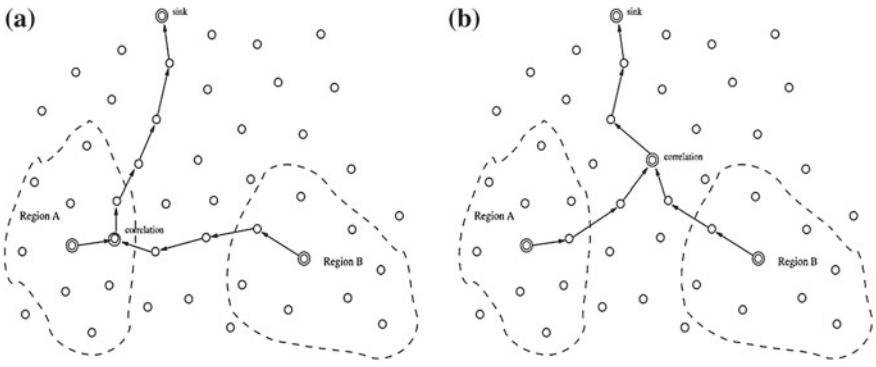


Fig. 3 Optimal operator placement examples for different data rate scenarios: **a** high data rate from region A, **b** more or less similar data rates from the two regions [24]

sensor nodes and where edges represent communication links, and a query is a operator tree with a tree structure [24]. The placement problem is modeled as the assignment of operators onto nodes that minimizes the global cost:

$$min \sum_{(i,j)\in\lambda} x_{ip} x_{jq} S_{pq}(d_{ij}) \tag{3}$$

subject to

$$\sum_{p\in\pi} x_{ip} = 1, \quad \forall i \in \eta, \forall p \in \pi : x_{ip} \in \{0,1\} \tag{4}$$

λ is the set of all edges in the operator tree. $S_{pq}(d_{ij})$ is the data rate between nodes p and q processing operators i and j respectively. $x_{ip} = 1$ if operator i is placed on node p, and $x_{ip} = 0$ otherwise. η is the set of all operators in the operator tree and π is the set of all physical hosts.

4.2.2 Network Usage

Most CEP systems are either real time or near real time, which means that low latency is an important metric that should be part of the objective to optimize. Another characteristic of mobile DCEP is the high amount of data, which requires significant system resources. In particular, the shared bandwidth in mobile systems becomes a scarce resource, and its consumption must be optimized to achieve the expected degree of scalability.

As such, the placement mechanism needs to find an optimal placement assignment that achieves the right balance between optimal bandwidth consumption for scalability and minimal end-to-end latency. Some research papers use the bandwidth delay product as the objective to minimize in order to find a resource efficient and low latency query processing scheme [16, 20, 29, 30]. The bandwidth delay product is referred to as the network usage and defined as follows:

$$\sum_{l\in L} dr(l) Lat(l) \tag{5}$$

where L is the set of all links in the operator network, $dr(l)$ is the data rate on link l in the operator network, and $Lat(l)$ is the delay on link l. The objective above includes both the scarce resource and the main performance metrics to optimize. By minimizing such an objective, it is possible to find an optimal solution both in terms of bandwidth consumption and end-to-end latency.

Pietzuch et al. [30] present a placement algorithm that aims to minimize the network usage of a query processing scheme while maintaining low latency. The goal for this algorithm is twofold: on the one hand it should achieve good streaming application perceived performance such as low delay, and on the other hand it should

at the same time optimize the consumption of scarce resources by minimizing the bandwidth consumption in order to support a large number of streams.

Satisfying the application performance needs while minimizing the overall bandwidth usage is particularly challenging as techniques to optimize one can produce sub-optimal performance for the other. In particular, a technique to minimize the application perceived delay would choose the shortest paths between data sources and consumers and use them to transfer all data between them. One technique to optimize bandwidth consumption is to balance bandwidth usage in the network by routing data through potentially longer routes in order to distribute the network load. When choosing only the shortest paths, certain links in the network will quickly be overloaded with data and either fail (node failure due to lack of battery energy) or start dropping data.

The bandwidth delay product (network usage) metric is used in [30] to model an objective function to calculate the optimal solution in terms of bandwidth utilization and end-to-end latency. The network usage $u(q)$ to minimize is modeled as in Eq. 5.

4.2.3 Constrained Optimization

Another way to consider network related parameters like latency and bandwidth consumption and other parameters in operator placement is to use these parameters as constraints. For example, a maximum allowed latency can be expressed as a constraint for the operator placement mechanism [20]. The constraint defines a maximum allowed end-to-end latency which effectively reduces the set of eligible placement assignment solutions. Only those placement assignment solutions with an end-to-end delay below the predefined maximum are eligible for the optimal solution. This works well with the network usage objective function, as it eliminates solutions with poor end-to-end latency no matter how efficient they might be in terms of network usage. As such, the solution to the objective expressed in Eq. 5 is found from placement assignments that meet the following latency restriction:

$$L(G) \leq R \tag{6}$$

$L(G)$ is the end-to-end latency experienced by the application and R is the maximum end-to-end latency.

Rizou et al. [20] optimize the network usage within a predefined end-to-end delay constraint. A maximum allowed end-to-end delay is important for real time and near real time applications. The placement problem is addressed in a two-stage approach. In the first stage, an optimal solution is found based on the objective alone. In the second stage, the optimal solution found in the first stage is modified to satisfy the latency constraints while ensuring that the initial network usage is only slightly increased. The unconstrained optimization phase is performed in a centralized manner, while the constrained optimization phase is performed in a distributed manner.

4.2.4 Classification of Placement Mechanisms

The optimization goals are a good foundation for a classification of the most prominent operator placement approaches for Mobile DCEP. Table 1 shows the resulting classification. It can be clearly seen in Table 1 that most approaches target energy consumption. Energy consumption is considered as the most important optimization goal in WSNs and its optimization is a means to prolong the lifetime of the mobile DCEP systems that are deployed in networks with limited energy.

Network usage is the second most important optimization metric addressed by a significant number of operator placement mechanisms. Its popularity is due to its ability to capture both the limited bandwidth resource and application latency requirements. Furthermore, most research works in this category target mobile networks where energy consumption is not as crucial as in WSN, for example because it is easier to recharge the battery of a smart phone compared to sensors deployed at remote locations.

Few research papers have yet addressed the placement problem as a constrained optimization problem. This is however a natural next step towards optimal placement schemes to effectively address both the need for efficient resource consumption and low latency in mobile DCEP systems. Furthermore, constraints provide an easy means to implement triggers for placement adaptation (see Sect. 4.4).

Table 1 A classification of placement mechanisms based on their optimization goals

Placement mechanism algorithms	Energy optimization	Network usage	Constrained optimization
Lu et al. [31]	X		
Ying et al. [28]	X		
Rizou et al. [20]		X	X
Rizou et al. [16]		X	
Ottenwälder et al. [29]		X	
Pietzuch et al. [30]		X	
Bonfils et al. [24]	X		
Chatzimilioudis et al. [25]	X		
Chatzimilioudis et al. [27]	X		
Chatzimilioudis et al. [26]	X		
Starks et al. [32]	X		

4.3 Algorithm Design

The main goal of a placement mechanism is to find a placement assignment of an operator tree to networked nodes which optimally satisfies a predefined objective function subject to one or more constraints [19]. The information used to achieve this goal varies in terms of scope and variability. On the one hand, some placement algorithms have access to the entire network topology in addition to workload and resource availability information. This makes it possible to perform placement assignment in a centralized manner [26, 33–35]. In some edge networks such as MANETs, it has been shown that certain routing protocols such as OLSR are able to maintain a rather complete view of the network topology on each node [36]. This information is stored in the routing table and would be available to a placement mechnism for free, i.e., no extra messages need to be exchanged to use this information. On the other hand, some placement algorithms cannot assume knowledge of the entire network state and resource availability due to various reasons, like the costs are too high to maintain this information, or in delay tolerant networks it might take quite some time to get information from another network partition. These algorithms must perform placement assignment in a decentralized manner based on local information [16, 20, 23–31]. As such, the scope of the input data provided to the placement mechanism has a direct impact on its inherent structure. Centralized placement mechanisms rely on a single node with global information about the system to perform placement, while distributed placement algorithms rely on local information to gradually find an optimal placement for the operator tree. Parts of the input data for the placement mechanism are subject to change across time due to mobility and other reasons. Consequently, it is important for a placement mechanism to include an adaptation strategy in order to maintain the target system performance (see Sect. 4.4).

4.3.1 Centralized Placement Algorithm

Centralized placement mechanisms perform placement assignment of the entire operator tree on a single node, which is typically the sink [26, 33–35]. Consequently, the cost of query dissemination is considered insignificant and therefore ignored [26].

It is relatively easy and straightforward for a centralized placement approach to find a global optimal placement assignment [33]. However, such approaches do not scale well in large-scale scenarios even if global resource information is available. Therefore, in cases where network resources availability changes over time, the centralized approach can incur substantial communication overhead and delay and lead to the deterioration of the overall system performance. Consequently, some works apply a two step operator placement approach where the initial centralized placement assignment is iteratively updated towards a good respectively the optimal scheme [33].

To exemplify centralized placement mechanisms, we briefly present the core idea of two centralized placement algorithms introduced by Chatzimilioudis et al. [26].

The first algorithm basically analyses the entire search space for the optimal solution, which is guaranteed to be found. The algorithm uses dynamic programming to build a matrix of operators and all nodes in the network and systematically considers all possible placement assignments. This solution is obviously computationally demanding and inapplicable for large problems. To combat this scalability issue, Chatzimilioudis et al. propose a heuristic-based algorithm which is able to find a near optimal solution. The algorithm has a two stage approach, where the first stage is performed in a centralized manner and the second decentralized. In the first stage, an operator tree is built and used as input to the second stage. In the second stage, the placement of the operators in the evaluation tree is iteratively optimized in a top down manner. This algorithm assumes that each node performing operator placement has knowledge of the entire network. Their evaluation shows an improvement in total query processing cost of 10–95% compared to the naive approach. This is due to both the reduced communication cost and near optimal placement assignment from the heuristic based algorithm.

4.3.2 Decentralized Placement Algorithm

In a decentralized placement mechanism scheme, the placement assignment is performed based on local information shared between neighbor nodes. The scope of the local information varies from neighboring nodes (one hop neighbors for example) to an entire network cluster.

Some decentralized placement mechanisms start with an initial processing cost exchange between neighbors before proceeding with the actual placement assignment [28, 31]. In such schemes, all nodes in the network are participating in the cost information exchange. Additionally, approaches such as [27, 31] allow any node to directly broadcast their cost information in case the local resource availability changes or they can re-broadcast overheard cost information from neighbor node(s). Due to their reliance on flooding techniques, the communication cost for these approaches can quickly dwarf the incentives of in-network processing especially when the rate of change is too high due to mobility or other reasons. One approach which reduces the message overhead related to operator placement is presented in [23]. The proposed placement mechanism uses an area-restricted flooding mechanism in order to limit the number of network nodes involved in operator placement and therefore reduces the inherent message overhead. However, in a highly dynamic network environment, the need to synchronize cost information between neighbors in order to perform an optimal assignment can quickly incur a high message cost and even fail to converge.

Another approach suggested in [32] uses the location of the data sources that will host the leaves of the operator tree, to direct the distributed placement scheme. Only relevant candidates for processing a part of the operator tree participate in the placement scheme. Relevant candidates are those nodes that are part of the routes from the data sources to the sink(s). Moreover, the decision to place an operator on a specific network node does not require any synchronization between neighbor nodes. The

proposed placement mechanism assumes network knowledge, but it can easily be extended to support only local network information. The main goal of this algorithm is to incur as low overhead for the operator placement as possible and to be able to choose a good placement scheme. However, it does not need to be the optimal placement, because mobility will probably change (and mostly reduce) the performance of a selected operator placement scheme. Therefore, it is more important to have a light-weight operator placement algorithm, which in turn allows to perform a new operator placement with low costs, than to spend a lot of resource to find the optimal placement, which might be sub-optimal after a short time due to mobility.

The approach presented in [27] also aims to reduce the communication cost related to initial placement. The distributed techniques for operator placement achieve this aim by:

- identifying special cases where no flooding is needed to perform placement,
- limiting the size (number of nodes) of the neighborhood to be flooded

The core idea behind the algorithm is the concept of candidate nodes, i.e., physical host in the network, which are better suited to host a given operator. The candidate nodes are elected from a set of neighboring nodes in the network. The set of candidate nodes for a given operator is kept to the minimum (using a cost threshold) in order to limit the number of message exchanged of the network during placement information exchange between them. This effectively reduces the communication cost related to the placement of the operator.

The set of candidate nodes for an operator is created in a centralized manner without network communication, this allows the algorithm to detect special cases where there is no candidate node which is better suited to host the given operator. In this particular case (according to their experiments, 56–85% of the time, there is no candidate node which is better suited to host the given operator), there is no need to initiate the distributed operator host election algorithm. The radius for flooding during initial neighbor discovery is also limited, and it ensures that the optimal physical host for an operator can be found in the set of nodes that are part of the limited flooding. Results from experiments show a 50–100% reduction in the communication cost compared to naive flooding techniques.

4.4 Placement Adaptation

Mobile systems are inherently dynamic and changes of all kind can occur, like number of nodes, availability of links, resource availability, data rates, and many more. These changes are classified by [19] in three categories: changes concerning network infrastructures, changes concerning data characteristics, and changes concerning operator tree information. Any of these changes can have a negative impact on the performance of an operator network. A placement adaptation strategy aims to adjust the operator network after a change such that it fulfills again the application requirements.

Changes concerning the network infrastructure represent scenarios where the network topology changes due to node failure, mobile nodes, link failure (due to network congestion or node failure), or new node(s) joining the network [27, 29, 30]. There might also be changes in the local resources for a network node, e.g., the battery might be drained, or other computationally intensive software implies a high workload for a node [27, 37].

Changes conerning the network load happen for example when the data rates from sensors or source filters change, or other background traffic increases. For example, the increase in data rate at the input of one or more operators in the network might result in an increase in the total cost of in-network processing if bandwidth consumption is part of the objective function [24, 25]. The network load can also change due to new application traffic in the network or a change in data rate for other applications using the same network infrastructure.

Changes concerning the operator tree occur when the number of operators changes due to new queries submitted or previous ones are terminated. Additionally, the operator tree might be updated due to changes in the user's interest (location) requiring the adaptation of the corresponding operator network (see Sect. 5).

As the query processing scheme performance deteriorates, the placement adaptation strategy consists in picking new hosting node(s) for one or more operators in the flow graph. Two main approaches are identified in [27]: operator migration and placement update.

With operator migration the placement adaptation for an operator is performed by moving it from one node to another until an optimal placement assignment is found [24, 25, 29, 30, 38]. During operator migration, every node involved in the process uses local information exchanged between neighbors to determine which one of them is better suited to host the current operator. The limited scope of the information used makes the approach relatively easy. However, in a highly dynamic environment, it could be difficult for the migration process to converge towards an optimal or even good sub-optimal placement.

The placement update approach aims to find the best host for an operator immediately. This can be done in a centralized manner as in [30], or decentralized [27, 37] manner by reusing initial placement techniques for the single operator instance.

Different approaches are used to determine when to trigger the operator migration or placement update. One approach is to monitor the processing cost related to each operator and exchange this information between neighbors. When a predefined threshold is reached for a given operator, its migration process is triggered [24, 25, 27, 30, 37, 38]. Other approaches monitor constraints violations in addition to a predefined performance threshold based on the applied objective [29]. Another approach is to periodically trigger the placement adaptation of the entire operator tree based on a predefined time interval. In all cases, an operator migration or placement update will potentially trigger subsequent operator migration(s) or placement updates.

The cost of the actual migration or placement update should be worth the operator placement adaptation, which means that the increase performance of an adapted operator network gives higher benefits that the adaptation costs. This is not always

Table 2 Classification table for different adaptation scheme

Adaptation schemes	Monitored change			Adaptation techniques		Adaptation trigger	
	Network topology	Data rate	Logical graph	Operator migration	Placement update	Performance threshold	Constraints violation
Oikonomou et al. [38]	X			X		X	
Bonfils et al. [24]			X	X		X	
Chatzimilioudis et al. [25]			X	X		X	
Pietzuch et al. [30]		X			X	X	
Chatzimilioudis et al. [27]	X	X			X	X	
Z. Abrams et al. [37]		X			X	X	
Ottenwälder et al. [29]	X			X		X	X

the case as the placement adaptation process requires transferring the state information of all operators that are hosted on a new broker. This state information can be as large as several GBs [29]. As such, in certain scenarios, the migration of placement update for an operator might incur a significant cost, especially in terms of network usage. In some cases, however, the migration or placement update for an operator might be unavoidable, e.g., the battery of the hosting node will soon be depleted. It is also possible to experience a sort of freeze period during placement adaptation or operator migration. The freeze period occurs as the operator and its state are in transit from their previous host towards their new host [29].

To exemplify operator placement adaptation, we refer to the work by Pietzuch et al. [30]. In this work, a placement update is used to regularly solve the placement optimization problem for each unpinned operator. In particular, every network host regularly attempts to find a better placement assignment for each unpinned operator using local cost information exchanged between neighbor operator host in the operator network.

Predefined threshold(s) are used to determine whether an operator placement update should take place or not. One threshold determines when the difference in performance between the newly found optimal placement and the previous one is high enough to incentivize the placement update. Another threshold determines whether the cost of the placement update is low enough given the expected gains from the new placement assignment. Additionally, the longevity of the query to which the operator belongs is taken into consideration by the latter threshold in order to make sure the query will run long enough to amortize the cost endured by the operator

placement update process. The cost thresholds are used to ensure that both the network resources consumed and the operator placement update delay do not cripple the overall system performance. If the placement of operators is updated frequently, the adaptation cost might grow higher then performance gains. Additionally, if the placement of operators is updated for insignificant gains, the overall performance of the system might be degraded.

To evaluate the performance of the placement update scheme, 24 queries are created. The performance of the operator network for each query is evaluated two times, i.e., with adaptation enabled and without adaptation. Overall results show a 75% decrease in network usage (see Sect. 4.2.2) when operator adaptation is enabled. Finally, the aggregated query delay is reduced by 10.5% through adaptation.

While the results show clear gains in terms of both the application perceived performance and system resource consumption, the evaluation system model considered is rather simplistic compared to typical scenarios with Mobile Big Data.

The operator tree comprises only 3 nodes, which introduces some uncertainty whether the results are representative for large scale scenarios for Mobile Big Data. The migration rate experienced in the experiments is in average 3.5 adaptations per query, which indicates that the results are probably not representative for higly dynamic Mobile Big Data systems. However, evaluation of Mobile DCEP with placement adaptation is rather hard, because appropriate methodologies and tools are missing (see Sect. 6).

5 Mobile Queries

The topic of spatio-temporal data and mobile range queries has been extensively studied in the database community. The overall goal is to provide continuously updated information, typically to a mobile consumer, e.g., the five closest bus stops to the current location of the consumer. The survey by Ilarri et al. [39] gives an excellent overview of challenges and approaches to enable location-dependent query processing in traditional database settings, i.e., the data is materialized on secondary storage before processing. Traditional approaches to store, query, or index spatio-temporal data are insufficient to handle the high data rates and potentially very large data sizes in Mobile Big Data [40]. This insigth motivated researchers to combine the two worlds of traditional spation-temporal data management and Data Stream Management Systems. The systems [41] and [40] are to the best of our knowledge the first published DSMS with support for moving range queries. Research on CEP support for mobile range queries is still in its infancy and pioneering work is recently published in [14, 29, 42–44]; and to a larger part summarized in the Thesis presented by Ottenwalder [29]. Therefore, we base our description of challenges in mobile DCEP introduced by spatio-temporal data issues and new solutions on the terminology and model of [29, 43]. The overall goal of this work is to enable location based situational awareness for consumers in a mobile setting.

To achieve this situational awareness mobile CEP queries, called *MCEP queries*, they need to be registered at the MCEP system. A MCEP query Q has the following structure:

$$Q = \{G, fo, R, \delta, PoI\} \tag{7}$$

G represents an operator tree, fo is focal object of the consumer, R a function to calculate the spatial interest based on fo, δ a lifetime parameter, and PoI the delivery semantics. That means that in case of a mobile focal object fo the function R needs to be recalculated if fo has a new position to adapt the spatial interest, i.e., the region of interest. The function R is by purpose not defined in this model in order to enable regions of interest with arbitrary shapes. As such, a sequence on location updates from fo, i.e., $(l_1, l_2, l_3, l_4, l_5, \ldots)$ results in a sequence of changing spatial interests $(R_1, R_2, R_3, R_4, R_5, \ldots)$ where $R_i = R(l_i)$ for each $i \in \mathbb{N}$.

Ottenwaelder et al. [43] use the example of traffic awareness in which a consumer is driving a car and aims to avoid traffic jams. As such the consumer is interested in all accidents that happened within the last 30 min within 500 m of the consumers current location. In this case, the consumer or the consumer's car is the focal object fo which continuously reports location updates. The function R calculates each l_i a circle with a radius of 500 m and l_I as the center. The parameter δ in the query has the value 30 min.

A change of spatial interest from R_i to R_{i+1} requires to update the operator tree G accordingly since the set of sensors that are deployed in R_i and R_{i+1} is typically not equal and the sensors are represented as leaves in G. The update of the spatial interest and the following switch to a new operator tree introduces new challenges:

- For traditional CEP systems, the temporal order of events can be for many operators crucial to perform correctly. A change in spatial interest with a swicth of the operator tree implies that in mobile CEP with spatio-temporal data also the spatial order and spatio-temporal order is important. To achieve a spatially ordered event stream all events from R_i need to be delivered before events from R_{i+1} are delivered. A spatio-temporal order requires spatial event order and temporal event order for events from each R_i.
- The concepts of consistency and completeness need to be extended for MCEP. Spatial consistency ensures that all nodes in one operator tree process only input data that is based on one region of interest. This can be atomic events stemming from one particular region of interest or composite events that are based on these atomic events. Temporal completeness requires that situational information for one region of interest is delivered in spatio-temporal ordering for the temporal interest δ. Thus temporal completeness with a large δ leads to large latency.
- CEP operators can, in contrast to DSMS, be stateful. As such an operator cannot just proceed to process the incoming data after an operator switch. Instead, the operator state that was established when processing input data for R_i needs to be deleted, respectively the operator needs to be restarted.

- To detect events of interests in the new region R_{i+1}, historical events, i.e., those that happened before the operator tree switch are useful for two reasons: (1) a window over the input data needs to be filled up before the operator can start processing, which obviously introduces a start-up latency. If historical data is available, the window can be filled up much faster, which in turn reduces the start-up latency. (2) Historical events are useful for the consumer. In the traffic awareness example, accidents that happened in the new region of interest R_{i+1} before the switch to G_{i+1} are useful for the consumer, because roads will be congested for some time after the accident.

Any CEP system supporting mobile queries needs to know the location of data sources, consumers, and brokers. The MCEP system [43] comprises a location and performance monitor that continuously monitors the location of data sources and consumers. A location update of a consumer from l_i to l_{i+1} triggers a query configurator which initiates a switch to a new operator tree based on the new region of interest R_{i+1}. The data sources in R_i are instructed to stop streaming atomic events, and the set of data sources in R_{i+1} is identified and these data sources are instructed to start streaming atomic events. Please note that there is a high probability that the sets of data sources in R_i and R_{i+1} overlap. To achieve spatial consistency and spatio-temporally ordered results, so-called *markers* are inserted in the event streams. Markers are special messages that separate in each atomic event stream from the data sources that are in R_i and R_{i+1} the atomic events that are relevant for R_i and R_{i+1}. The arrival of a marker at an operator implies that now atomic events from a new region arrive. It is possible that the operator is still waiting for atomic events from the old region of interest to achieve completeness. Before processing the atomic events from the new region of interest, the operator is reset to avoid spatial inconsistencies. A marker is inserted in the operators' outgoing event stream before the first event from R_{i+1} is inserted. In this way, markers are inserted by each operator in the operator tree and enable spatial consistency and spatio-temporally ordered results for all operators. Several optimization techniques are leveraged in MCEP to produce timely results. In the proactive version of an operator switch, the future location of the focal object is predicted and the system starts to process historical results for a future region of interest. Once the focal object is in the predicted region of interest, all historical events are already available and processed. Another optimization is related to the overlapping sets of data sources of subsequent regions of interest and reuses the events for processing. The delivery semantics *PoI* are used to specify how to trade Quality of Information against streaming and processing costs.

Earlier work on the SOLE system [40] considers so-called regions of uncertainty which is caused by the fact that the system does not know about all data sources in a region of interest. The reasons for this uncertainty can be that new queries are installed and no historical data is available, as such it takes some time until the windows are filled and results can be produced. Furthermore, moving queries imply continuously changing regions of interest, which in turn leads to new data sources. This is similar for mobile data sources that move into a region of interest. To handle the last two cases of uncertainty, SOLE applies a caching strategy for all moving objects

that are predicted to be at some point in time in the region of interest. As such this solution is similar to the proactive approach in MCEP.

Query optimization is a classical research problem in databases and also investigated in spatio-temporal databases. Mobile queries have a huge potential for optimization. Consider the application for a car driver to get continuous information about traffic congestions in the vicinity of the driver. The fact that there is not a single car driver on the road, but instead a large amount of drivers introduces severe scalability issues. Each driver represents one focal object fo_i with its unique location l_i. Therefore, i mobile queries need to be executed and i can be very large considering the number of cars that are travelling on roads in major cities during rush hour. Two mechanisms are presented in [29] to address this scalability issue:

- Reuse of processing results: If all car drivers use the same or rather similar mobile queries to achieve situational awareness there is a substantial overlap of the operator graphs that are used to process these queries. The regions of interest of focal objects that are close to each other will also overlap substantially and as such the operators in the different operator graphs will process to a certain degree the same input events.
- Relax the requirement for a fully accurate set of input events to calculate situational awareness. This idea is similar to the well-known technique of load shedding in CEP. By relaxing the need for accurate input data it is possible to increase the number of operator graphs for different focal objects that can share the same set of input events.

The solution presented in [29] is based on a reuse-aware operator tree in which some operator graphs process input events on behalf of other operators and a system component called *selection manager*. The selection manager analyzes the degree to which input events of operators overlap. Given that the overlap is large enough with respect to a predetermined quality metric, the selection needs to be processed only once and can be reused for the other operator graphs.

Combining this kind of query optimization with operator placement could allow for more improvements if the computational complexity and the deployment costs could be kept low enough. However, this challenge is subject to future work.

6 Mobile DCEP Evaluation

The usefulness of mobile DCEP systems depends on their performance. Due to their large scale and complexity, their performance evaluation constitutes a challenge on its own, requiring the use of well-established, systematic methodologies. This section introduces the most commonly used techniques for performance evaluation of distributed systems, then summarize how these are used to evaluate mobile DCEP systems.

6.1 Basic Requirements and Approaches

We very briefly summarize the common approaches for performance evaluation of distributed systems. Consult [45] for a more elaborate treatment of the subject.

Before we describe the approaches, we explain the desirable properties of the evaluation approach and results. A basis for performance evaluation is sufficiently *representative, accurate and understandable* data that enables a proper analysis of the system under test. With experimental approaches, it is important that the experiments are *repeatable*, e.g., to enable third party verification of the results or to investigate the results of incremental system improvements. Results obtained from systems with similar purposes should furthermore be *comparable*, e.g., by applying similar models and/or (values of) parameters and metrics. To facilitate acquiring results with all these properties, the evaluation techniques and tools should require a *low effort and cost*.

The most representative results are obtained from real world experiments using workloads, configurations and environments similar to that expected during the final deployment. For large-scale and/or complex systems, this approach is often too expensive and time consuming. Instead, evaluation is performed using abstract mathematical or simulation models. The quality of model-based evaluation rests on how well the used model captures the characteristics of the system under test that determine its performance. When this cannot be achieved to a satisfactory degree with mathematical formula, due to system complexity or scale, simulation and emulation provides a popular alternative. Simulation is by far the most common evaluation approach in computer systems' evaluation due to its low cost, support for abstractions that facilitate understanding and a controllable experimentation environment. Emulation combines real and simulated components, e.g., running real applications on virtual network nodes, and thereby benefits from the advantages of both real and simulation experimentation. However, since emulation experiments involve real components, they also suffer from scalability limitations.

6.2 Performance Evaluation for Mobile DCEP

This section summarizes 19 performance evaluation reports presented in 13 key publications on mobile DCEP [14, 16, 18, 20, 23–28, 30, 32, 46] in terms of the applied approach, tools, parameters, and metrics.

All publications include at least one simulation study, except for two publications based on emulation that employ both simulated and real components [20, 32]. As a result, 14 of 19 evaluation reports are based on either simulation (11 reports) or emulation (three reports). Of the remaining five reports, three are based on real world experiments (in [18, 30]), e.g., with the well-known PlanetLab test bed [47], and two are based on mathematical analysis (in [14, 23]). The fact that simulation is the most popular approach is not surprising since a proper evaluation of placement algorithms

typically requires networks with several hundred nodes, making real-world experiments unfeasible. This problem is further exacerbated for wireless networks where the shared medium and node mobility implies a high degree of network dynamicity, making it exceedingly difficult to conduct controlled and repeatable experiments. We find that seven of the simulation experiments are performed with simulators that are created for the specific experiments at hand, and that the remaining four experiments are performed with the popular network simulators J-Sim [48], OMNeT++ [49] and PeerSim [50] using real world or generated topologies, mobility patterns and network traffic as input. Our survey indicates that there exists no common simulation platform to enable the evaluation of mobile DCEP systems in general.

Some reports involve parameters and metrics that cannot readily be found in other reports, e.g., model-specific paramters like the α in [23] and metrics like the *stretch factor* in [16]. For results from such reports to be comparable with those for other similar systems, such parameters and metrics must first be translated. This might be cumbersome or even impossible, limiting the comparability of results. There are, however, metrics that are widely used within a subset of the reports. For instance, solutions for WSN [23–28, 46] address the common challenge of resource constraints on nodes and are thus typically evaluated in terms of energy consumption and/or processing cost. Comparability is however somewhat limited by the fact that the metrics can be defined slightly different between works, or because the applied parameter types and values differ significantly from study to study. For example, the number of nodes in the simulated networks ranges from less than 10 [46] to several hundred [25, 27]. The most common metric for the remaining six works [14, 16, 18, 20, 30, 32] is *network usage* based on the bandwidth-delay product presented in Sect. 4.2.2. This is nevertheless only used in three of these six works [14, 18, 30] and can therefore hardly be considered as a common ground for comparison. In general, we cannot identify any clear consensus in terms of metrics and parameters that facilitate comparability among different mobile DCEP systems.

6.3 Future Work on the Evaluation of Mobile DCEP

Due to factors such as cost and effort, the de-facto standard evaluation approach for mobile DCEP is simulation, mostly with simulators created for the paper at hand. There are several disadvantages of this extensive use of custom simulators. First, their models are not subjected to the rigorous validation that models in more general purpose simulators are subjected to. Examples of such general purpose simulators include the widely-used, de-facto standard simulators used in the networking community, i.e., Ns-2, Ns-3, and OMNeT++. These popular simulators have been available for decades during which their models have been subjected to continuous validation to assess realism and comparability. Second, for the experiments to be repeatable, the custom simulators need to be very simple to allow sufficient, yet brief description in an evaluation report. This problem is exacerbated by the fact that the simulators are rarely made available for download online. To accomplish this, the simulations

are often based on overly simplified assumptions, e.g., not accounting for complex network phenomena like link interference and bit-error rates, which in turn affects the credibility of the results. Comparability is also compromised since different simulators are based on different models, parameters and metrics that produce results that cannot readily be compared. In contrast, the above mentioned network simulators Ns-2, Ns-3 and OMNeT++ are freely available for download online, and are maintained by a well-established, code review-based developer community that provides a channel through which researchers can contribute and distribute their models world-wide. The mobile DCEP community is a relatively new one compared to the networking community. This might be the reason behind the lack of a corresponding de-facto standard simulator for mobile DCEP. Our findings suggest that such a generic DCEP-simulator, facilitating the evaluation of a wide variety of mobile DCEP solutions, would help improve the quality of the simulation results in terms of *repeatability* and *comparability*. Since the performance of DCEP is largely affected by the characteristics of computer networks, models for a DCEP-simulator would benefit in terms of *accuracy* and *realism* from the reuse of these computer network models. We argue that this is best approached by extending existing network simulators with DCEP-models, rather than vice-versa, in order to benefit from the large base of models, knowledge and support available in the network simulation community.

7 Summary and Conclusion

CEP is a promising technology to enable situational awareness in real-time in the Internet of Things, because it provides a declarative interface to mobile DCEP application programmers, abstracting away data processing intricacies in the distributed environment. Therefore, we are convinced that mobile DCEP can play an important role in future Mobile Big Data systems.

However, mobile DCEP need to handle unstable infrastructure with limited resources, because mobility implies the use of wireless networking technologies with potential bandwidth limitations, dependency on battery lifetime in mobile devices, and a dynamic network topology. Consequently, effective data handling and efficient resource consumption is a prerequisite for such systems. The most important mechanism to handle these issues in mobile DCEP and to meet application QoS requirements is operator placement.

The classification of the state-of-the-art in operator placement in Sect. 4.2.4 shows that most works aim to minimize system resource consumption such as energy and bandwidth. Some researchers address application QoS requirements such as low latency together with efficient system resource consumption. Including constraints in the operator placement, like security concerns, is in its infancy and the potential that operator placement has for privacy protection has to the best of our knowledge not been addressed yet. A lot of decentralized placement mechanism exists, but very few address issues related to a dynamic topology in mobile DCEP. Consequently, none

of the solutions proposed is applicable in highly dynamic environments. A lot of work has been done in enabling placement adaptation to deal with change in DCEP systems. The adaptation strategies vary by the monitored change, adaptation techniques applied and adaptation triggers. The network topology, data rate and change in the operator tree are the main elements monitored to determine when it is time to trigger adaptation using predefined performance threshold. Some works consider constraints violation as a means to ensure application QoS through adaptation. Two main adaptation techniques are applied: operator migration and placement adaptation. It is our belief that, to enable QoS for mobile DCEP applications, it is necessary to further study constrained violation based adaptation triggers.

Operator placement itself is also not sufficient for mobile environments, since mobility of devices, including brokers hosting operators, can render an initial and near optimal placement in short term sub-optimal or even result in a very inefficient system. Therefore, an efficient placement adaptation is required to ensure the performance and efficiency of the system.

Mobility does not only cause challenges at the infrastructure level, but also at the user and application level. For example, mobile consumers of location-based services are often interested in events in their vicinity. Due to mobility the region of interest and the sensors in these regions continuously change. To support mobile queries in CEP for this kind of applications mobile DCEP needs to properly handle dynamic data sources or producers, dynamic or mobile range queries, spatio-temporal event ordering, spatial consistency and temporal completeness, CEP operator state transition and management, and location awareness.

There is currently very limited work on mobile queries for mobile DCEP. More research needs to be done to explore all the issues introduced by mobile consumers with location based interests. Furthermore, operator placement techniques need to address challenges introduced by mobile queries in order to ensure efficient and effective event processing networks.

Mobile DCEP is mostly evaluated with either simulation or emulation. The parameters used vary across evaluation reports, making it difficult to compare them. Custom simulators are used for evaluation, making them less reliable compared to more established simulators. Furthermore, the custom simulators are rarely made available to reproduce the results from the evaluations. It appears that the DCEP research community would benefit from a generic DCEP-simulator facilitating the evaluation of a wide variety of mobile DCEP solutions. Such a simulator would enable repeatable experiments with results that are representative of the simulated system as well as comparable with results from experiments with other systems conducted with the same simulator. Additionally, reusing already existing and mature simulation tools from the networking community would ensure accuracy and realism of mobile DCEP evaluations due to their large scale networking characteristics.

Even thought there are many open research issues in mobile DCEP, this chapter shows that results and useful systems for certain scenarios exist. Since mobility can

have many forms it will probably turn out in the future that it is not possible to design one mobile DCEP system that can handle all the challenges which are caused by mobility. Instead, proper solutions for cases in which only the edge network is mobile might be earlier ready than proper mobile DCEP solutions for infrastructure less networks.

References

1. Golab, L., Özsu, M.T.: Issues in data stream management. SIGMOD Rec. **32**(2), 5–14 (2003). https://doi.org/10.1145/776985.776986
2. http://sqlstream.com/intro. Accessed 30 Dec 2016
3. Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., Motwani, R., Srivastava, U., Widom, J.: Stream: The stanford data stream management system. Technical Report 2004-20, Stanford InfoLab (2004). http://ilpubs.stanford.edu:8090/641/
4. Abadi, D.J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., Zdonik, S.: Aurora: a new model and architecture for data stream management. VLDB J. **12**(2), 120–139 (2003). https://doi.org/10.1007/s00778-003-0095-z
5. Stegmaier, B., Kuntschke, R., Kemper, A.: Streamglobe: adaptive query processing and optimization in streaming p2p environments. In: Proceeedings of the 1st International Workshop on Data Management for Sensor Networks: In Conjunction with VLDB 2004, DMSN '04, pp. 88–97. ACM, New York, NY, USA (2004). https://doi.org/10.1145/1052199.1052214
6. http://www.espertech.com/esper/. Accessed 30 Dec 2016
7. Kazemitabar, S.J., Demiryurek, U., Ali, M., Akdogan, A., Shahabi, C.: Geospatial stream query processing using microsoft sql server streaminsight. Proc. VLDB Endow. **3**(1–2), 1537–1540 (2010). https://doi.org/10.14778/1920841.1921032
8. http://evam.com/platform/. Accessed 30 Dec 2016
9. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '02, pp. 1–16. ACM, New York, NY, USA (2002). https://doi.org/10.1145/543613.543615
10. Luckham, D.C.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA (2001)
11. Cugola, G., Margara, A.: Processing flows of information: from data stream to complex event processing. ACM Comput. Surv. **44**(3), 15:1–15:62 (2012). 10.1145/2187671.2187677. http://doi.org/10.1145/2187671.2187677
12. Eugster, P.T., Felber, P.A., Guerraoui, R., Kermarrec, A.M.: The many faces of publish/subscribe. ACM Comput. Surv. **35**(2), 114–131 (2003). https://doi.org/10.1145/857076.857078
13. Etzion, O., Niblett, P.: Event Processing in Action, 1st edn. Manning Publications Co., Greenwich, CT, USA (2010)
14. Koldehofe, B., Ottenwälder, B., Rothermel, K., Ramachandran, U.: Moving range queries in distributed complex event processing. In: Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems, DEBS '12, pp. 201–212. ACM, New York, NY, USA (2012). https://doi.org/10.1145/2335484.2335507
15. Zhang, B., Mor, N., Kolb, J., Chan, D.S., Lutz, K., Allman, E., Wawrzynek, J., Lee, E., Kubiatowicz, J.: The cloud is not enough: saving iot from the cloud. In: 7th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 15). USENIX Association, Santa Clara, CA (2015). https://www.usenix.org/conference/hotcloud15/workshop-program/presentation/zhang

16. Rizou, S., Durr, F., Rothermel, K.: Solving the multi-operator placement problem in large-scale operator networks. In: 2010 Proceedings of 19th International Conference on Computer Communications and Networks (ICCCN), pp. 1–6. IEEE (2010)
17. Rizou, S., Durr, F., Rothermel, K.: Providing qos guarantees in large-scale operator networks. In: 2010 12th IEEE International Conference on High Performance Computing and Communications (HPCC), pp. 337–345 (2010). https://doi.org/10.1109/HPCC.2010.53
18. Cardellini, V., Grassi, V., Lo Presti, F., Nardelli, M.: Optimal operator placement for distributed stream processing applications. In: Proceedings of the 10th ACM International Conference on Distributed and Event-based Systems, DEBS '16, pp. 69–80. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2933267.2933312
19. Lakshmanan, G.T., Li, Y., Strom, R.: Placement strategies for internet-scale data stream systems. IEEE Internet Comput. **12**(6), 50–60 (2008). https://doi.org/10.1109/MIC.2008.129
20. Rizou, S., Diirr, F., Rothermel, K.: Fulfilling end-to-end latency constraints in large-scale streaming environments. In: 30th IEEE International Performance Computing and Communications Conference, pp. 1–8 (2011). https://doi.org/10.1109/PCCC.2011.6108086
21. Cipriani, N., Lbbe, C., Moosbrugger, A.: Exploiting constraints to build a flexible and extensible data stream processing middleware. In: 2010 IEEE International Symposium on Parallel Distributed Processing, Workshops and Ph.D. Forum (IPDPSW), pp. 1–8 (2010). https://doi.org/10.1109/IPDPSW.2010.5470847
22. Anastasi, G., Conti, M., Francesco, M.D., Passarella, A.: Energy conservation in wireless sensor networks: asurvey. Ad Hoc Netw. **7**(3), 537–568 (2009). https://doi.org/10.1016/j.adhoc.2008.06.003, http://www.sciencedirect.com/science/article/pii/S1570870508000954
23. Lu, Z., Wen, Y., Fan, R., Tan, S.L., Biswas, J.: Toward efficient distributed algorithms for in-network binary operator tree placement in wireless sensor networks. IEEE J. Sel. Areas Commun. **31**(4), 743–755 (2013)
24. Bonfils, B.J., Bonnet, P.: Adaptive and decentralized operator placement for in-network query processing. Telecommun. Syst. **26**(2–4), 389–409 (2004)
25. Chatzimilioudis, G., Cuzzocrea, A., Gunopulos, D., Mamoulis, N.: A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement. J. Comput. Syst. Sci. **79**(3), 349–368 (2013)
26. Chatzimilioudis, G., Hakkoymaz, H., Mamoulis, N., Gunopulos, D.: Operator placement for snapshot multi-predicate queries in wireless sensor networks. In: 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, pp. 21–30. IEEE (2009)
27. Chatzimilioudis, G., Mamoulis, N., Gunopulos, D.: A distributed technique for dynamic operator placement in wireless sensor networks. In: 2010 Eleventh International Conference on Mobile Data Management, pp. 167–176. IEEE (2010)
28. Ying, L., Liu, Z., Towsley, D., Xia, C.H.: Distributed operator placement and data caching in large-scale sensor networks. In: INFOCOM 2008. The 27th Conference on Computer Communications. IEEE (2008)
29. Ottenwälder, B., Koldehofe, B., Rothermel, K., Ramachandran, U.: Migcep: operator migration for mobility driven distributed complex event processing. In: Proceedings of the 7th ACM International Conference on Distributed Event-based Systems, pp. 183–194. ACM (2013)
30. Pietzuch, P., Ledlie, J., Shneidman, J., Roussopoulos, M., Welsh, M., Seltzer, M.: Network-aware operator placement for stream-processing systems. In: 22nd International Conference on Data Engineering (ICDE'06), pp. 49–49. IEEE (2006)
31. Lu, Z., Wen, Y.: Distributed and asynchronous solution to operator placement in large wireless sensor networks. In: Proceedings of the 2012 8th International Conference on Mobile Ad-hoc and Sensor Networks, MSN '12, pp. 100–107. IEEE Computer Society, Washington, DC, USA (2012). https://doi.org/10.1109/MSN.2012.23
32. Starks, F., Plagemann, T.P.: Operator placement for efficient distributed complex event processing in manets. In: 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 83–90 (2015). http://doi.org/10.1109/WiMOB.2015.7347944

33. Jain, N., Biswas, R., Nandiraju, N., Agrawal, D.P.: Energy aware routing for spatio-temporal queries in sensor networks. In: IEEE Wireless Communications and Networking Conference, 2005, vol. 3, pp. 1860–1866 (2005). https://doi.org/10.1109/WCNC.2005.1424795
34. Pathak, A., Prasanna, V.K.: Energy-efficient task mapping for data-driven sensor network macroprogramming. IEEE Trans. Comput. **59**(7), 955–968 (2010). https://doi.org/10.1109/TC.2009.168
35. Srivastava, U., Munagala, K., Widom, J.: Operator placement for in-network stream query processing. In: Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '05, pp. 250–258. ACM, New York, NY, USA (2005). https://doi.org/10.1145/1065167.1065199
36. Drugan, O., Plagemann, T., Munthe-Kaas, E.: Dynamic clustering in sparse MANETs. Computer Communications **59**, 84–97 (2015). https://doi.org/10.1016/j.comcom.2014.12.005, http://www.sciencedirect.com/science/article/pii/S0140366414003703
37. Abrams, Z., Liu, J.: Greedy is good: on service tree placement for in-network stream processing. In: 26th IEEE International Conference on Distributed Computing Systems (ICDCS'06), pp. 72–72 (2006). https://doi.org/10.1109/ICDCS.2006.45
38. Oikonomou, K., Stavrakakis, I., Xydias, A.: Scalable service migration in general topologies. In: Proceedings of the 2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks, WOWMOM '08, pp. 1–6. IEEE Computer Society, Washington, DC, USA (2008). https://doi.org/10.1109/WOWMOM.2008.4594891
39. Ilarri, S., Mena, E., Illarramendi, A.: Location-dependent query processing: where we are and where we are heading. ACM Comput. Surv. **42**(3), 12:1–12:73 (2010). https://doi.org/10.1145/1670679.1670682
40. Mokbel, M.F., Aref, W.G.: Sole: scalable on-line execution of continuous queries on spatio-temporal data streams. VLDB J. **17**(5), 971–995 (2008). https://doi.org/10.1007/s00778-007-0046-1
41. Xiong, X., Elmongui, H.G., Chai, X., Aref, W.G.: Place: a distributed spatio-temporal data stream management system for moving objects. In: 2007 International Conference on Mobile Data Management, pp. 44–51 (2007). https://doi.org/10.1109/MDM.2007.16
42. Hong, K., Lillethun, D.J., Ramachandran, U., Ottenwälder, B., Koldehofe, B.: Opportunistic spatio-temporal event processing for mobile situation awareness. In: The 7th ACM International Conference on Distributed Event-Based Systems, DEBS '13, Arlington, TX, USA—June 29–July 03, 2013, pp. 195–206 (2013). https://doi.org/10.1145/2488222.2488266
43. Ottenwälder, B., Koldehofe, B., Rothermel, K., Hong, K., Lillethun, D.J., Ramachandran, U.: MCEP: A mobility-aware complex event processing system. ACM Trans. Internet Techn. **14**(1), 6:1–6:24 (2014). https://doi.org/10.1145/2633688
44. Ottenwälder, B., Koldehofe, B., Rothermel, K., Hong, K., Ramachandran, U.: RECEP: selection-based reuse for distributed complex event processing. In: The 8th ACM International Conference on Distributed Event-Based Systems, DEBS '14, Mumbai, India, May 26–29, 2014, pp. 59–70 (2014). https://doi.org/10.1145/2611286.2611297
45. Jain, R.: The art of computer systems performance evaluation. In: Limoncelli, T., Hogan, C., Chaiup, S. (eds.), The Practice of System and Network Administration, vol. 3, pp. 978–032, lSBN-I. Wiley (1991)
46. Kakkad, V., Santosa, A.E., Scholz, B.: Migrating operator placement for compositional stream graphs. In: Proceedings of the 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 125–134. ACM (2012)
47. Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., Bowman, M.: Planetlab: an overlay testbed for broad-coverage services. ACM SIGCOMM Comput. Commun. Rev. **33**(3), 3–12 (2003)
48. Kačer, J.: Discrete event simulations with j-sim. In: Proceedings of the Inaugural Conference on the Principles and Practice of programming, 2002 and Proceedings of the Second Workshop on Intermediate Representation Engineering for Virtual Machines, pp. 13–18. National University of Ireland (2002)

49. Varga, A., et al.: The omnet++ discrete event simulation system. In: Proceedings of the European Simulation Multiconference (ESM2001), vol. 9, p. 65. sn (2001)
50. Montresor, A., Jelasity, M.: Peersim: a scalable p2p simulator. In: 2009 IEEE Ninth International Conference on Peer-to-Peer Computing, pp. 99–100. IEEE (2009)

Electromagnetic Interference and Discontinuity Effects of Interconnections on Big Data Performance of Integrated Circuits

Seyi Stephen Olokede and Babu Sena Paul

Abstract An antenna-in-package solution has recently been the ultimate technology offering innovation and perhaps the most highly integrated radio miniaturization surface-mounted chipset device for short-range, high-speed, high-gain, and large-scale big data hyper-performance server platforms. Electromagnetic interference (EMI) arises as a result of discontinuity of the interconnections between the antenna and the integrated circuit (IC) chips, which limits their efficiency considerably, as it increases the mutual coupling and initiates and propagates surface waves, thus limiting the radiation efficiency in particular at the far-field. The backplanes, on which the IC boards containing data communication chips and processors are densely installed, are interconnected with high-speed integrated transceiver circuits using wire traces and connectors. As a result, transmission losses become considerable, in particular for backplanes operating at transfer speeds greater than 10 Gbps. In effect, the signal distortion becomes so significant that accurate data transmission without distortion is near impossible. Techniques to ameliorate the drawbacks of the side effects of parasitics are investigated in this chapter. Existing solutions to mitigate such effects are assessed to determine the extent of their efficacy. Alternative coupling techniques are examined. The effects of grounding, filtering, guard rings, shielding and decoupling are studied. The implication of process technology in eliminating EMI is also examined.

S.S. Olokede (✉) · B.S. Paul
Faculty of Engineering and the Built Environment,
Department of Electrical and Electronic Engineering Technology,
University of Johannesburg, Doornfontein Campus, Johannesburg 2028,
South Africa
e-mail: s.s.olokede@ieee.org; solokede@gmail.com

B.S. Paul
e-mail: bspaul@uj.ac.za

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_8

1 Introduction

Data-intensive technologies, otherwise known as big data, are currently the technology of choice, necessitated by present and future needs brought about by the data explosion with an unprecedented increase in diverse data sources. Big data technologies are currently the board-level technologies, where an enormous volume of data traffic is generated through a deluge of data driven by many available smartphones, mobile video and video streaming, tablets, social media, remote sensing, global positioning systems, smart metering, smart cities, the internet, healthcare, a wide range of sensors, connected cars, user-generated content, machine-to-machine communication, etc. A new lexicon of units to measure memory capacity, such as petabyte and even exabyte, is being used as everyday terminology. The massive volume of data may even grow to hundreds of exabyte from the proliferation and confluence of these data sets emanating from the deluge of diverse information sources. Thus, the rapid proliferation of these massive data sets is driven by data handling and transfer, in particular over internet protocol in the order of 100 times of that obtainable in 2010, with a projected volume growth to about half a million exabyte by 2020 [5], with a projected constant growth rate of between 20 and 40% [8].

Advances in complementary metal-oxide semiconductor (CMOS) technologies and their derivatives have paved the way for the recent information revolution via highly integrated circuits (ICs) to support high-density, high-performance, large-scale multiprocessor-based data centers in a bid to meet the constant demand for multi-gigabit applications. The fast and massive data system transfer, internet access workloads, big data applications and diversity vis-à-vis its complexity, etc., have necessitated the current migration to millimeter wave (mm-wave) band with excellent throughput rates and data streams in order to satisfy requirements for huge bandwidth, mobility and low-cost devices. It is evident therefore that integrated mm-wave communication technology offers the largest amount of available spectral bandwidth of about ~10 GHz with robustness for greater data rates. The dynamics of silicon (Si) technologies' operation with respect to Si germanium (SiGe) bipolar CMOS and CMOS at mm-wave frequencies, originally reserved for III–V compound semiconductors, has engendered enthusiasm in recent times. They are less sensitive to interconnect parasitics as a result of complementary transistor terminal impedances for a given power consumption and towering transconductance. To meet consumer marketplace requirements, alternative but efficient solutions must be demonstrated to be realizable, cheap and compact in order to effect mass deployment with higher transmission speeds and longer transmission channels for the large-scale data center transceivers.

The antenna-in-package (AiP) alternative is a contemporary, unconventional success in wireless systems miniaturization. Unlike antenna-on-package (AoP), AiP has been identified as the utmost auspicious antenna alternative for extremely densified integrated mm-wave applications, albeit in short-range but very high-speed wireless communications. It alleviates the difficulties in interconnections between

the chip and antennas, as well as the motherboard. The interconnection can be accomplished by using either the flip-chip or wire-bonding technique. As technology progresses, it becomes necessary to isolate various active and passive elements from one another in the IC structure, especially since the integration of dissimilar signals requires large isolation between them. Isolation problems, such as the widely observed and understood latch-up phenomena fostering crosstalk between active circuits and antennas, have become issues of concern. Because of poor isolation of the Si substrate, the isolation properties of the Si substrate result in capacitive and resistive parasitics that decrease the speed of the transistor. The AiP mechanism of placing antenna elements in an arrangement of stacked chips and thus feeding them through vias reduces the complexities of EMI as a result of discontinuities along these interconnections. Nonetheless, the distributed magnetic and electric fields are thus experienced along the interconnecting vias, the effects of which become appreciable at mm-wave band, and they consequently suffer from fringing field effects and losses due to abrupt changes in terms of short/open stubs, junctions and bends. In effect, signal distortions escalate with longer transmission channels and higher transmission speeds. As a result, the link budget limit at about 10 GHz communications is less than 1 m in existing multichannel transceivers, which makes it extremely difficult to extend the distance of transmission, in particular as regards high-speed transceivers. It also poses a serious challenge of slow response to calls from large-scale servers, causing severe difficulty.

For big data to be sustained with respect to the internet-of-things (IoTs), backplanes must be implemented in such a way that transfer speeds of no less than 10 Gbps are sustained. The interconnections must be implemented such that discontinuities are minimized at device level to avoid low speed, signal distortion and inability to recover clock-signal components correctly. In this chapter therefore, we intend to develop electromagnetic (EM) code to model and characterize interconnections' discontinuity. We intend to investigate their response to excitations further, and to determine to what degree they can inhibit signal dispersion. The chapter is written based on two fundamentals, namely the interrelatedness of big data with systems-on-chip (SoCs), and the second part that examines the propagations delay response occurring on ICs due to interconnection parasitic and EM interactions. The chapter therefore investigates the effects of interconnect parasitic and EM interactions on propagation delay response with respect to big data highly integrated large-scale ICs platforms. The concept of SoC with respect to big data is introduced in detail, and a mathematical model to predict propagation delay based on an *RC* model is examined. While the *RLC* model has been gaining popularity in recent times to predict the effect of parasitics on ICs' performance, this chapter instead focuses on the *RC* model since the model is sufficient to predict the effects of capacitive coupling and resistance on the delay time approximately. The intent is to investigate the interconnections parasitic effect on the big data performance of ICs to ascertain whether there is indeed a substantial delay. If the delay is substantial, then signal distortion occurs, in particular where the propagation signal is the communication of diverse and massive data across highly densified large-scale servers that are built based on SoC technology.

2 Big Data and System-on-Chips

The proliferation of big data with an unprecedented explosion of data traffic as a result of diverse domain sources will definitely require flexible, robust and decentralized network architectures to guarantee adequate and quick-response storage, process, request, and query, for it to be sustainable and efficient. Though the big data six latest developments industrial chain (namely data source or extraction using sensor data; aggregation-based open source infrastructure using either NoSQL or NewSQL data analysis and mining via unstructured data analysis, or/and data visualization; cross-platform infrastructure; and finally, marketing and advertising marketing applications [29]) comprises core technologies upon which big data and its applications are based, its efficiency and sustainability may be contingent on the response and speed of the densely integrated heterogeneous interconnect sensors and high-speed computer servers. Data acquisitions requested from these hyper-densified heterogeneous servers demand low latency roundtrip and ultra-fast arrival time from the internet access workloads and massive data technologies in order to support this deluge of big data. Moreover, the complexities of the massive interconnected *RF* hardware chains simultaneously serving these densified servers put more pressure on the possibility of relaxed latency, and much more on energy consumption complexity.

The situation becomes more precarious as this massive RF hardware chain increases in magnitude. Efforts are made at software level in connection with communication capabilities with the intent to achieve relaxed latency, in particular using many available and even novel machine learning algorithms to enhance the speed of the highly densified multi-servers. However, inadequacies such as parallel slowdown and race condition effects with respect to parallel programming rather than the traditional human brain sequential way of thinking, render the software solution inefficient [7]. Ultra-high density integrated multi-functional Si SoCs enabled essentially to power big data technologies, backplanes, and IoTs to satisfy sufficient bandwidth, latency requirements and avoid signal distortions are necessities. While extreme densification of heterogeneous networks could be beneficial for bandwidth enhancement, the need for more bandwidth cannot be over-emphasized even in the face of offloading procedures for optimal performance [5]. Little research has been reported at the level of hardware enhancement and systemic performance efficiency.

3 Millimeter Wave

The mm-wave band provides promising wireless digital interface data rate capabilities in excess of 20 Gbps [12]. The vast available bandwidth in this idle mm-wave band is more than the sum total of all other licensed spectra available for wireless communication [2]. This band has thus been standardized to facilitate a

robust platform for the next generation of wireless multimedia applications. The wavelength specifics (1–10 mm) of this band with attendant reduced form factor support densification of low power, high-speed, ultra-high ICs and sensor networks to support big data communication. As advances in SiGe and CMOS technologies progress, highly integrated multi-chip alternative solutions in Si become progressively realizable, thus making integrated mm-wave technology more attractive. Indeed, the commercial success of SiGe technology can be attributed to its ability to provide a very competitive integrated radio solution on a single chip using the high-volume manufacturing capability of a standard CMOS processing fabricator. This ever increasing, rapidly growing, expanding, and evolving wireless world with new high-speed demands and technological breakthroughs has had a huge impact on the market, such that laptops, personal computers, printers, cellphones, and voice-over-internet-protocol phones, MP3 players at homes, in offices and even in public areas are incorporating this wireless technology. It is then evident that the exorbitant cost of deployment is not unconnected with the historical expensive cost to implement the hardware operating at mm-wave frequencies.

Moreover, the systemic RF front end has been considered a substantial sphere of risk with respect to yields and fabrication tolerances, because of the high costs of implementation, semiconductor technology/chipset and packaging. Nonetheless, technology breakthroughs, development and advances may be consolidated to ensure cost cutbacks in semiconductor technology, packaging and implementation. Since this technology is gradually becoming ubiquitous considering the high-frequency application domain, the implication is that it is already on a declining cost curve. Besides, current advances in semiconductor technology could be supplemented for low-cost alternatives to warrant mm-wave wideband wireless applications' costs. The unending advances in integrated mm-wave packaging technologies are driving different innovative and packaging alternatives that are in turn driving low-cost solutions.

Si IC implementations of emerging 3D IC, ultra-high density catch memory on multiple layers, and radio standards are promising smaller, lower cost next generation chips enabling mobility and feature-rich big data robust capability [31]. SoC, which also consists of highly integrated AoP and AiP, is the most important technology for next generation big data technologies. SoC makes it possible to integrate all required circuits and devices on a chip, such that smaller and lower-power tuners offer tremendously attractive solutions, while substantially driving down costs. Today high-density, high-performance, large-scale servers require large SoCs integrating multi-Gbps transceivers with clock data recovery to handle large-volume multimedia data [56]. This requires high operating performance and high package density with low cost and low power consumption. High operating performance and flexibility for many applications demand extremely low power consumption in spite of high throughput operation. Uses of dedicated operation engines, vector pipelining operation and parallelism with low voltage operation are well-known techniques to ensure low power yet high throughput processing.

4 MIMO Based on Millimeter Wave Technology

Multiple input, multiple output (MIMO) is particularly beneficial at mm-wave band because of the smaller form factor as a result of the extremely short wavelength specificity of this frequency band. Consequently, massive MIMO systems are not only realizable in this band, but are reasonable as the antenna form factor dwindles with respect to frequency. By implication therefore, highly integrated arrays of a substantial bandwidth-gain product can be realized with a considerable reduction in aperture size, in particular when compared with the microwave band of frequency, since for instance the wavelength of a 60 GHz antenna is about 5 mm. According to estimates, close to 70 antennas can be co-located on a single die, such that close to 200 antennas can be co-located on a serial array of about 0.5 m at a center resonance of 60 GHz. Implementing MIMO technology based on mm-waves creates an opportunity to co-design the antenna, the chip and the package such that the single-chip radio combines an antenna with a single-chip die in order to evolve a benchmarked surface installed device. Alternatively, it may be regarded as an antenna integrated in a chip package. When highly IC technology is implemented, both the board area of the antenna and the assembly cost can be saved [58]. The performance of the resulting single-chip radio is thus maximized, with increasing mass deployment capabilities. The benefits of this technology are compactness and cost-effectiveness.

Such electrically small antennas exhibit very poor radiation efficiency, a substantial quality factor (Q-factor) and low available power [9, 53]. Substantial radiation losses due to conductive current and substrate absorption and the near effect of metal structures in the vicinity of the antennas (which influences the phase, the magnitude and input impedance of the received signals) are other challenges [17, 19, 26, 55]. There have been many challenges regarding system packaging (due to high-resolution photo-lithography, accurate alignment, or high-precision machining), simulation, physical realization, design, integration, complete system testing, etc. Nevertheless, the advantages of a mm-wave single-chip integrated antenna still drive a substantial amount of research to determine optimal alternatives to solving these performance challenges. Tsutsumi et al. [52], for instance, introduced a three-dimensional 60 GHz triangular antenna chip with the output and input pad connected to a metal pad mounted on a substrate with bonding wires. The result offered remarkably improved radiation efficiency as a result of the distance created between the strong electric current and the chip.

5 Internet-of-Things

It is certain that the many existing available traditional software tools and hardware environment are grossly insufficient to accommodate, manage, retrieve and process the large amount of data required by big data technologies. Efficient conventional

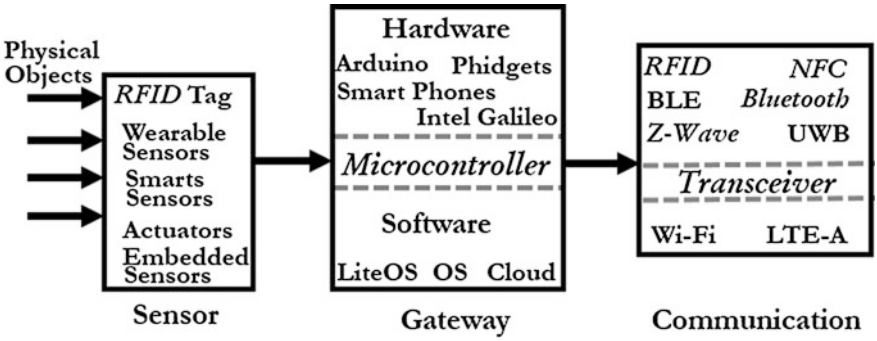


Fig. 1 Basic IoT system technology

sensors and massive heterogeneous smart sensors could be beneficial by providing additional and complementary process and storage platforms [34]. Interconnections of several millions of internet-enabled smart sensors, otherwise known as IoTs with low roundtrip latency, and minimum power consumption could serve as complementary but additional alternatives to support the existing software and hardware tools. Figure 1 depicts the basic architecture of IoTs' technology. The immediate focus of this work is to examine the communication implications of IoT technology, with a view to investigate the many available efficient alternatives for transmitting the massive amount of data to the data centers. For optimal performance, it is expected that the IoT technology design specifications must exhibit a reasonable link budget, low power consumption, and finally, cost-effectiveness. To ensure a reasonable link budget, available communication technologies will be reviewed.

5.1 Near Field Communication

Near field communication (NFC) is a two-way low-bandwidth wireless technology that supports contactless data transfer in the neighborhood of 400 Kbps. Wireless interaction of machine-to-machine and information exchange takes place in a range of 0.1 m. It is a low-powered technology that supports transactions between electronic gadgets in particular through mobile payment systems (using Android pay, Apple pay, or Samsung pay) such as smart phones, tablets, payment transaction through contactless point of sale, etc. Unfortunately, the data rate capability of technology of less than 500 Kbps makes the technology inadequate for a big data system because of the low-bandwidth limitation. The NFC technology relies on near field EM induction, the region of which experiences a dominant magnetic field. Because the magnetic field density is high in this region, the performance is limited

by a short link budget of a few centimeters. Moreover, the EM compatibility (EMC) effect is significant. To be adaptable for big data, hundreds of billions of interconnections may be required in order to increase the channel capacity, and also accommodate such massive data. In effect, the cumulative effect of the EMC may be dysfunctional to optimal performance of NEC technology with respect to big data hardware resources. The drawback of such a scenario is expected to become obvious as its deployment enters the mainstream in the next few years. To address these shortcomings, attempts to forestall such a scenario are being made. For instance, Kim et al. foresaw this shortcoming and hence proposed near field MIMO based on heterogeneous multipole antenna arrays [27]. Their resulting design successfully enhanced channel capacity and reduced crosstalk between the transmitter and receiver.

5.2 Radio Frequency Identification

Radio frequency identification (RFID) is wireless technology consisting of a miniaturized IC chip and an antenna. While this electronic device is not different from the magnetic strip on an ATM card, credit/debit card, or a bar code in terms of application, purpose and object identification, the alignment problem between the reader and the tag, reminiscent of what is obtainable in these above-mentioned applications, is uncommon with RFID. It exhibits a good budget link range of 0.0001–0.2 km [6]. However, it has a low data rate capability, besides the usual tag and/or reader collision challenges, occurring because more readers overlap, as a result of simultaneous queries.

5.3 IEEE 802.15.4

Wireless home automation networks standardized by IEEE 802.15 wireless task group 4 are low-power, low data through-put technology predicated on battery-powered RF transceivers. The applications include ZigBee, z-wave, 6LoWPAN, Thread, Wavenis, INSTEON, etc. On average, they demonstrate low data throughput in the range of 20–250 Kbps, and a link budget of about 10–50 m for indoor and 45–1000 m for outdoor operations. They operate at an average frequency band of 433 MHz–2.4 GHz. Though these applications are similar, nonetheless there have been considerable differences in terms of physical/link layers, network/application layers, communication modes, implementation size, modulation technique, expected latency, etc. Notwithstanding this, they

demonstrate low latency on average, except the emerging INSTEON and its derivatives with an average latency of between 100 and 200 ms, and will consequently pose a significant challenge when deployed for big data.

5.4 Cellular Technology

IoTs requiring a good link budget essentially rely on cellular technology such as GSM, 3G, or 4G (Long-Time Evolution LTE, Long-Time Evolution—Advance LTE-A). The link budget ranges from 40 km to close to 200 km for high-speed packet access. The technology demonstrates wider coverage, high-speed data throughput, low latency, and enhanced bandwidth, in particular for LTEs with a data throughput of about 3–10 Mbps. However, it operates at lower frequency bands of the spectrum, such as 900/1800–1900/2100. These frequency bands cause the electrical length of the antennas to be large in terms of their aperture size, thus making highly integrated circuit nearly impossible. Deployability as IoTs for big data may pose a significant challenge in terms of hardware technology of their ecosystem. Furthermore, the power consumption is enormous, and the cost may be exorbitant.

5.5 WiFi and RF Technology

Low cost, low power consumption, higher data throughput and long link budget connectivity are competing design requirements to implement successful highly densified IoTs for big data environments. In order to transmit massive data to the large-scale servers (gateway), alternative robust connectivity such as that offered by Bluetooth, WiFi (2.4/5.8 GHz), and the emerging 60 GHz technology, otherwise called sub-wavelength transceivers, provides better data throughput, higher speed and low cost. The 2.4/5.8 GHz support data throughput of 0.5–1 Gbps, a link budget of less than 100 m and reduced power consumption. While the antenna form factor of 5 GHz is reasonable, the 2.5 GHz aperture size is substantial and requires higher power consumption. 5.8 GHz transceivers support higher channel bandwidth with an attendant low link budget, selectivity, and sensitivity limitations. They also demonstrate inferior penetration. The congestion of the 2.4 GHz spectrum due to many competing applications causes considerable EMI infractions from neighboring applications.

To satisfy the big data requirement of ultra-high speed and highly integrated connectivity SoCs using IoTs, emerging mm-wave radios will be appropriate. Channel bandwidth in excess of 12 GHz, substantial data throughput of about 1 Gbps, good narrow-band transmission capability and reasonable power consumption are the many advantages of sub-GHz radios. Because of their extremely small form factor and advances in Si technologies for scalability, highly densified integrated transceivers can be realized.

6 Integrated Antenna and System-on-Chip Transceiver Chipset

A CMOS exhibits optimal integration alternatives and is cheapest. Device scaling via bandgap shrinking and thinning of Ge with respect to Si permits optimal performance enhancement in terms of low cost, single-chip integration capability and lasting battery life. Research continues on semiconductor materials on Si [40], with a focused interest in the development of transistor-scale heterogeneous integration processes to combine high-power CS devices intimately with high-density Si CMOS circuits, in order to proffer solutions or alternatives to numerous challenges facing today's engineers and research scientists attempting to offer low-cost, efficient and tunable designs. CMOS technology applications (including the RF front-end) are ubiquitous and gradually becoming standard practice at 10 GHz and below. SiGe is an advanced technology over the traditional Si mainstay semiconductor technology. Being a state-of-the-art technology, it offers improved and optimized power consumption capability, a reduced number of external components, better sensitivity, high speed, better gain and an enhanced dynamic range. The design of the antenna, in particular its integration, becomes much easier at 60 GHz, as its size is in the neighborhood of 5 mm. Thus, the integration of such an SoC into a package becomes less cumbersome. Unfortunately, the antenna performance metrics in terms of beamwidth, impedance bandwidth, antenna efficiency and directive characteristics are degraded as a result of the reduced form factor. Substrate low resistivity, high permittivity, surface waves and thickness are factors that depreciate their performance profile. The radiation efficiency degrades with respect to the choice of substrate. AiPs offer a promising and more efficient alternative.

7 Antenna-in-Package

The AiP alternative is a contemporary unconventional and vital success in wireless systems' miniaturization. The AiP solution, unlike AoP, is identified as the utmost auspicious antenna alternative for extremely integrated mm-wave applications, albeit at short-range but very high-speed wireless communications. This is because of the available broad bandwidth [3, 13, 20, 28, 30, 38, 39, 44, 52, 60, 65]. It alleviates the difficulties in interconnections between the chip and antennas, as well as the motherboard [42, 62]. The interconnection can be accomplished by using either the flip-chip or wire-bonding technique. Both interconnection techniques are implemented based on inductive interconnections and exhibit optimal frequency bounds, notwithstanding interconnection via capacitive- or EM-coupling techniques. The two coupling techniques are deployable where and when necessary. These are the very reasons why the AiPs have outperformed the AoCs, and are considered superior, in particular in terms of low cost, compact size, high gain and radiation efficiency [35].

8 Antenna-in-Package Technology

As advances in SiGe and CMOS technologies progress, single-chip alternative solutions in Si become progressively realizable, thus making integrated mm-wave technology more attractive. The short wavelength specificity of the mm-wave band allow antennas to be integrated on a chip or embedded within a package. Thus, a large number of antennas can be combined in a densely integrated radio die.

The implementation can be done in either horizontal geometry or vertical type, as illustrated in Fig. 2 and further illustrated in the work of Zhang and Duixian [59]. As a result, AiP has been identified as the most promising alternative antenna solution, in particular with respect to its inherent enormous impedance bandwidth, extremely high speed, reasonable gain and substantial data throughput. For instance, considerable bandwidth efficiency is achievable by prototyping a planar antenna radiating element on a superstrate [4, 10, 18] or suspending an antenna element in air [14]. In a bid to achieve further low-cost integrated mm-wave AiP solutions, a proof-of-concept antenna prototyped on fused silica was developed, and is suitable for mass production as reported by Pfeiffer et al. [38]. A similar design, prototyped on different packaging, was realized in the work of Huang and Wentzloff [21]. The realized cavity-backed folded dipoles demonstrated a robust radiation efficiency well above 90%, a gain of 7 dBi and an impedance bandwidth greater than 20 GHz.

In Fig. 3a a conceptual build-up of AiP based on a land grid array by IBM is presented [38, 38]. The antenna is flip-chipped using wire-bonding interconnect. The Toshiba research group also implemented an AiP solution using wire bonding [65], as demonstrated in Fig. 3b. The bond wires interconnect the chip input-output (I/O) pads and the substrate in such a way as to form a 3D triangular loop shape [63]. A similar procedure as shown in Fig. 3c was employed by NTU Singapore, also using bond wire [63]. Interconnections to interface the chip to the package is done by using the bond wires or flip-chip. It is evident that transmission losses due to discontinuity, bends and junctions are unavoidable as the excitation signal propagates along the interconnections. The effect of this becomes significant in the mm-wave spectrum, which is the subject of consideration in this chapter, in order to provide a robust platform for big data, and thus has a considerable influence on performance in terms of speed and data throughput. Though flip-chip technology yields a better performance profile compared to bond wire, bond wire is robust, tolerant to chip thermal expansion and cost-effective. Because of the severity of this problem, several alternative solutions to ameliorate the challenges have been proposed and novel solutions are being investigated. The common solution is to reduce either the interconnect length of the bond wire or the chip-package inter-spacing. However, reduction of the bond wire length is limited by manufacturability procedures, whereas wider chip-package spacing must be maintained to enhance the performance, especially in the mm-wave band.

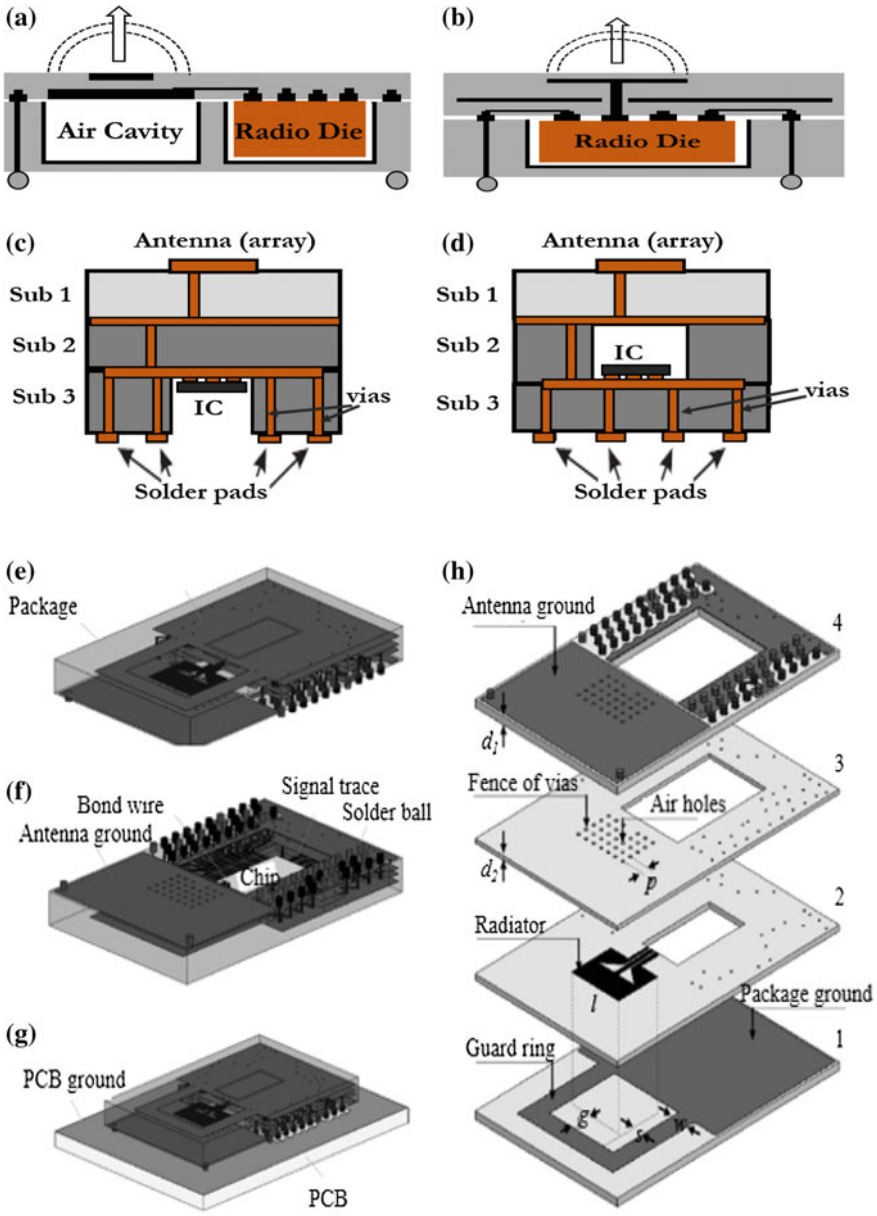


Fig. 2 An AiP solution **a** horizontal, **b** vertical [62], **c** flip-chip mounted IC with open cavity, **d** flip-chip mounted IC with closed cavity, **e** encapsulated AiP top view [23], **f** encapsulated AiP bottom view, **g** encapsulated AiP landing view [63], **h** encapsulated AiP exploded view [61, 63]

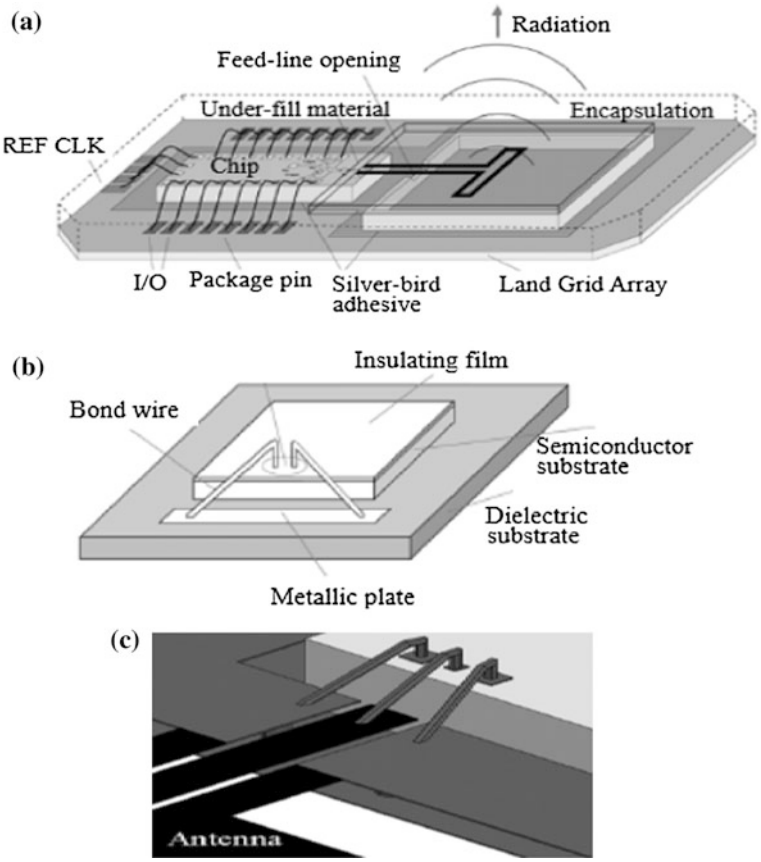


Fig. 3 AiP configurations **a** IBM [38, 39], **b** Toshiba, **c** NTU [63]

Zhang et al. [63] introduced Budka's bond wire compensation scheme (shown in Fig. 4) to mitigate the effect of transmission loss as the frequency increases at mm-wave frequency, whereas a 3D triangular loop was proposed by Toshiba engineers, as shown in Fig. 4b. Toshiba engineers indirectly elongated the interconnection between the chip and the substantial electric current by following a triangular loop in order to accommodate sufficient distance between the chip and the package such that the strong electric current along the path decays while simultaneously maintaining compactness. Figure 4a depicts the compensation technique introduced by Zhang et al. [63] using Budka's bond wire compensation scheme. Though these alternative solutions were able to reduce transmission losses, the EMI problem is largely unsolved and solutions for transmission losses proposed by these research projects have not been standardized by any professional body.

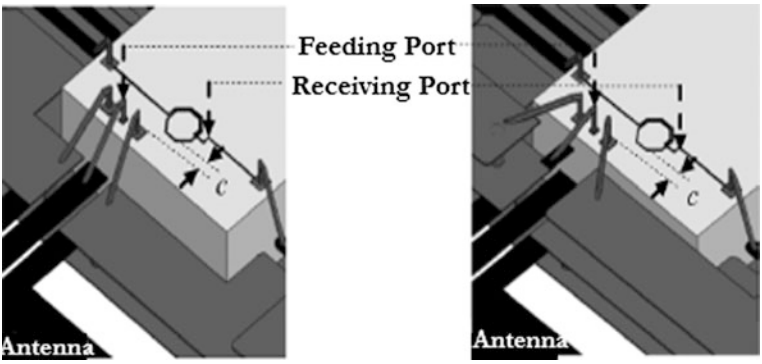


Fig. 4 Budka's bond wire compensation scheme [63]

9 Electromagnetic Interference Implications of Interconnections

When *RC* interconnection lines are etched on a substrate, as shown in Fig. 5b, the planar interconnection structure is similar to a radiator, such as a passive transmission line or an antenna. As these transmission line structures bend, discontinue, terminate abruptly or form junctions, fringing fields are created, in particular when spacing between the interconnection structures and the subsequent conductive layers is interleaved with a dielectric material. The radiation characteristics of these structures increase their interference susceptibility to EM fields emanating from the structure upon excitation. To understand this phenomenon better, a few methods of analysis are often used to investigate the EM interaction challenges on the circuits. Analysis using transmission line coupling theory based on injected EM has been popular in the past. Because of the inefficiency of this analysis, alternative solutions based on a full-wave technique predicated on 3D finite-domain time-domain (FDTD) codes are now commonly used. Though the FDTD methods are accurate, they require considerable computational resources in particular when applied to high-speed, high-gain, and, large-scale big data hyper-performance server platforms, as are discussed in this chapter.

To proffer alternative efficient solutions, diverse methods have been proposed, as reported in many publications. Rather, we propose a delay response extraction model based on the equivalent circuit representation of the interconnected structures on multi-layer substrates. A matrix-based extraction algorithm is employed using externally applied field coupling excitation as a result of current sources (i_i) and applied voltages (v_i) across the interconnections. While this method investigates the time delay response on the long *RC* interconnection lines in the vicinity of EM interactions, we will focus on the implications of *RC* interconnection delays (for the

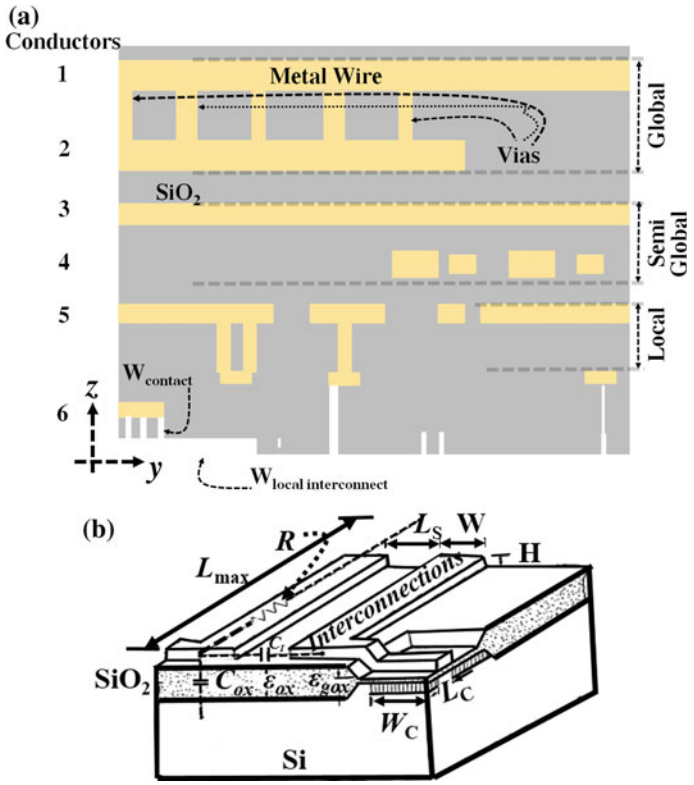


Fig. 5 A typical cross-section schematic diagram showing hierarchical contacts, interconnects and vias. **a** *yz*-plane, **b** *xyz*-plane interconnect [15, 43]

purpose of brevity) as demonstrated in Sect. 10, and rather interpolate the effect of EM interactions on circuit delay using the work of Shen et al. [46] and Tang et al. [49]. We go further to validate these interpolations using 3D EM numerical code based on advanced numerical methods. Our interpolation indicates that the availability of passives in the vicinity of the *RC* interconnect lines depreciate marginally the gain of the would-be antenna oriented in parallel to these interconnect lines. This gain may seem marginal, but the effect is catastrophic, since the gain of the traditional antenna on/in-package on silicon substrate is intrinsically in the negative territory owing to high substrate permittivity and low resistivity of the dielectrics. In addition, the *RC* interconnect lines are capable of shifting upward the resonance response of integrated antennas due to EM interactions and fringing effects. The metallization thickness of the interconnect passives on the dielectrics substantially affects the transmission gain, such that its increase substantially reduces the transmission gain. Moreover, the EM radiations from the interconnects affect the

CMOS functionalities due to EM interactions. The side effect of the EM interactions is also evident between I/O adjacent interconnect coupling lines with respect to signal integrity, such that the signal propagation becomes weaker. This side effect causes crosstalk between adjacent channels. The effect gets worse in the mm-wave frequency spectrum and for a high-speed data network capable of supporting big data server platforms.

10 Analysis of Interconnections

Interconnects are conductive wires or flip-chips that interconnect functional circuit elements and devices. It is evident from the above sections that either low-cost bond wires or flip chips are the standard procedures employed to connect the lead frame to the on-chip electronics. These interconnects consist of resistance, capacitance and inductance. As a result, these identified intrinsic components introduce a capacitive, resistive, and inductive parasitics contribution. By implication, side effects such as power distribution challenges, energy consumption, delay, transmission losses due to bends, junctions, discontinuities and noise are common occurrences with attendant system unreliability. As feature size reaches $<0.35\lambda$ microns, the parasitics effect becomes substantial, with associated poor signal propagation integrity along the interconnects. The computational run time becomes significant, with undesirable trade-off in terms of accuracy. Figure 5a depicts a typical cross-section of a backend structure, showing different interconnections, contacts and vias in hierarchical topography. Global interconnects known to have low resistivities usually travel above the local interconnects plane and much longer distances across different circuit elements and devices. However, the local interconnects do not travel over long distances because of their higher resistivities. Figure 5b demonstrates *RC* analysis of interconnect structures on the SiO_2 layer. It is anticipated that if global interconnects should travel a longer distance, the discontinuity effect, transmission loss, delay, and signal distortion will be considerable. To confirm these assertions, we present the parasitic characterization effect of interconnections based on equivalent circuit representation of interconnections depicted in Fig. 6. Consider an IC with m number of conductors and dielectrics. Assume that both the conductor surfaces and the dielectric interface are discretized into n cells (or tiles), where $n = n_c + n_d$ such that n_c is the number of cells on the conductors, and n_d is the number of cells in the dielectrics. Let the length of each interconnection be L_I. It is expected that that a charge q_i will be uniformly distributed on each cell (i) when potential is applied across the two conductors x_i interspaced with a dielectric interface.

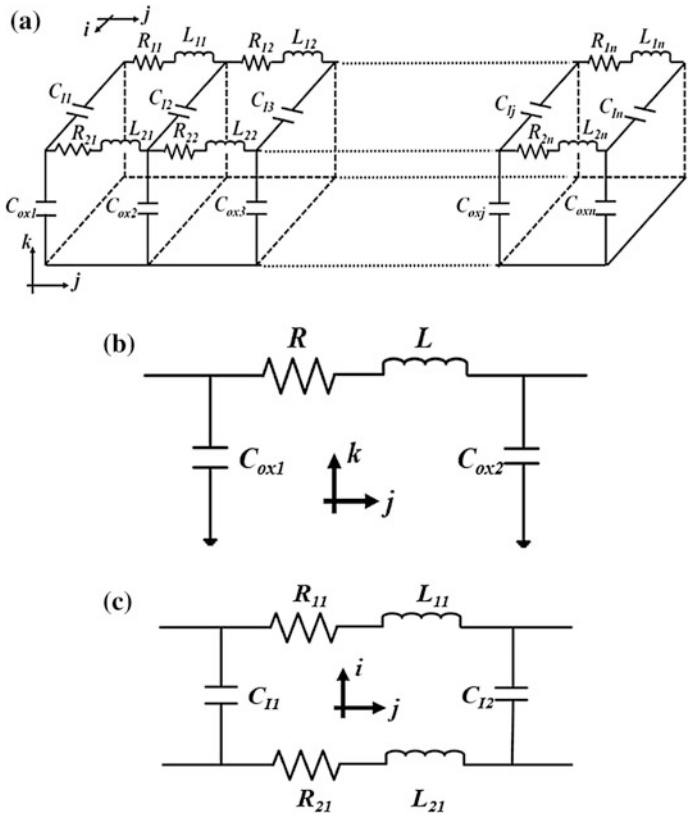


Fig. 6 Equivalent circuit of a typical cross-section interconnect. **a** *xyz*-plane, **b** *yz*-plane, **c** x*y*-plane

Applying Gauss's law, the charge distribution on n cells can be determined. The charge contribution on a cell j to the potential at the center of cell i is as stated in Eq. (1).

$$\overline{v^2}(q_i P_i) = 0$$

$$q_i\left(\overline{v^2}.P_i\right) = q_i\left(\frac{\partial^2 P_i}{\partial x^2} + \frac{\partial^2 P_j}{\partial y^2} + \frac{\partial^2 P_k}{\partial z^2}\right) = 0$$

$$\frac{\partial^2 P_i}{\partial x^2} = \frac{P_{i+1,j,k} - 2P_{i,j,k} + P_{i-1,j,k}}{\Delta x}$$

where

$$
\begin{aligned}
&i_i(x_i,t) - i_i(x_i+\Delta x,t) = C_i \Delta x \frac{\partial v_i(x_i+\Delta x,t)}{\partial t} \\
&v_i(x_i,t) - v_i(x_i+\Delta x,t) = R_i \Delta x i_i(x_i,t) \\
&\frac{\partial^2}{\partial x^2} v_i(x_i,t) = -R_i \frac{\partial}{\partial x} i_i(x_i,t) \\
&\frac{\partial v_i(x_i+\Delta x,t)}{\partial x} = -R_i i_i(x,t), \quad \frac{\partial^2}{\partial x^2} v_i(x_i,t) = -R_i \frac{\partial}{\partial x} i_i(x_i,t) \\
&\frac{\partial}{\partial x}\left(\frac{\partial v_i(x_i)}{\partial x}\right) = -R_i \frac{\partial}{\partial x} i_i(x,t), \quad \frac{\partial^2}{\partial x^2} v_i(x_i,t) = -R_i C_i \frac{\partial}{\partial x} v_i(x_i,t)
\end{aligned} \tag{1}
$$

$$
P_{ij} = \frac{q_i}{a_j} \int\limits_{a_j} \frac{1}{4\pi\varepsilon_0 \| x_i - x' \|} da'
$$

$$
P_i(x_i) = P_{i1} q_1 + P_{i2} q_2 + \cdots + P_{in} q_n \tag{2}
$$

$$
P_{ij} \approx \frac{1}{a_j} \int\limits_{cell\ j} \frac{1}{4\pi \| x_i - x' \|} da'. \tag{3}
$$

Equation (2) is the potential (P_i) at the center of the i-th cell with respect to sum of the contribution on all other n cells on the surface of the conductor. Similarly, the displacement field difference normal to the center of the i-th dielectric interface cell is as stated in Eq. (4), in particular as cell i lies at the dielectric-dielectric interface with dielectric constants ε_1 and ε_2.

$$
\varepsilon_1 \frac{\partial v_i(x_i)}{\partial n_i} - \varepsilon_2 \frac{\partial v_i(x_i)}{\partial n_i} = 0 \tag{4}
$$

Substitute Eq. (4) into Eq. (2) while evaluating the potential gives Eq. (5),

$$
\varepsilon_1 E_{i1} + \varepsilon_2 E_{i2} + \cdots + \varepsilon_j E_{ij} + \cdots + \varepsilon_n E_{in} = 0 \tag{5}
$$

where

$$
E_{ij} \approx (\varepsilon_1 - \varepsilon_2) \frac{\partial}{\partial n_i} \frac{1}{a_j} \int\limits_{cell\ j} \frac{1}{4\pi\varepsilon_0 \| x_i - x' \|} da'. \tag{6}
$$

Simplifying Eq. (6) gives Eq. (7). Equations (2) and (5) are re-expressed in a matrix form and as shown in Eq. (8).

$$D_{ii} \approx \frac{(\varepsilon_1 + \varepsilon_2)}{2a_i\varepsilon_0}, E_{ii} \approx \frac{(\varepsilon_1 + \varepsilon_2)}{2a_i\varepsilon_0(\varepsilon_1 - \varepsilon_2)} \tag{7}$$

$$\begin{bmatrix} v_1 \\ \vdots \\ v_{nC} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ P_{nC1} & P_{nC2} & \cdots & P_{nCn} \\ E_{n_C+1,1} & E_{n_C+1,2} & \cdots & E_{n_C+1,n} \\ \vdots & \vdots & \cdots & \vdots \\ E_{n_1} & E_{n_2} & \cdots & E_{n_n} \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_{nC} \\ q_{nC+1} \\ \vdots \\ q_n \end{bmatrix} \tag{8}$$

Further simplifying Eq. (8) gives Eq. (9).

$$\begin{bmatrix} v \\ 0 \end{bmatrix} = [q] \begin{bmatrix} P \\ E \end{bmatrix} \Rightarrow A\vec{q} = \vec{b} \tag{9}$$

where $i \neq j$, and A is a dense matrix equivalent to the number of cells × the number of cells. Thus, the capacitances of line-to-other-line, or the coupling capacitances (C_I) are given in Eq. (10)

$$\begin{gathered} A \approx \begin{bmatrix} P \\ E \end{bmatrix}, b \approx \begin{bmatrix} v \\ 0 \end{bmatrix} \\ [C_I] = [q] \begin{bmatrix} v \\ 0 \end{bmatrix}^{-1}, [C_I] = \sum_j^n C_{ij} \end{gathered} \tag{10}$$

$$\begin{gathered} [C_I] = \varepsilon_{ox}\varepsilon_0 \frac{\sum_j^n H_j L_j}{\sum_j^n L_{Sj}} \\ C_T = k_I \left([C_I] + \varepsilon_{ox}\varepsilon_0 \frac{\sum_j^n W_j L_j}{X_{ox}} \right) \end{gathered} \tag{11}$$

$$R_{ij} = \rho \frac{\sum_j^n L_j}{\sum_j^n W_j H_j} \tag{12}$$

The total capacitances (C_T) with respect to Fig. 5b are as stated in Eq. (11), whereas the equation to determine the line resistances of the interconnects is stated in Eq. (12), where X_{ox} is the oxide thickness, k_I is the factor responsible for the

substrate fringing field, ρ is the interconnect resistivity, and ε_{ox} is the dielectric permittivity of the oxide. The second term of Eq. (11) represents the substrate-to-line capacitance. To determine the delay time (τ_L) using $\tau_L = 0.89\ RC$ gives Eq. (13).

$$\tau_L = R_{ij} k_I \left([C_I] + \varepsilon_{ox} \varepsilon_0 \frac{\sum_j^n W_j L_j}{X_{ox}} \right) \tag{13}$$

The sizes of W and L_S are contingent on the feature size and lithographic capabilities. It is not uncommon to see a few laboratories running lithographic technologies with etching capabilities around 45–180 nm. In this work, we assume $X_{ox} = 0.18\lambda$ micron and $H = 0.045\lambda$ micron.

11 Performance Metrics of Interconnections

Figure 7 depicts the performance metrics of *RC* line interconnections. In Fig. 7a, the delay time against the line resistance is plotted. It is evident that the delay time increases as resistance increases, in particular as the line length of the wire interconnection increases. The delay time also increases as the line width decreases. A steep time delay response is observable in the medium to long range as the line length increases. Figure 7b demonstrates the effect of total capacitances (C_T) on the delay time. The delay time responds rather sharply in a short time to about 200 fF/mm, and subsequently increases steadily to the medium and long range. The influence of the line-to-line capacitance (C_I) on the response time is dominant. Interestingly, the capacitance is in a way influenced by the degree or extent of the dielectric permittivity. Low dielectric permittivity lowers the delay as a result of the *RC* wire interconnection. It also lowers the power consumption and crosstalk. Once these interconnection lines are energized, the propagated energy from the source down the line dissipates owing to line loss as the power propagates along the *RC* line, with attendant reduced data throughput. Tapering the cross-sectional area of the *RC* interconnection lines (but not below the line effective resistance) improves the response time, in particular in the medium to long distance. The delay time becomes considerable as the interconnection line length increases, as demonstrated in Fig. 7c. The delay times also appreciate as the resistivity of the lines increases with increasing length. The delay time is almost linear until the line length is about 5 mm, after which the response becomes progressively steeper. The response gradient is only contingent on the feature size as depicted. The gradient is highest in the case of a feature size of 45 nm, and lowest at 180 nm. Therefore, the delay time will remain a challenge as the lithographic technologies progress. Figure 7d is a graph of delay time versus wire width (in mm). The delay response appreciates as

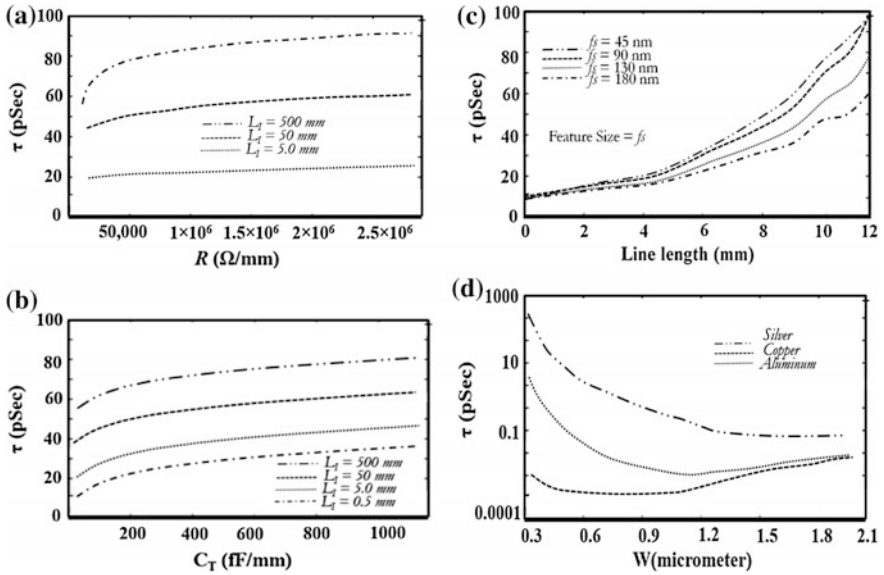


Fig. 7 Performance assessment of interconnections. **a** Delay versus C_T, **b** Delay versus R, **c** Delay versus *line length*, **d** Delay versus W

the interconnection wires increases. This effect stabilizes at $W = 1$ μm, and subsequently mildly linearizes at 0.0001 $<\tau>$ 5 psec. The *RC* delay curve characteristics stipulate its response dependency on the resistive and capacitive contributions. Though their effects on the delay time vary from short, medium and long distances of the interconnections line length and width, the delay specifically impedes the circuit speed performance and data throughput capability. The longer the line length, the more the delay, and also the power consumption.

12 Performance Improvement of Interconnections

It is evident therefore that the delay response on *RC* interconnection lines is substantial. It is also obvious that the crosstalk noise is unavoidable because of the presence of EM interactions on the lines. Distortion tendencies of this effect on the communicated data are supported by Tang [50] who observes byte-swap of the most significant byte of a counter owing to EM interactions. He further points out that the EMI problems affect both high-speed and low-speed systems. Dynamic power consumption consequently occurs and increases with *RC* interconnection

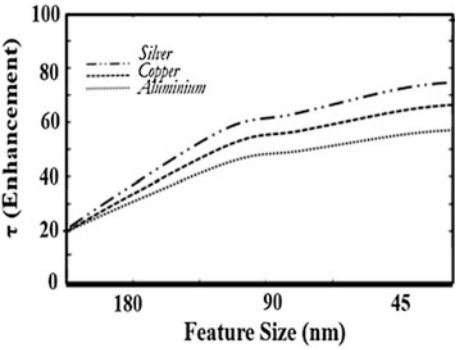


Fig. 8 Performance improvement of interconnections delay response

length. Figure 8 demonstrates the performance enhancement possibility of these interconnections with respect to the feature size progressions. It is clear that silver-etched *RC* interconnects demonstrate overt enhancement compared to copper, followed by aluminum. While silver could be expensive, aluminum exhibits substantial delay in the neighborhood of 40% of the signal propagation delay. While copper demonstrates brilliant enhancement over aluminum, the lithographic etching process using copper is very challenging with respect to the low dielectric constant and low-*k* materials of wire isolation projections of the National Technology Roadmap for Semiconductors.

The large power consumption by the *RC* interconnect lines, along with a considerable amount of delay, makes them inefficient for use in high-speed networks. Hitherto, traditional medium-to-long-distance interconnects have often been subdivided into several repeaters to improve the delay response. Instead, transmission lines with copper interconnect traces demonstrated enhanced delay response and faster signal propagation, with moderate power consumption even when the transmission line interconnects were longer, and where repeater technique was employed. However, the relatively wider surface area of the traces in transmission line interconnects (compared to the *RC* interconnect structures) leads to extremely large C_I as the number of transmission line interconnects multiply as exemplified in highly heterogeneous large-scale and high-speed servers. Though transmission line interconnects enhance power consumption of long interconnects and also improve the delay response, the increase of C_I on the flip side circumvents the advantages otherwise achieved. Therefore a compromise becomes mandatory in the light of the signal propagation delay challenge to either target to improve delay response or save power consumption. The medium-to-long interconnects could be replaced with transmission lines to improve power consumption, whereas the delay response time could be improved by replacing the longer interconnect lines on critical paths with transmission line interconnects.

13 Coupling Techniques of Interconnections

Inefficient power coupling via the interconnections to the ICs degrades the ICs' performance to a large extent owing to substantial signal propagation losses. Poor matching; losses due to bends, junctions, discontinuity, and abrupt changes along the interconnection lines; high substrate permittivity and coupling of power-to-substrate modes owing to surface waves are a few of the many factors responsible for signal degradation as a result of poor coupling. The signal distortion becomes so significant that accurate data transmission cannot be ascertained. Incidentally, the efficiency of the coupling techniques is a measure of the chip-to-substrate gap relative to the height of the chip. Various efforts to ameliorate these problems have been reported in the literature, with many of them bedeviled by numerous challenges. In this section, we identify the types and determine their advantages.

13.1 Traditional Wire-Bonding Interconnection

Thermal and electrical properties coupling to ICs via bond wires have been inefficient because of poor impedance mismatch and insertion losses with respect to signal propagation along the interconnect structures. The mismatch creates a bonding circumference about the soldered joint that restricts the channel bandwidth aggregate data throughput. Besides, the effect also limits the speed performance for data transfer because of insertion losses. Importantly, the bond wire invariably induces (1) considerable signal time delay, which increases with an increase in I/O count, (2) a significant inductance contribution, which increases with an increase in signal frequency, and (3) very high power dissipation.

13.2 μB Technology Interconnections

μB coupling technology was subsequently introduced as an efficient alternative solutions to wire bonding. However, the interconnection coupling based on this technology using either gold bumps or copper solder makes the resulting circuitry very bulky and unfit for highly integrated antenna on/in-chip systems. In essence, these seeming alternatives exhibit deficiencies with respect to their higher vertical stacked density of interconnection when compared to bond wire interconnections.

13.3 Through-Si Vias Interconnections

The through-Si vias (TSVs) support 3D IC integration to facilitate high bandwidth, system scaling, low power utilization and good systemic performance. The

resulting chips with TSV coupling interconnects are miniaturized owing to technology that relies on the vertical stacking of the ICs and uses TSVs to interconnect the chips along the shortest paths. The interconnection lengths are thus reduced with attendant enhanced delay time and reduced power utilization. The technology is notorious for high frequency loss that eventually degrades the system performance through EM interaction effects, as several interconnection traces are crowded into limited 3D space. Proximity and skin effects become prominent as frequencies increase, thus creating alternate current paths.

13.4 Capacitive Coupling Interconnections

Technology pull to satisfy low noise requirement, high bandwidth, low power and low delay design considerations now poses a need for alternative coupling interconnect technology for highly densified and high I/O bandwidth chips. The alternative coupling to support highly integrated large-scale multi-gigabits ICs with miniaturized chip area, increased packaging density, high pin counts, and extremely low power utilization can certainly not be met by traditional bond wire-based electrical interconnects. As the scaling of CMOS technology continues, this technology pull can only increase in monumental proportions. Contactless (proximity) power coupling proves to be robust especially when compared with the traditional low-speed bond wire interconnects. Capacitive (AC) coupled interconnects based on contactless technology circumvent the mechanical limitations of other contact interconnect technologies. However, testing becomes onerous as the pitch decreases in order to satisfy the standoff height requirement in the face of escalated growth of I/O signal connections. The reliability of the trench geometry is another cause for concern, in particular since the coupling element density and its performance are dependent on the height of chips' substrate gap. The side effect is the degradation of the coupling element performance with respect to the square of the gap height.

13.5 Inductive Coupling Interconnections

Inductive coupling interconnect technology also employs contactless (proximity) coupling technology in the similitude of the capacitive coupling interconnects. It permits medium-to-long distance power coupling between chips. The extent of the distance of coupling is dependent on the inductor layout area as well as the driven current. While the capacitive coupling interconnects use the capacitance occurring between the dielectric interleaved two conducting plates representing the chip metal traces and the substrate ground, the inductive coupling interconnects rather rely on the separate interfaces through vertically stacked structures. Thus, it is current-driven as against voltage-driven capacitive coupling interconnect technology.

14 On-Chip Radio Frequency Signal Propagation Enhancement Techniques

14.1 Effects of Grounding

When a substrate is grounded, it is connected to the ground plane (which could be solid or mesh) through vias. The Ohmic connection is thus established between the ground plane and the substrate, unlike that obtained in the floating substrate. Usually, the rate of decrease of dielectric space between the interconnect passives and the ground conductor creates more proximity effects. These effects cause the concentration of part of the return currents directly below the interconnect passives. The effect of this becomes substantial at the mm-wave frequency spectrum. The interconnect resistance (R) reduces considerably as the interconnect passive ground plane space increases. Thus, the propagation signal delay and the IC power utilization improved. An increase of substrate conductivity further increases this performance profile. The use of tiled Si has proven to offer comparative advantage over single-chip Si in terms of interconnect delay enhancement [16].

14.2 Effects of Guard Rings

I/O interconnect structures experience high signal propagation delay, especially when they are large in number as applicable to densified highly integrated large-scale server platforms for big data communication. Often, the delay is due to the radiated EM interactions and near field coupling. Theoretically the guard ring, when carefully implemented, has the capability to shield the interconnect line passives from the radiated EM interactions and near field coupling. Figure 9 is a typical example of a guard ring. While Fig. 9a depicts the geometry of the guard ring in the *xy*-plane, Fig. 9b demonstrates the implementation on a substrate in the *xyz*-plane with its equivalent circuit. The mutual capacitive contribution is created because of the capacitive coupling field between these many interconnect passives. In effect, the delay time depreciates with an increase in C_T. In principle, the parasitic effects due to the capacitive contributions with respect to electric field lines between the interconnect passives terminate owing to low potential (as the guard rings creates a low impedance path to ground) of the grounded guard rings. In effect, the EM field interactions between these interconnect passives are thus confined such that the capacitive contributions cancel out. This reduces coupling and hence improves isolations and delay time. Table 1 review the reported implications of the guard ring technique on performance enhancement of interconnection delays, in particular when or not the guard ring is grounded. Findings indicate that isolation becomes substantial when the guard ring is grounded, and inferior otherwise, as reported by Xu and Wang [54]. Generally, the application of the guard ring technique adds value to the improvement profile of the delay time by

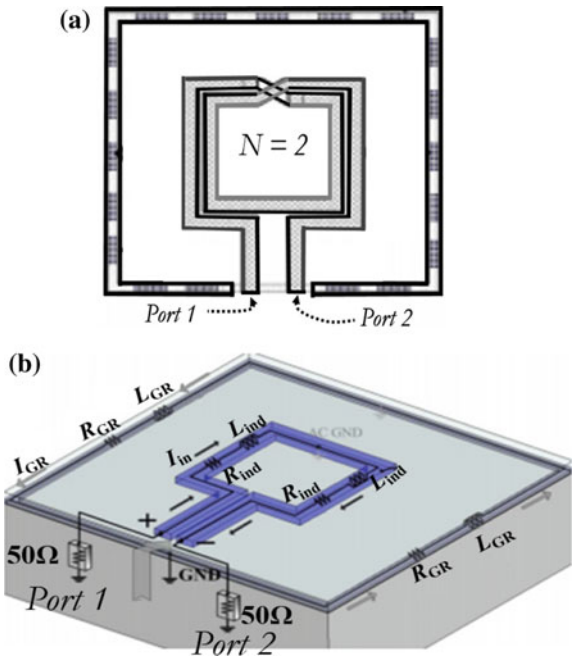


Fig. 9 Guard ring improvement technique of interconnections. **a** *xy*-plane geometry, **b** xyz-plane geometry and its circuit interpretation [57]

Table 1 Performance improvement using guard ring

Authors	Frequency	Feature size	Isolation dB
Zhang et al. [64]	400 MHz–40 GHz	130 nm	10–30
Noh et al. [65]	Not stated	200 μm	31.8
Xu and Wang [54]	10–20 MHz	0.23 mm	7–11 (No grounding)
			10–15 (With grounding)
Kim et al. [25]	10 MHz	10 μm	19
	10 GHz	50 nm	7
Tsai and Ker [51]	Not stated	0.6 μm @ 5 V	Not stated
Shen et al. [46]	0.04–10 GHz	0.18 μm	Varying guard ring width: 15–50 15–50 to victim distance 15–40 aggressor to victim distance
You and Huang [57]	55.7 GHz	90 nm	Not stated

Table 2 Performance improvement using EBG

Authors	Freq. (GHz)	Technology	Packaging	Isolation dB
Park et al. [37]	0–20	Double-stacked DS-EBG is embedded between the power and ground	LTCC + 10 μm Gold + DS-EBG	−100 @ 3.5–6 GHz
Shahparnia and Ramahi [45]	0–15	Non-symmetrical EBG structure + high dielectric material embed	Top layer: High dielectric material $\varepsilon_r = 30$	−70 @10 GHz with EBG
			Bottom layer: EBG implemented on FR4	−45 @10 GHz without EBG
Roger [41]	0–12	Parallel-plate waveguide (PPW) interleaved with 2 dielectric layers + EBG	Conventional PCB	−100 @ 3–7 GHz
Abhari and Eleftheriades [1]	0–12	FR4-based PPW + EBG surfaces	Conventional PCB	−90 @ 3.8–4.2 GHz with EBG
				−43 @ 8 GHz without EBG
Ma et al. [32]	9–16	PCB-based EBG with via	Conventional PCB	>−40 with EBG
				<−30 no EBG

increasing isolation and thus reducing capacitive coupling. The use of the guard ring-to-victim distance, and guard ring-to-aggressor-to-victim distance substantially improves the effect of the guard ring on isolation enhancement. Unfortunately, the effect of the guard ring on the isolation function improvement depreciates with an increase in frequency (Table 2).

14.3 Effects of Shielding

It has been documented in Sect. 11 how total capacitance (C_T) affects the delay response time, in particular in Fig. 7b. Applying the shielding technique reduces the effective capacitance of the *RC* interconnects. Theoretically, it is expected that the resistance of the interconnect structure will obstruct the propagation signal through the interconnect, thus partly shielding C_T, and eventually enhance the delay response. In principle, the capacitive coupling occurring along a segmented RC interconnect passive will be marginal, provided the spacing is small in distance and cross-sectional area. Hence, such a capacitive coupling function may not be substantial enough to change with respect to orthogonal interconnect utilization [33]. The concept is such that when the *RC* interconnect passives are signal lines segmented in interdigitated form, and perhaps grounded, such interconnect line passives are isolated from the neighboring passives. The consequence is that capacitive coupling is eliminated because the technique will inadvertently screen the interconnect line passives [33]. However, a glitch voltage appearing across the interdigitated spacing is a challenge. Importantly, increasing the frequency of operations substantially increase the series resistance along the passive, thus depreciating the efficiency of this technique.

14.4 Effects of Decoupling Capacitor

Figure 10 depicts the fundamentals of the decoupling capacitor technique, where R_S is the insulation resistance of the interconnection line passives, R is the equivalent series resistance of the interconnecting coupling parasitic, and C_T is the total capacitances occurring between the interconnect lines, as depicted in Fig. 10a. Theoretically, the frequency response shown in Fig. 10b increases as the impedance function of the interconnect coupling capacitance C_T decreases monotonic-wise as depicted by x. Essentially, resistance R flattens the impedance function as its value increases; the value increases as frequency increases with respect to an increase in L. The overall effect of this technique reduces the IR drop, and hence improves the delay response and power utilization.

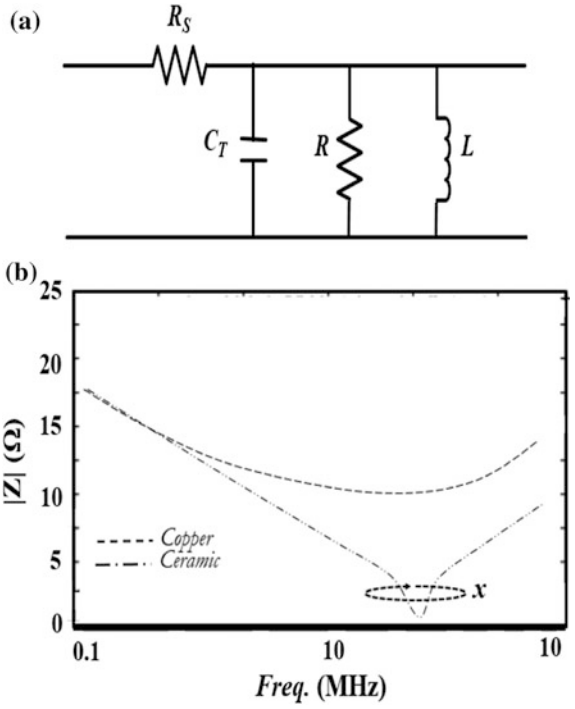


Fig. 10 Concept of decoupling capacitor improvement technique of interconnections. **a** Equivalent circuit, **b** Decoupling response

15 Optimization Techniques to Mitigate EMI at Process Level

A sizeable number of optimization techniques to minimize the EMI parasitics on circuit propagation delay of systems on/in-chips (SiCs/SoCs) have been reported in the literature. Removal of the EMI source(s), if it were possible, would have been the optimal solution. Owing to the impossibility of this method, the only feasible available alternatives, such as protecting (shielding) the affected interconnects lines, are only palliative measures, though research on solving these challenges continues at the modelling level, circuit level, board level, and even at the interconnect lines. This chapter addresses EMI and discontinuity effects on propagation delay that may disqualify SiCs/SoCs as promising solutions to integrate high-speed and large-scale servers to support massive data owing to slow speed, anticipated signal distortion and low return latency. However, stereotyped focus on the interconnect EMI optimization alone may not yield a holistic solution to this challenge. Therefore, we will focus more on the review of existing optimization techniques at interconnect line level (while we periodically proffer solutions at circuit level), empirically assess the efficacy of these alternatives, and finally make our judgment on the reliability and efficacy of such optimizations.

15.1 Fibre-Optic Alternative to Interconnect Lines

RC interconnect lines or transmission line passives (traces) behave like antennas (with increased tendencies to radiate as the interconnect line increases) upon excitation when the passives are supposed to transmit the current. This effect worsens when high-speed signals are propagated through such interconnect passives. To minimize such occurrences, the passive lengths and the ratio of passive width to substrate thickness (w/h) must be controlled. A value of w/h in the interval between 1 and 3 is expected to reduce the passives' radiation, and become unpredictable subsequently. Alternative techniques, such as effective EMI filtering, ground copper fills of unused trace areas (to prevent floating interconnect lines), short ground return interconnects, avoidance of 45° angled interconnect lines, avoidance of stubs on the lines, transient shields, trace separation, setting ground pour directly underneath the lines, grounding, using the guard ring, and the inclusion of an ancillary decoupling capacitor along the lines are other methods to reduce the EMI challenges. While these techniques have minimized the EMI parasitics on propagation delay, the techniques have added extra weight to the design, in addition to their cost-ineffectiveness. Worse still is the fact that these techniques are inefficient in supporting the propagation of high-speed signals. Using optic fibers instead of the passive lines will totally eliminate the EMI problems. However, the cost is daunting.

15.2 Shielding Optimization of Interconnect Lines

The *RC* interconnect lines or transmission line passives (traces) can be encapsulated by an optic fiber, other materials that exhibit magnetic properties, or conductive packaging materials. The materials mechanically shield the interconnects lines, thus preventing EM interaction with the lines. In essence, shielding techniques absorb both the magnetic and electric fields of the EMI. When such a shield is connected to ground, two effective advantages of interconnect line reductions (as fringing fields since the magnetic field densities could extend the electrical length, and hence the physical length of the interconnect lines are absorbed), and the reflection of the would-be EM radiation are achieved. However, shielding procedures add additional weight to the ICs, and hence extra cost.

A high-impedance surface (HIS) has been proved to offer an alternative bumpy periodic surface (texture) that is capable of reducing these EM interactions. In principle, the HIS forms an artificially engineered periodic bumpy lattice of EBG which in turn acts like a bandstop filter that forbids EM propagation in a particular bandgap under consideration [48], Yablonovitch (1999). Table 3 summarizes the HIS effect on the mitigation of the EMI with enhanced isolation well above −70 dB. Other alternatives use metal impedance suppression techniques, Faraday cage isolation structure, photonic bandgap microstrip waveguides,

Table 3 Performance improvement using HIS

Authors	Freq. (GHz)	Technology	Packaging	Isolation dB
Kamgaing and Ramahi [24]	0–4	Replacing parallel-power plane (PPP) pair with HIS	PBC	−70 @ 1.8 GHz with HIS
				−40 @ 1.0 GHz with PPP
				−20 with *RC* wall
Sievenpiper et al. [47]	0–30	Use of high impedance ground planed	Not stated	−60 with HIS
				−40 with conventional flat metal ground plates
Kamgaing and Ramahi [24]	0–4	Inductive/capacitive (instead of vias) enhanced HIS	Not stated	−80 PPP + HIS
				−80 PPP + HIS + wall of RLC
Clavijo et al. [11]	0.4–2.4	Artificial anisotropic magneto-dielectric material based on via array embedded in a dielectric with a capacitive FSS	Low dielectric foam	−60 @ 1–1.35 GHz
Islam and Alam [22]	2–5.5	Meander-line bridge high impedance electromagnetic structure	FR4 glass-epoxy	−70 Waveguide method
				−60 Transmission line method

silicon-on-insulator techniques, suppression based on stripline and microstrip structures, via hole fences, microstrip to stripline transitions, etc. These techniques are embedded in the substrate, and in particular in multilayer radio frequency circuits. In spite of their many enhancement advantages, a few of these possible alternatives exhibit various deficiencies either in terms of additional weight, design complexities, extra cost or overall system inefficiency, or are simply not realizable.

15.3 EBG Optimization of Interconnect Lines

The use of an electromagnetic bandgap (EBG) to create isolation by suppressing EM parasitic coupling is realizable at the process level of the board. Conducting parallel planes embedded with EBG materials could suppress EM coupling. The periodicity of EBG presents an intrinsic stopband phenomenon, which is sufficient to inhibit signal propagation in a frequency band. The EBG behaving as a bandstop filter prohibits EM couplings via the power-to-ground layers. Table 2 documents the existing performance enhancements reported by different authors using the EBG to mitigate the effect of EMI. Substantial isolation well above −40 dB was recorded in the table.

15.4 High-Impedance Surface Optimization of Interconnect Lines

A traditional flat conductive metallic ground plane has demonstrated poor radiation propagation characteristics owing to significant surface wave propagation along its plane. This surface wave propagates as EM signals, and radiates on reaching abrupt edges, bends and discontinuities. The radiations subsequently interfere with the circuits and in turn affect their performance as the EM radiations induce currents in the surface of the ground plane.

16 Conclusion

The chapter examines signal propagation delay through *RC* interconnection lines etched on compact multi-gigabits chips. For big data technology to be robust, several million interconnections of internet-enabled multi-gigabits chipsets will be required to accommodate and manage large amount of data, and to serve as storage platforms. If the signal propagation delay response along two capacitively coupled *RC* interconnect lines is considerable, delay response through several million multi-gigabit chipsets will be unimaginable. The consequence of such a quantum propagation delay will be evident in serious signal distortions, inability to recover the post-processed data precisely and perhaps severe power consumption owing to the length of the several million interconnect lines involved. In this chapter, a simple *RC* propagation delay model is developed by modifying the existing formulae available in the literature to examine the extent of the delay response for several million chipsets using *RC* π-network lumped element equivalency. We observed that interconnection transmission lines using thin film traces are more beneficial for propagation delay enhancement. We noted that traces made of silver, copper, or aluminum respectively reduce propagation delay in ascending order. In addition, we reviewed existing delay response enhancement techniques, identified their performance profile and noted their operational penalties.

References

1. Abhari, R., Eleftheriades, G.V.: Metallo-dielectric electromagnetic bandgap structures for suppression and isolation of the parallel-plate noise in high-speed circuits. IEEE Trans. Microw. Theory Tech. **51**(6), 1629–1639 (2003)
2. Adhikari, P.: Understanding millimeter wave wireless communication. In: VP of Business Development for Network Solutions, pp. 1–6. Leoa Corporation, San Diego (2008). http://www.loeacom.com/pdf
3. Ahmadi, M.R.N., Safieddin S. N., Zhu L.: On-chip antennas for 24, 60, and 77 GHz single package transceivers on low resistivity silicon substrate. In: Proceedings of IEEE Antenna Propagation Symposium. Honolulu, HI, 9–15 June 2007

4. Alexopoulos, N.G., Jackson, D.R.: Fundamental superstrate (cover) effects on printed circuit antennas. IEEE Trans. Antennas Propagat. **32**(8), 807–816 (1984)
5. Andrews, J.G., Buzzi, S., Choi, W., et al.: What Will 5G Be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
6. Atzori, L., Lera, A., Morabito, G.: IoT: A Survey. Comput. Netw. **54**(15), 2787–2805 (2010)
7. Bauer, H., Ranade, P., Tandon, S.: Big data and the opportunities it creates for semiconductor players. https://www.scribd.com/document/283642630/Big-Data-and-the-Opportunities-It-Creates-for-Semiconductor-Players
8. Char, K.: Internet of Things System Design with Integrated Wireless MCUs. Silicon Labs (2015). http://www.mouser.co.za/applications/internet-of-things-integrated-wireless-mcu/
9. Charthad, J., Weber, M.J., Chang, T.C., et al.: A mm-sized implantable medical device (IMD) with ultrasonic power transfer and a hybrid bi-directional data Link. IEEE J. Solid-State Circuits **50**(8), 1741–1753 (2015)
10. Choi, W., Pyo, C., Cho, Y.H., et al.: High gain and broadband microstrip patch antenna using a superstrate layer. In: IEEE Antennas and Propagation Society International Symposium, Columbus, OH, USA, 22–27 June 2003 (2003)
11. Clavijo, S., Diaz, R., McKinzie, W.: Design methodology for Sievenpiper high-impedance surfaces: an artificial magnetic conductor for positive gain electrically small antennas. IEEE Trans. Antennas Propag. **51**(10), 2678–2690 (2003)
12. Derhacobian, N.: One chip to rule them all? The Internet of Things and the next great era of hardware (2016). https://techcrunch.com/2016/05/28/one-chip-to-rule-them-all-the-internet-of-things-and-the-next-great-era-of-hardware/
13. Desclos, L.: V-band double slot antenna integration on LTCC substrate using thick film technology. Microw. Opt. Technol. Lett. **28**(5), 354–357 (2001)
14. Faiz, M.M., Wahid, P.F.: A high efficiency L-band microstrip antenna. In: IEEE International URSI Conference, Orlando, FL, USA, 11–16 July 1999
15. Gardner, D.S., Meindl, J.D., Saraswat, K.C.: Interconnection and electromigration scaling theory. IEEE Trans. Electron Devices **34**(3), 633–643 (1987)
16. George, A.G., Krusius, J.P., Granitz, R.F.: Packaging alternatives to large silicon chips: tiled silicon on MCM and PWB substrates. IEEE Trans. Compon. Packag. Manuf. Technol. Part B **19**(4), 699–708 (1996)
17. Guo, X., Li, R., Kenneth, K.O.: Design guidelines for reducing the impact of metal interference structures on the performance of on-chip antennas. In: Proceedings of IEEE AP-S International Symposium USNC/URSI National Radio Science Meeting, Columbus, OH, June 2003
18. Gürel, C.S., Yazgan, E.: Bandwidth widening in an annular ring microstrip antenna with superstrate. In: IEEE Antennas and Propagation Society International Symposium, Newport Beach, CA, USA, 18–23 June 1995
19. Harun-ur Rashid, A.B.M., Watanabe, S., Kikkawa, T., et al.: Interference suppression of wireless interconnection in Si integrated antenna. In: Proceedings of the IEEE International Interconnect Technology Conference, San Francisco, CA, 5–5 June 2002
20. Hoivik, N., Liu, D., Jahnes, C.V., et al.: High-efficiency 60 GHz antenna fabricated using low-cost silicon micromachining techniques. In: Proceedings of IEEE Antennas Propagation Symposium, Honolulu, HI, 10–15 June 2007
21. Huang, K.-K., Wentzloff, D.D.: 60 GHz on-chip patch antenna integrated in a 0.13-μm CMOS technology. In: Proceedings of IEEE International Conference on Ultra-Wideband, 1–2 Sept 2010
22. Islam, M.T., Alam, M.S.: Design of high impedance electromagnetic surfaces for mutual coupling reduction in patch antenna array. Materials **6**, 143–155 (2013)
23. Johannsen, U.: Technologies for integrated millimeter-wave antennas Eindhoven: Technische Universiteit Eindhoven (2013). https://doi.org/10.6100/IR754833
24. Kamgaing, T., Ramahi, O.M.: Development and application of physics-based compact models for high-impedance electromagnetic surfaces integrated in a power plane configuration. In: IEEE AP-S Symposium Digest, June 2003

25. Kim, Y., Noh, I., Noh, H., Park, J., Yang, K.: A low dark current planar-type InGaAs guard-ring PIN photodiode using an ALD-Al2O3 passivation for short-wave infrared imaging applications. Compound Semiconductor Week 2016 (CSW) [Includes 28th International Conference on Indium Phosphide & Related Materials (IPRM) & 43rd International Symposium on Compound Semiconductors (ISCS), pp.1–2 (2016) MoP-IPRM-027
26. Kenneth, K.O., Kim, K., Floyd, B.A., et al.: On-chip antennas in silicon ICs and their application. IEEE Trans. Electron. Devices **52**(7), 1312–1323 (2005)
27. Kim, H.-J., Park, J., Oh, K.-S., et al.: Near field magnetic induction MIMO communication using heterogeneous multipole loop antenna array for higher data rate transmission. IEEE Trans. Antennas Propag. **64**(5), 1952–1962 (2016)
28. Lee, Y.C., Chang, W., Park, C.S.: Monolithic LTCC SiP transmitter for 60 GHz wireless communication terminals. In: IEEE MTT-S International Microwave Symposium Digest, Long Beach, CA, 12–17 June 2005
29. Liu, Y., He, J., Guo, M., et al.: An Overview of Big Data Industry in China. China Commun. **11**(12), 1–10 (2014)
30. Liu, D., Gaucher, B.: Design consideration for millimetre wave antennas within a chip package. In: Proceedings of IEEE International Workshop Antenna Technology, Xiamen, China, 21–23 Apr 2007
31. Long, J., Montalvo, T.: Wireless receivers for consumer applications. In: Proceedings of IEEE International Conference on Solid-State Circuits, Digest of Technical Papers, pp. 424–425, Philadelphia, 10–10 Feb 2005
32. Ma, K.-P., Kim, J., Yang, F.-R., et al.: Leakage suppression in stripline circuits using a 2-D photonic bandgap lattice. In: IEEE MTT-S International Microwave Symposium Digest, Anaheim Convention Center. Anaheim, California, June 1999
33. Massoud, Y., Majors, S., Kawa, J., et al.: Managing on-chip inductive effects. IEEE Trans. Very Large Scale Integr. VLSI Syst. **10**(6), 789–798 (2002)
34. Ngu, A.H.H., Gutierrez, M., Metsis, V., Nepal, S., Sheng, M.Z.: IoT middleware: a survey on issues and enabling technologies. IEEE Internet of Things J. **PP**(99):1–1 (2017)
35. Noh, I., Noh, H., Kim, Y., Lee, K., Yang, K.: A novel deep guard-ring InGaAs PIN photodiode structure reducing a crosstalk in SWIR imaging detection. 2016 Compound Semiconductor Week (CSW) [Includes 28th International Conference on Indium Phosphide & Related Materials (IPRM) & 43rd International Symposium on Compound Semiconductors (ISCS), pp. 1–2 (2016)
36. Ohata, K., Maruhashi, K., Ito, M., et al.: Wireless 1.25 Gb/s transceiver module at 60 GHz band. In: Proceedings of IEEE Solid-State Circuits Conference, 7–7 Feb 2002
37. Park, J.B., Lu, A.C.W., Chua, K.M., et al.: Double-stacked EBG structure for wideband suppression of simultaneous switching noise in LTCC-based SiP applications. IEEE Microw. Wirel. Compon. Lett. **16**(9):481–483 (2006)
38. Pfeiffer, U., Grzyp, J., Liu, D., et al.: A chip-scale packaging technology for 60-GHz wireless chipsets. IEEE Trans. Microw. Theory Tech. **54**(8), 3387–3397 (2006)
39. Pfeiffer, U., Grzyp, J., Liu, D., et al.: A 60-GHz radio chipset fully-integrated in a low-cost packaging technology. In: Proceedings of 56th Electronic Components Technology Conference, San Diego, CA, 2 June 2006
40. Raman, S., Chang, T., Dohrman, C., et al.: The DARPA COSMOS program: the convergence of InP and silicon CMOS technologies for high-performance mixed-signal. In: Proceedings of International Conference Indium Phosphide Related Materials (IPRM), Kagawa, Japan, 1–4 June 2010
41. Rogers, S.D.: Electromagnetic-bandgap layers for broad-band suppression of TEM modes in power planes. IEEE Trans. Microw. Theory Tech. **53**(8), 2495–2505 (2005)
42. Saiz, N., Dolats, N., Arbabian, A.: A 135 GHz SiGe transmitter with a dielectric rod antenna-in-package for high EIRP/channel arrays. In: Proceedings of the IEEE Custom Integrated Circuits Conference, San Francisco, California, 15–17 Sept 2014
43. Sarawat, K.C., Mohammadi, F.: Effect of scaling of interconnections on the time delay of VLSI circuits. IEEE Trans. Electron Dev. **17**(2), 275–280 (1982)

44. Seki, T., Nishikawa, K., Toyoda, I., et al.: Millimeter wave high-efficiency multilayer parasitic microstrip antenna array for system-on-package. NTT Tech. Rev. **3**(9), 33–40 (2005)
45. Shahparnia, S., Ramahi, O.M.: Miniaturized electromagnetic bandgap structures for broadband switching noise suppression in PCBs. Electron. Lett. **41**(9), 519–520 (2005)
46. Shen, M., Mikkelsen, J.H., Zhang, K., et al.: Modeling and design guidelines for P + guard rings in lightly doped CMOS substrates. IEEE Trans. Electron Devices **60**(19), 2854–2861 (2013)
47. Sievenpiper, D., Zhang, L., Broas, R.F.J., Alexopolous, N.G., Yablonovitch, E.: High-impedance electromagnetic surfaces with forbidden frequency band. IEEE Trans. Microw. Theory Tech. **47**(11), pp. 2059–2074 (1999)
48. Sievenpiper, D., Yablonovitch, E.: Circuit and method for eliminating surface currents on metals. U.S. Patent 60/079953, 30 Mar 1998
49. Tang, M., Lu, J.-G., Mao, J., et al.: A systematic EM-circuit method for EMI analysis of coupled interconnects on dispersive dielectrics. IEEE Trans. Microw. Theory Tech. **61**(1), 1–13 (2013)
50. Tang, H.K.: EMI-induced failures in microprocessor-based Counting. Microprocess. Microsyst. **17**(14), 248–252 (1993)
51. Tsai, HW., Ker, MD.: Active guard ring to improve latch-up immunity. IEEE Trans. Electron Devices. **61**(12) (2014)
52. Tsutsumi, Y., Nishio, M., Sekine, S., et al.: A triangular loop antenna mounted adjacent to a lossy Si substrate for millimeter-wave wireless PAN. In: Proceedings of IEEE Antenna Propagation Symposium, Honolulu, HI, 10–15 June 2007
53. Wheeler, H.A.: Fundamental limitations of small antennas. Proc. IRE **35**(12):1479–1484 (1947)
54. Xu, J., Wang, S.: Investigating a guard trace ring to suppress the crosstalk due to a clock trace on a power electronics DSP control board. IEEE Trans. EM Compat. **57**(3), 546–554 (2015)
55. Yoon, H., Kim, K., O, K.K.: Interference effects on integrated dipole antennas by a metal cover for an integrated circuit package. In: Proceedings of IEEE AP-S International Symposium USNC/URSI National Radio Science Meeting, Salt Lake City, UT, July 2000
56. Yoshikawa, T., Hirata, T., Ebuchi, T., et al.: An over-1-Gb/s transceiver core for integration into large system-on-chips for consumer electronics. IEEE Trans. Very Large Scale Integr. VLSI Syst. **16**(9), 1187–1198 (2008)
57. You, P.-L., Huang, T.-H.: A switched inductor topology using a switchable artificial grounded metal guard ring for wide-FTR MMW VCO applications. IEEE Trans. Electron Devices **60**(2), 759–766 (2013)
58. Zhang, Y.P.: Antenna-in-package technology for modern radio systems. In: IEEE International Workshop on Antenna Technology, Small Antennas and Novel Metamaterials, Chiba, Japan, Mar 2008
59. Zhang, Y.P., Duixian Liu.: Antenna-on-Chip and Antenna-in-Package Solutions to Highly Integrated Millimeter-Wave Devices for Wireless Communications. IEEE Trans. Antennas Propag. **57**(10) (2009)
60. Zhang, Y.P., Sun, M., Chua, K.M., et al.: Antenna-in-package in LTCC for 60-GHz radio. In: Proceedings of IEEE International Workshop Antenna Technology, Cambridge, U.K., pp. 279–282, 21–23 Mar 2007
61. Zhang, Y.P., Sun, M., Chua, K.M., et al.: Integration of slot antenna in LTCC package for 60 GHz radios. Electron. Lett. **44**(5), 330–331 (2008)
62. Zhang, Y.P., Sun, M., Chua, K.M., et al.: Antenna-in-package design for wirebond interconnection to highly integrated 60-GHz radios. IEEE Trans. Antennas Propag. **57**(10), 2842–2852 (2009)

63. Zhang, L., EP, Li., XP, Yu.: Frequency-response-oriented design and optimization of N+ Diffusion Guard Ring in Lightly Doped CMOS Substrate. IEEE Trans. Electromagn. Compat. **59**(2) (2017)
64. Zwick, T., Liu, D., Gaucher, B.: Broadband planar superstrate antenna for integrated millimeter-wave transceivers. IEEE Trans. Antennas Propag. **54**(10), 270–2796 (2006)

Evaluating Decision Analytics from Mobile Big Data using Rough Set Based Ant Colony

Soumya Banerjee and Youakim Badr

Abstract The significance of mobile centric data from various sensors, mobile phones and from other corresponding sources has already been identified across different sections of applications from commercial services to decision making applications. However, uncertainty and volume of mobile big data solicits appropriate analytics and decision making ability to be inferred from such data sources. Primarily, the data source and analytics to be chosen from the perspective of adaptive yet intelligent technique. The proposed chapter elaborates such solution while deploying rough set, which is capable of handling imprecise and uncertain contexts of mobile big data. In addition to, ant colony pheromone deposition and evaporation process assists in optimal feature selection mechanism for resolved decisions. The proposed model is supported by case study of hazards event and the information of the event is propagated through mobile data derived from social network. The data is represented as social tweets and posts. It has been analyzed with rough set based ant colony.

Keywords Mobile big data · Rough set based feature selection · Ants pheromone

1 Introduction

Mobile big data (MBD) is an emerging concept, which considers a substantial yet huge amount of mobile data in different forms from smart phones, mobile sensors and even to social networks. Conventionally, it cannot be processed using a single machine. According to recent CISCO report, it has been revealed that [6],

S. Banerjee (✉)
CNRS, LIRIS, INSA de Lyon, Lyon, France
e-mail: soumyabanerjee@bitmesra.ac.in

S. Banerjee
Birla Institute of Technology, Mesra, India

Y. Badr
CNRS, LIRIS INSA, UMR5205, INSA Lyon, Universite De-Lyon, Lyon, France
e-mail: youakim.badr@insa-lyon.fr

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_9

half a billion mobile devices were globally sold in 2015, and the mobile data traffic grew by 74% yielding 3.7 exabytes (1 exabyte = 1018 bytes) of mobile data per month. In the past few years smart phones remarkably started to carry sensors like GPS, accelerometer, gyroscope, microphone, camera and Bluetooth devices. Relevant application and service offering encompasses from basic information search, to entertainment or healthcare [11, 12]. MBD contains useful information for resolving many complex problems such as fraud detection, predictive marketing and targeted advertising, context-aware computing and healthcare. Therefore, at present MBD analytics is a highly demanding and focused domain aiming at extracting meaningful information and patterns from raw mobile data from multi-sources. Therefore, inclusion of a decision analytics becomes mandatory and in addition to, considering the diversification and multi-sources of mobile big data, an ubiquitous and computationally intelligent methodology for formulating such analytics also has become major concern.

The research challenge includes similarity of mobile big data, their interaction pattern and the different dimensions of data under different contexts. While investigating those relevant intelligent techniques to process mobile big data, couple of straight as well as purely soft computing driven approaches are available, e.g. Chi-FRBCS-BigData algorithm: A MapReduce design based on the fusion of Fuzzy Linguistic Rules and MR-EFS: Evolutionary Feature Selectionfor Big Data Classification: A MapReduce approach [26, 27]. The first model deploys the building of the fuzzy partition using equally distributed triangular membership functions where as the later approaches towards the particular feature selection methods mentioning the Support Vector Machine, Logistic Regression, and Naive Bayes implemented within the Spark framework to address big data problems. Inspired by such evolutionary intelligence for analyzing the big data, contemporary decision analytic framework could be proposed. The rough set based ant colony is another novel bio-inspired yet evolutionary initiative to analyze massive mobile big data and can be finally tuned up with an adaptive and computationally feasible decision analytics framework. The generic feature selection is accomplished by using evolutionary algorithms [14, 31]. Usually, the set of features is encoded as a binary vector, where each position determines, if a feature is selected or not. It allows to perform feature selection with the exploration capabilities of evolutionary algorithms. Conventionally, evolutionary feature selection algorithms are one of the class of algorithms, seldom represent smart evolutionary learning algorithm. In this chapter also, the proposed model, driven by rough set based ant colony, has been presented a hybrid evolutionary class of algorithm, backed up with appropriate learning abilities. The learning ability can assist in suitable classification and in analysis of mobile big data, without generalizing with a particular mapping framework like MapReduce or Hadoop framework. The remaining part of the chapter is organized as follows: Sect. 2 details on the motivational aspects to represent mobile big data with respect to intelligent techniques followed by a snap on emergency management concerning the mobile big data content in Sect. 2.1. Section 3 describes the proposed algorithm on mobile big data focusing a specific context of emergency management. Section 4 discusses the

post implementation of result and sources of test data followed by conclusion and open research challenge of mobile big data paradigm in Sect. 5.

2 Motivation

Manifold and voluminous data are generated every day with the rapid computations and sensors through various range of domains, e.g., search engines, social media, health care organizations, insurance companies, financial industry, retail, and many others [16, 18]. The paradigm of such Big Data, is characterized by 5Vs, i.e., Volume, Velocity, Variety, Value and Variety. Volume implies that the amount of data that needs to be treated could be quiet significant and big. Velocity represents that the speed of data processing should be very high. Variety means that the data is varied in nature and there are many different types of data that need to be properly combined to make the most of the analysis. Value means high yield will be achieved, if the big data is processed correctly and accurately. Veracity means the inherent trustworthiness of data, namely, the uncertainty about the consistency or completeness of data and other ambiguities [20]. We investigate different sources of mobile data, which could be adhered to different contexts and flavor of information. They are likely to be:

- Data at rest
- Data at motion
- Data in various formats and forms
- Data with Low value density
- Uncertain data, which could participate in future inference process.

The sources of data are identified as a data center of mobile data acquisition [21]. It has been explored that till now, computational intelligence approaches and optimization (primarily, the bio-inspired techniques in the form of ant colony) and other evolutionary methodologies are well addressed and deployed in the feature selection and reduction mechanism. There are few relevant contributions, where fuzzy and Dominance-based Rough Set Approach (DRSA) [19] can process information with preferences and the attributes and features are ordered. The approach has been successfully applied in multi-criteria decision analysis and other related issues. Still, purely decision and inference mechanism to combat emergency situations, are not being addressed through intelligence and adaptive decisions, especially, where data interpretations for learning are contextually different and diversified in operation. Hence, there are unique blend of hybrid computational intelligence approaches could be appreciated, which can assist the inherent learning of features part of a problem from such existing mobile contextual data. Subsequently, certain plans or decisions can also be evolved from such mobile data repository. The approach can be precise than conventional statistical approaches and data classification. If the Table 1 data indicator is considered, can there be any resolution of emergency management

Table 1 Data types and variants [21]

Data type	Volume	Suitability for learning
Calls (in/out/missed)	240,227	Data storage & interpretations
MS (in/out/failed/pending)	175,832	Data storage & interpretations
Photos	37,151	Storage
Videos	2,940	Storage & communicate
Application events	8,096, 870	Storage & mark up
Calendar entries	13,792	Storage & mark up
Phone book entries	45,928	Storage & Retrieval
Location points	26,152, 673	Mark up
Unique cell towers	99,166	Storage & mark up
Bluetooth observations	38,259, 550	Storage & communicate
Audio samples	595,895	Storage & Retrieval

scenario, based on social network media or through mobile data set towards specific time stamps for a specific location? This query could be resolved in next subsection.

2.1 Problem Formulation

The paper solicits a synthetic case instances, where certain evaluation indicators and criteria for emergency decision making are considered for an emergency preparedness plan on the basis of mobile big data characteristics. The problem could be more precise, if we consider a scenario where in a part of city location, a major accident outburst while ago. The authority needs to apply the rescue plan and decisions on the affected area as fast as possible. The historical and archival data set is recorded from mobile sources of data at the instance of accident and evaluation indicators such as specificity, completeness, quick response to an emergency, and other related characteristics of the emergency preparedness plan are regarded as a set or universe, denoted V. That is, the universe V stands for all characteristics of the emergency preparedness plan, i.e., $V =$ (strong pertinence (y_1), soundness of personnel and resources allocation (y_2), good intersectorial collaboration (y_3), . ., reasonable cost(in)). Generally speaking, V is finite because the indicators describing the basic features of the plan are limited in number. Meanwhile, we group all the emergency preparedness plans into aset or universe, denoted U, i.e., $U = \{x_1, x_2, ..., x_m\}$, where x_i stands for the ith emergency plan. We call a subset R of $U \times V$, the compatibility relation between the emergency preparedness plan set U and the characteristics set V. That is, for any $x \in U$, $y \in V$, there exists $y_0 \in V$, $x_0 \in U$ Satisfying $(x, y_0), (x_0, y) \in U \times V$. Effectively, the binary compatibility relation R defines a set-valued mapping on universes U and V as follows:

$$M : U \rightarrow 2V, x \rightarrow y \in V | (x, y) \in R \tag{1}$$

That is, for any emergency plan $x(x \in U)$, the basic characteristic is:

$$M(x) : M(x) \in 2V \tag{2}$$

This part of the problem is treated as rough set basis and decisions could be made on the pattern of mobile sourced data, out of which some of them could be indicator of learning or training for the live agents in the application domain known as ants. The decision of ants could here must represent a a risk or loss. if the decision is visualized as action taken by Sensor/IoT based agents. The the skeletonal approach is to sense the appropriate decisions of a context, perceived through mobile big data. Here, ants' pheromone is an indicator for sensing the internal communication. We can investigate certain specific to emergency management pertained with mobile big data and computational intelligence. Georgios Chatzimilioudis et.al mentioned crowdsourcing as a major vertical to smartphones, a taxonomy that classifies the emerging field of mobile crowdsourcing and three in-house applications that optimize location-based search and similarity services over data generated by a crowd [2]. Recently, mobile phone data to tackle problems related to economic development and humanitarian action. In this research, the suitability of indicators derived from mobile phone data as a proxy for food security indicators has been presented [9]. The snaps of mobile big data and its suitability of metrics derived from mobile phone data and call data records can easily analyze the food security and poverty indicator. The analysis comprises of several months of mobile phone activity of a significant portion of the population in specific locational interest, from one large mobile phone carrier. Each call data records contains the following information: caller ID, callee ID, cell tower initiating the call, date and time [5, 33] Substantial tweet records as mobile big data records can predict the status of mental health, depression and suicidal rate. Hence, the application of mobile big data and analytics may reveal even other contemporary methods of characterizing the intrinsic and relational data. Those applications are capable with Spark-based framework and they are intended towards deep learning parameters driven e.g. learning partial models, parameter averaging and dissemination for extracting and analyzing such large data set.

2.2 Challenges of MBD Analytics

Large-scale and high speed mobile networks, portability, and crowdsourcing redefines the trends of mobile big data especially phone data. Recent studies have shown the value of mobile phone data to tackle problems related to economic development and humanitarian action. The rapid usage of mobile devices and high-speed mobile networks is represented through Wi-Fi as well as cellular networks. They introduce massive yet density driven enhancing mobile data traffic. In 2015, 3.7

exabytes of mobile data was generated per month, which is expected to significant rise through 2017 [6]. MBD flows at a high rate, it reciprocates the anticipated delay towards serving mobile users. Long queuing time of requests results in dissatisfied subscribers and increased tariff of delayed decisions. The reason of data veracity is visible in mobile device due to its high portability independently among many locations. Therefore, MBD is represented as nonstationary (volatility) data type. Due to inherent portability, the time duration for which the collected data is valid for decision making can be relatively reduced and thus MBD analytics should be frequently executed to cope with the newly collected data samples. In addition to portability, crowdsourcing could be another highlighted trend of mobile applications for pervasive sensing, which includes a massive data collection from many participating users. Crowd sensing differs from conventional mobile sensing systems as the sensing devices are not owned by one institution but instead by many individual identities across different locations. Definitely, MBD quality is not be ensured (veracity) with diversified blends of locations and users. This aspect is critical for assessing the quality uncertainty of MBD as mobile systems do not directly manage the sensing process of mobile devices. Since most mobile data is crowdsourced, MBD can contain low quality and missing data samples due to noise, malfunctioning or uncalibrated sensors of mobile devices, and even intruders, e.g., impreciselylabeled crowd-sourced data. Subsequently, low quality of data points, affect the analytical accuracy of MBD. Besides, MBD is available in different data types due to the manifold sensors to support mobile devices. Hence, MBD analytics (value) can contribute to better service management pivoted to any application and case study. The appropriate extraction of knowledge and patterns from MBD with different adaptive and intelligent measures could also improve the control and analytics of end users concerning MBD.

2.3 Case Study Plan

To demonstrate the attricultaion, analytics and trend of mobile big data, we consider a simple case to combat emergency situation and context of a disastrous event to a region. The relevant data and input is gathered from mobile bog data source on that event through tweets, posts and shares of individuals. Ideally, the context should be the formulation of an emergency plan with respect to a sudden event like fire hazard in a city and information of the event is being propagated through social media i.e. tweeter. The input of tweeter has been recently projected as a prime tool for managing disaster [7, 13, 29]. However, solution has been demonstrated with rough set based ant colony algorithm. Rough set is primary used for imperfect data, similar to mobile centric big data. Rough set based data analysis starts from a data table called a decision table, columns the table are labeled by attributes, rows. The objects of interest and entries of the table are denoted as attribute values. Attributes of the decision table are divided into two disjoint groups called condition and decision attributes, respectively. Each row of a decision table induces a decision rule, it specifies decision (action, results and outcome), after satisfying a set of specific

conditions. If a decision rule uniquely determines decision in terms of conditions the decision rule becomes certain and deterministic. In contradiction, the decision rule should be uncertain [25].

2.4 Contributions and Parameters of Rough Set Based Ant Colony

The contribution of rough set and ant colony can be demonstrated from the trends and feature selection of mobile big data and thereby improvising the process of ant heuristics to find out the minimal reduces. This is possible with given a dataset for discretized attribute values. As mentioned the algorithm finds a subset (termed a reduct) of the original features using rough sets that are the most informative; all other features can be relinquished from the dataset with minimal information loss. Previous methods employed an incremental hillclimbing algorithm to discover such reduces. However, this often led to feature subsets of a non-minimal size [4]. From their work, Yumin Chen conceived the plan of using rough set relatively in smaller dimension of graph [3]. However, with respect to emergency service rough set driven ant colony can be an optimum tool primary for two reasons: firstly, an emergency decision must often be made in a short period of time using partial or incomplete and inaccurate information, especially in the early stages of the disaster and secondly in most of the cases decisions during emergency has serious results [15]. Classically, rough set as decision theoretic approach and feature selection mechanism as described in [30] established well positioned application pertaining to knowledge discovery, decision analysis and conflict analysis. In this chapter, a novel framework of emergency plan with respect to rough set is proposed. Following the equation (2), rough set model over two universes is composed of stat binary states and 3 corresponding inferences. The set of states is given by $= \{Y, Y^C i\}$ $(Y \in V)$, indicating that an object $y \in M(x)$ $(x \in U, M(x) \subseteq V)$ is in Y and not in Y, respectively. Here, we use the same symbol to denote both a subset of Y and the corresponding state so that no confusion arises. Y denotes a set of the basic characteristics of the ideal emergency plan given by decision makers according to real-time scenarios. Let $A = \{P, B, N\}$ be the action set where P, B and N represent the three actions in classifying an object $x(x \in U, F(x) \subseteq V)$, namely, deciding $x \in POS$ (Y), deciding $x \in BND(Y)$, and deciding $x \in NEG(Y)$, respectively. Hence, therefore to conceive an emergency plan in output stage may comprise of with the following metadata.

- Compactness of context,
- strong pertinence,
- reduced disposal time,
- comprehensiveness of post-disposal context and situation,
- executed precautionary measures,
- scientific analysis of causes,
- optimal resource consumption and minimum rescue workers in operation.

These output parameters could be mapped with associated risks against each action(s). Subsequently, the risk can be declared as dynamic constraints. Corresponding the action aP, aB and aN with dynamic constraint of risk C over rough set of universe, the followings maps can be formulated:

$$C(a_P \mid M(x)) = \Omega_{LP} P(Y \mid M(x)) + \Omega_{PN} P(Y^C \mid M(x)) \tag{3}$$

$$C(a_B \mid M(x)) = \Omega_{LB} P(Y \mid M(x)) + \Omega_{BN} P(Y^C \mid M(x)) \tag{4}$$

$$C(a_N \mid M(x)) = \Omega_{LN} P(Y \mid M(x)) + \Omega_{NN} P(Y^C \mid M(x)) \tag{5}$$

where, $P(Y \mid M(x))$ and $P(Y^C \mid M(x))$ expresses the value of conditional probability, the additive terms represents the risk when the specific sum of the risks and constraints for the respective actions of selecting the particular emergency plan, delaying decision making and not selecting the emergency plan if the characteristic of the emergency plan does not satisfies the requirement of the ideal emergency response plan with respect to rough set universe. However, decision making process consult regular Bayesian probabilistic theory, which delegates optimized risk covariance in emergency plan. To choose a particular $x(x \in U)$, accept or reject an universe or not to choose the specific universe, depending the dynamic constraint C with conditional probability. At this outset, the hybridization with ants' pheromone deposition and evaporation mechanism could be deployed. The necessary input for realizing the degree of emergency and efficiency of emergency combat plan is defined from mobile social big data (e.g. tweets, shares and posts) to be collected over a fixed span of time. The level of positive and negative pheromone is set.The objective is to monitor the trend of mobile big data and after following the trend, rough set based universe will execute the decision but the components of emergency plan and service with constraints is modeled with pheromone. Each context of emergency situation releases pheromone in proportion to the inverse of the output parameters. Adjacent neighborhoods and relevant input in the form of mobile data stream of social media and tweets are upstreams of data links for analytics. The proposed model considers following parameters:

i: current value of pheromone depending on real time input

e: execution value of pheromone, where control executes the actual value of pheromone

d: distribution of spreading the information to alter a specific value of pheromone at time t and place p.

$$e(t,p) = \frac{1}{\mid EM(t,p) \mid} + \sum_{i \in EM(t,p)} \frac{\alpha}{Exe(t,p)} \tag{6}$$

$$i(t+1.p) = Ev \times i(t,p) + d(t,p) \tag{7}$$

$$d(t+1,p) = \sum_{p \in Adj(p)} \frac{F}{\mid Adj(p) \mid} c(t,p) + d(t,p_{\sim}) \tag{8}$$

Algorithm 1: **High Level Description of Rough Set Based Ant Colony**

Input: Emergency decision-making information over two universes (U,V, R)
Output: The optimal emergency plan

1 Presenting the values of loss function λ ,for every emergency preparedness plan $x \in U$.
2 Computing the threshold pheromone deposition and evaporation levels (from mobile big data).
3 Computing the maximum threshold (from mobile big data).
4 Establishing the characteristics of optimal emergency plan Y.
5 Computing the conditional probability$P(Y|F(x))$and then making the decision according to the decision rules (P') based on pheromone measures as followed in Eq. (6), (7) and (8).

3 Proposed Algorithm

The chapter presents a hybrid application based approach, which has rough set as a basis and ant colony should act as learning from the universe and could sense different labels of inferences accordingly. The inference deployment could be carried out though pheromone deposition and evaporation mechanism. Algorithm has been coded in C++ Linux operating system. We implemented the proposed framework on a shared cluster system (https://www.acrc.a-star.edu.sg) running the load sharing facility (LSF) management platform on RedHat Linux. Each node has 8 cores (Intel Xeon 5570 CPU with clock speed of 2.93 Hz) and a total of 24 GB RAM. In the proposed experiments, we set the cores in multiples of 8 to allocate the entire node's resources. One partial model learning task is initialized per each computing core. Each task learns using a data batch consisting of 100 mobile data driven social media data samples for 100 iterations.

4 Discussion and Analysis of Results

In the current study, 100 ants were utilized with 1000 computational iterations for deposition and evaporation of pheromones as followed by the Eqs. (7) and (8). In this proposed model, each ant was assigned to its own set of binding patterns and it deposited an equal amount of pheromone to the assigned set. Pheromone preference factor and pheromone evaporation factor control the pheromone pattern of communication, from tweets and posts, and that will implicate to formulate achievable emergency plan Y. The value of preference signifies (either positive or negative pheromone, depending on the degree of emergency to be fixed for that particular zone) determines the preference to pheromones, and the value of evaporation regulates fading of the previous concentration of pheromone, if the degree of emergency is less, then evaporation time will be faster. However, from the concept used in the simulated environment, rough set constitutes the plan as output, while the pattern of

information or motor of information is searched through pheromone deposition of ant agents.

Ant algorithm explores the cluster of tweets for a group of the best possible real time input. in a space of individual mobile data trends. The former space is apparently more restricted than the latter combinatorial space. Therefore, the ant algorithm has deterministic context to terminate the search with maximum number of emergency plan meta data. In addition to, the advantage of the ant algorithm includes identification of search factor for different tweet cluster analysis towards effective and optimal emergency plan. For a given pheromone map, with different time instances, analytics demonstrate the spectrum, in higher entropy as shown in Fig. 1. A pheromone spectrum was then built from the sum while reproducibility was defined as correlation among the spectra (Refer Fig. 1). A higher selectivity in general implies that the limited number of attribute or pattern candidates across the tweets and posts on the particular instant. They receive a significantly higher pheromone concentration than most of the other candidates just immediate to the emergency context. The extreme case for preference pheromone could be 0 or it could be greater than 100, yielded reduced throughput to asset the rough universe for effective Y. Since, the executable pheromone is predominantly improvised by the ants in the first or the last generation of mobile big data repository. The experimentation demonstrate 3 types of tweets denoted as CU, CH & FF separated as different clusters. The axis referred in the simulation is the standard differentiation initiated as equal time instances for deposition of pheromone. The pheromone distribution mechanism is modeled again a particular case hazards, collected from tweet input. The input tuned with several attributes for

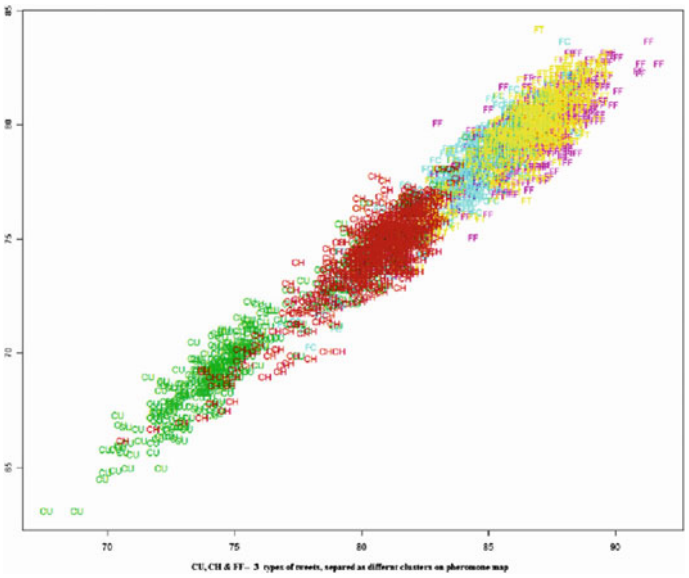


Fig. 1 Pheromone density with mobile data trend

emergency exit process can deliver the required plan as an output. The time and concentration of pheromone from social media tweets categorically modeled with 30, 100 and 170 cm (Fig. 2) in a restricted environment with a time stamp from 20.45 to 21.45 and it has been observed the different initial value of deposition of pheromone (tweets/posts), based on the catastrophic condition of hazards or events. The possible reason for such diversification could be the value differentiation in positive or negative pheromone for an improved label and classification of the context of hazard. In Fig. 3, analytics has been differentiated with respect to different constraints as refered by Eqs. (3), (4) & (5).

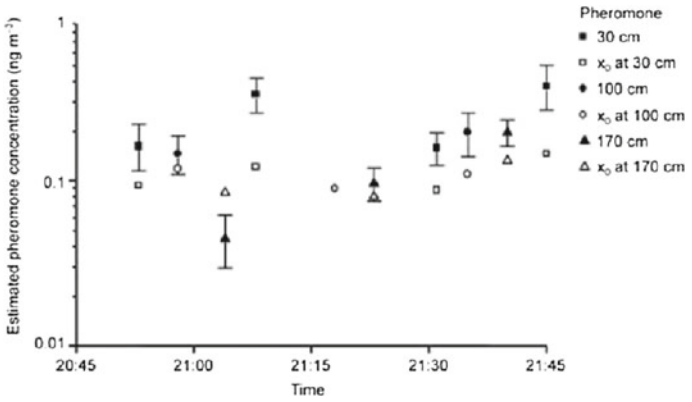


Fig. 2 Positive & negative pheromone concentration

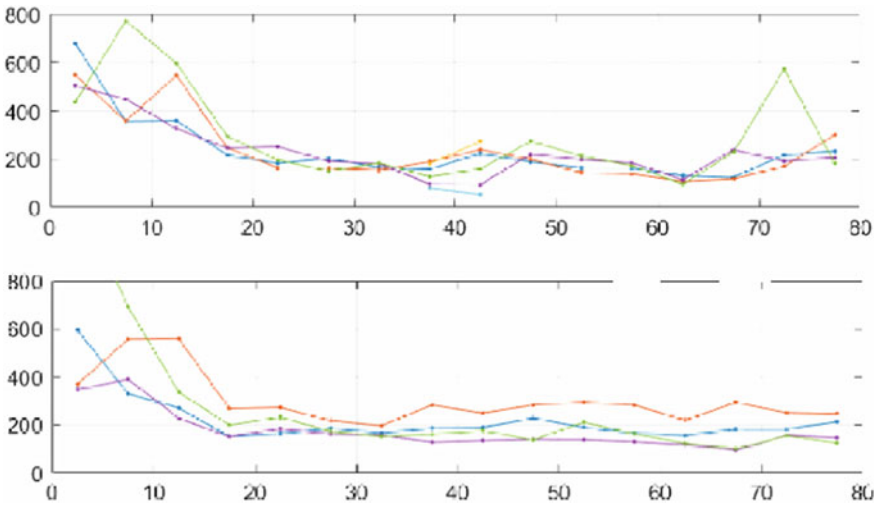


Fig. 3 MBD analytics with constraints

5 Comparison and Performance

Envisaging the rough set and rough set driven ant colony could demonstrate a contemporary comparison in the analysis of mobile big data. As the case study reveals here, the tweets/posts fetched from social media about an accident or hazard occurred at a specific time span. The standard dataset contains 2000 reviews comprising 1000 positive and 1000 negative reviews (considering a homogeneous distribution). Each span has 1000 positive and 1000 negative labelled reviews. Documents are initially pre-processed as follows:

- Performing with negative expression in sentences and opinion is accomplished in [24], NOT is inserted to every words occurring after the word of negative expression (e.g. no, not, isn't, can't, never, couldn't, didn't, wouldn't, don't) and first punctuation mark in the tweet/posts.
- frequency of occurrence of words in less than 3 documents are removed from the feature set. Binary weighting scheme has been identified as an efficient weighting scheme as compared to frequency based schemes for opinion classification.
- Noisy and irrelevant features are eliminated from the feature vector generated after pre-processing using various feature selection methods discussed before. However, prominent feature vector analysis is used by machine learning algorithms. Support Vector Machine (SVM) and Naive Bayes (NB) classifiers are the prominently classifier used for sentiment classification [32]. Therefore, the comparison is pivoted on the classification results of SVM and NB classifier for classifying review documents into positive or negative sentiment polarity about the event from media tweets/posts on a specific time instance. For the empirical evaluation of proposed methods, 10 fold cross validation method is used.

The results deploys F-measure as an analytical measure [35]. F measure can be viewed as a compromise between recall and precision. It is high only when both recall and precision are high. Practically, for any information retrieval (IR) system which has recall R and precision P on a test document collection and an information need. At this point the F-measure of the system is defined as the weighted harmonic mean of its precision and recall. Initially, unigram features are extracted from the review documents. Feature set without using any feature selection method is considered as a benchmark.

Feature vector lengths for various features used for sentiment classification of different datasets are shown in Table 2. In the simulations, Firstly, rough set algorithm is applied to fetch the best optimal feature subset for the set of given posts. Additionally, according to the feature subset size obtained from threshold, is set for ant's pheromone map, which is further used for classification. When hybrid features selection approach is used for application event dataset (Refer Table 1) and location points, F-measure is enhanced from 84.2 to 87.7% (+ 4.15) for SVM classifier as given in Table 1. Hybrid i.e. rough set combined with ants pheromone features gives better classification results as compare to other features with very small feature vector length. It is due to the fact that pheromone map (either positive or negative) can

Table 2 Statistical measures with data set and proposed model

Data types & set	Extraction	Unigram features	Pure rough set features	Proposed rough set based ant colony features
Application events (8,096,870)	SVM	84.2	85.9(+2.1)	86.6(+3.4)
	NB	76.9	77.8(+2.1)	80.5(4.3)
Location (26,152,673)	SVM	77.2	80.3(+3.1)	83.4(+6.1)
	NB	73.2	76.4(+2.2)	78.2(+5.0)

eliminate the irrelevant and noisy features and thus rough set based ant colony reduces redundancy among features and can extract the optimal feature subset. Accordingly, decision on mobile data could be made on this subset. Thus, in both instances, rough set based ant colony is found to be a better classifier for analyzing dynamic and voluminous of instant mobile data.

6 Conclusion and Future Research Scope on Mobile Big Data

Mobile big data storage and analytics has become a major proliferation in recent cyber physical activities. The present chapter is an initiative to introduce computational intelligence to manage and analyze mobile driven big data content. The mobile big data importantly contribute consensus in decision making and also can assist to formalize an optimal emergency plan. Rough set driven method yield to generate and compare effective rules under uncertainty in emergency fire hazard condition from given input feed through social media. The direction and message concerning the emergency is modeled as positive and negative pheromone. The proposed model thus presents an analytics comprising of Ants' pheromone deposition and evaporation mechanism and also rough set based decision making process. In brief, the dashboard and control can add on the proposed analytics to formulate different instances of emergency plan to combat the hazard. We find interesting propositions as mentioned in the s section with convergence and firing of different rules , interpreted through rough set. The final demarkation (refer Fig. 3) is denoted through pheromone spectrum. It signifies that if pheromone deposition and type is positive and rules are compatible Compactness of context, strong pertinence, reduced disposal time and optimal resource consumption and minimum rescue workers in operation, then an optimal yet timely emergency disaster management plan could be prepared. To support the proposed experimentation, public data set is used at different stages with respect to time, distribution of information and place of context. To demonstrate the

relevance of bio-inspired algorithm and rough set, other contemporary evolutionary and population based algorithms also could be inspected [1, 36]. It is observed that, ant colony and the pheromone deposition mechanism is adequate in sequential covering strategy for inducing classification rules [22, 23]. The major challenges in big data practice are identified as the data modeling, computing model, and implementation platform. However, off-shoot parameters and relationship of mobile big data for analytics could be trivial. As the main stream of big data research includes computing model, deep learning [17], MapReduce [8], and Platform, such as Hadoop [34] and Apache Spark [10], therefore proposed computational intelligence application and control mechanism can be leveraged in the form of development of application program interface (API) core library and thus pattern of mobile data trend can easily be analyzed. Even in emerging mobile data analytics, graph-based approaches to data analysis has become more pertinent. Thus, with the incorporation of Property Graph Query Language (PGQL) [28] graph pattern matching on mobile big data, may closely adopt syntactic structures of SQL, and provides regular path queries with conditions on labels and attributes for more accessibility and complex pattern finding queries.

References

1. Cheng, S., Z, Q.Q.Q. : Big data analytic with swarm intelligence. Ind. Manag. Data Syst. (2016)
2. Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., Zeinalipour-Yazti, D.: Crowdsourcing with smartphones. IEEE Int. Comput. **16**(5), 36–44 (2012)
3. Chen, Y., Miao, D., Wang, R.: A rough set approach to feature selection based on ant colony optimization. Pattern Recogn. Lett. **31**, 226–233 (2010)
4. Cheng, S., Liu, B., Ting, T.O., Qin, Q., Shi, Y., Huang, K.: Survey on data science with population-based algorithms. Big Data Anal. **1**(1), 3 (2016)
5. Choudhury De, M., Kiciman, E., Dredze, M., Coppersmith, G., Kumar, M.: Discovering shifts to suicidal ideation from mental health content in social media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 2098–2110 (2016)
6. Cisco: Cisco visual networking index: global mobile data traffic forecast update 2015–2020, White Paper (2016)
7. Cooper, G., Yeager,V., Burkle, F., Subbarao, I.: Twitter as a potential disaster risk reduction tool. part 1: introduction, terminology, research and operational applications. PLoS Curr. Disast. (2015)
8. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Commun. ACM **51**, 107–113 (2008)
9. Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J., Krings, G., Gutierrez, T., Blondel, V.D., Luengo-Oroz, M.A.: Estimating food consumption and poverty indices with mobile phone data. CoRR. https://doi.org/abs/1412.2595 (2014)
10. Donoho, D.: 50 Years of Data Science. Technical report, Stanford University, (2015)
11. Eagle, N., Pentland, A., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proc. Nat. Acad. Sci. **106**(36), 15274–15278 (2009)
12. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
13. Houston, J.B., Hawthorne, J., Perreault, M.F., Park, E.H., Goldstein Hode, M., Halliwell, M.R., Turner McGowen, S.E., Davis, R., Vaid, S., McElderry, J.A., Griffith, S.A.: Social media

and disasters: a functional framework for social media use in disaster planning, response, and research. Disasters **39**(1), 1–22 (2015)
14. Iglesia de la, B.: Evolutionary computation for feature selection in classification problems. Wiley Interdis. Rev. Data Mining Knowl. Disc. **3**, 381–407 (2013)
15. Jia, X., Tang, Z., Liao, W., Shang, L.: On an optimization representation of decision-theoretic rough set model. Int. J. Approx. Reason. **55**(1), 156–166 (2014)
16. Jin, X., Wah, B.W., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Res. **2**(2), 59–64 (2015)
17. LeCun, Y., Bengio, Y.: H.G: Deep learning. Nature **521**(4), 36–44 (2016)
18. Li, T., Lu, J., Luis, M.: Preface: intelligent techniques for data science. Int. J. Intel. Syst. **30**(8), 851–853 (2015)
19. Li, S., Li, T., Zhang, Z., Chen, H., Zhang, J.: Parallel computing of approximations in dominance-based rough sets approach. Know. Based Syst. **87**, 102–111 (2015)
20. Luo, C., Li, T.: Incremental Three-Way Decisions with Incomplete Information, pp. 128–135. Springer International Publishing, (2014)
21. Nokia: https://research.nokia.com/mdc, Nokia Research
22. Otero, F.E., Freitas, A.A.: Improving the interpretability of classification rules discovered by an ant colony algorithm. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation, GECCO '13, pp. 73–80 (2013)
23. Otero, F.E., Freitas, A.A., Johnson, C.G.: Inducing decision trees with an ant colony optimization algorithm. Applied Soft Computing **12**(11), 3615–3626 (2012)
24. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**, 1–135 (2008)
25. Pawlak, Z.: Rough Sets-Theoretical Aspects of Reasoning About Data. Kluwer, Boston, London, Dordrecht (1991)
26. Peralta, D., Rio, S., Gallego, S.R., Triguero, J.B.I., Herrera, F.: Evolutionary feature selection for big data classification: a mapreduce approach. Math. Prob, Eng (2015)
27. Rio, S., Lopez, V., Benitez, J., Herrera, F.: A mapreduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. Int. J. Comput. Intell. Syst. **8**, 422–437 (2015)
28. Sevenich, M., Hong, S., van Rest, O., Wu, Z., Banerjee, J., Chafi, H.: Using domain-specific languages for analytic graph databases. PVLDB **9**(13), 1257–1268 (2016)
29. Shannon, C.: Understanding community-level disaster and emergency response preparedness. Disaster Med. Public Health Prepared. **9**(3), 239–244 (2015)
30. Sun, B., Ma, W., Zhao, H.: A fuzzy rough set approach to emergency material demand prediction over two universes. Appl. Math. Model. **37**(10–11), 7062–7070 (2013)
31. Tan, I.W.T.M., Wang, L.: Towards ultrahigh dimensional feature selection for big data. J. Mach. Learn. Res. **15**, 1371–1429 (2014)
32. Tan, S., Zhang, J.: An empirical study of sentiment analysis for chinese documents. Expert System with Applications **34**(4), 2622–2629 (2008)
33. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 3187–3196, (2015)
34. White, T.: Hadoop: The Definitive Guide, 4th edn. O'Reilly Media Inc, Sebastopol (2015)
35. Zhang, E., Zhang, Y.: F-Measure, pp. 1147–1147. Boston, MA: Springer US, (2009)
36. Zhou, Z.H., Chawla, N.V., Jin, Y., Williams, G.J.: Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. IEEE Comput. Intell. Mag. **9**(4), 62–74 (2014)

Energy-Aware Issues for Handling Big Data in Mobile Cloud Computing

Chhabi Rani Panigrahi, Rajesh Kumar Verma, Joy Lal Sarkar and Bibudhendu Pati

Abstract The popularity of mobile devices has been growing at a very fast rate and it is evident from the fact that it is possessed by almost each and every person and some may have even more than a single mobile device. MCC helps in computation and running of various complex applications on the mobile device and also offloads to the cloud when it requires lot of resources for computation or storage purposes. However, as energy is limited in the mobile device, processing of complex applications using big data is a challenge that needs to be addressed using energy efficient architectures. In this work, we mainly focuses on identifying energy-aware issues for handling big data in Mobile Cloud Computing (MCC) environment and their current solutions. Also, we have included the review of few techniques available to handle big data in mobile devices. This chapter will also include a brief discussion of techniques available to process big data in MCC in an energy efficient manner. Finally, we conclude with an analysis of identified issues for handling big data in MCC and future scope of research.

1 Introduction

In today's world, around 2.5 Quintillion bytes of data gets generated every day [1]. In fact, data is growing at a very fast pace as there are huge number of data

C.R. Panigrahi · J.L. Sarkar (✉)
Department of Computer Science, Central University of Rajasthan, Ajmer, India
e-mail: joylalsarkar@gmail.com

C.R. Panigrahi
e-mail: panigrahichhabi@gmail.com

R.K. Verma
Cloud and Infrastructure Services (CIS) Lab, Infosys Limited, Hyderabad, India
e-mail: rajeshverma_chicago2004@yahoo.com

B. Pati
Department of Computer Science and Engineering, C.V. Raman College of Engineering, Bhubaneswar, India
e-mail: patibibudhendu@gmail.com

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_10

producing devices like micro sensing devices, software logs/alerts/notifications, cameras, Radio-Frequency Identification (RFID) and Wireless Sensor Networks (WSN) [2, 3]. The big data needs to be processed in an energy efficient manner and useful information must be derived from this big data, which will be beneficial for the society. Big data refers to data sets or combinations of data sets having features such as big size (volume), complexity (variability), and rate of growth (velocity) (3 V's according to Gartner's research). The conventional technologies and tools, such as relational databases cannot process this big data within the required period of time to make it useful. Most analysts and practitioners currently refer to data sets ranging from 30 to 50 terabytes to multiple petabytes as big data. Also, big data helps to resolve the current challenge of processing large amounts of data of heterogeneous type (it may be either structured, semi or un-structured data). Also, it helps by doing the processing in parallel (like a sorting of an entire country's census data) [4].

The growth of mobile devices in the world has been so rapid that it has far exceeded the human population of the world [5]. The number of smart phone users in the world is greater than 2 billion, and the number is expected to touch 2.7 billion by 2019. Also, greater than 60% of the web traffic is generated from mobile devices, which depicts a radical shift from the conventional desktops and laptops. This leads to generation of humongous volume of data which is coming at huge velocity and has variety in it. In order to handle this type of mobile data flowing to and fro, big data comes to the rescue, as it efficiently handles all related operations and delivers value to the user [6].

MCC refers to the computation which is generally initiated by the mobile device and done both on the Smart Mobile Device (SMD) and the cloud (through offloading process), and finally the results are displayed to the user on his mobile device [7]. The offloading approach is used in case where the computation is pretty complex and the vast cloud resources are required to perform the computation and storage activity. MCC emerges as a new computing paradigm where mobile devices exploit the available cloud computing platform for performing specific tasks and/or accessing data on demand. With the widespread exploitation of information, an increasing number of academic researches and industrial applications result in the appearance of big data from multiple heterogeneous sources in mobile clouds. Storage, transmission, analysis, and processing for such heterogeneous big data are crucially required [8]. Mobile devices have certain limitations such as limited battery energy, CPU, storage, and network bandwidth. So, it is important to consider the battery power of mobile devices in a MCC environment [9].

As per current research on Application Programming Interface, approximately 200 exabytes in 2014 and an estimation of 1.6 zettabytes in 2020 is supposed to be processed, 90% of these data are currently processed locally and the processing rate increases day by day. In the same time the risk of critical data theft, data and device manipulation, falsification of sensitive data as well as IP theft, manipulation and malfunction of server and networks also can not be avoided [10]. There is a great impact of data consolidation and data analytics in network configuration i.e. CISCO, HPE and others. Next, in application platform areas based on clouds and firewalls at the network boundaries are more prone from external attacks [10].

There can be several issues with respect to mobile big data such as latency, faster processing, energy etc. and among them energy plays a significant and indispensable part in all mobile related operations [11]. Hence, it is of utmost importance that one needs to concentrate on conserving energy in the mobile devices and also using it in an effective manner for successful running of applications for a good amount of time. When trying to run big data programs or storing huge amounts of data, it is generally not possible to do so on a mobile device. Under these circumstances, the offloading process comes to the rescue, wherein the computation and storage is done on the cloud. Of course, we need to weigh the advantages of performing the operations on the cloud with respect to the mobile device, as offloading to the cloud does involve a certain amount of energy besides the cost of using the cloud for our computational purposes. The current architecture for processing big data has several limitations. When the integration of big data with MCC is considered, it is advantageous as it can easily offload a huge chunk of computation and storage onto the cloud. For this several parallel programming paradigms have such as Map Reduce and others have been developed to process these big data. Several algorithms and architectures have been developed in the area of MCC that emphasize on energy savings [12]. The existing systems such as ThinkAir, Dream, eTime and others deliver efficient mechanisms of saving energy using offloading mechanism and help to execute mobile applications both in the mobile device and the cloud in an efficient manner [12, 13]. Also, several efficient Map Reduce algorithms have been implemented which require very less energy to execute the mobile big data [14, 15]. The vision here is to introduce a few techniques and architectures which will help to ensure that the big data related computations using MCC are handled in an efficient manner using optimum amount of energy of the mobile devices. Towards this direction of energy efficiency, the authors have made a humble effort to put forth the various architectures in order to achieve energy efficiency in case of MCC.

2 Energy Management Techniques in MCC with and without Big Data

2.1 Energy Efficiency of Mobile Devices Using MCC

Mobile Cloud Computing (MCC) refers to the integration of three different technologies, namely, cloud computing, mobile computing and wireless networks in order to accomplish the computational tasks required by mobile users [16].

The growth of mobile devices (Smart Mobile Devices—SMD's, tablets, laptops, etc.) over the years has been very rapid (around 7.22 billion mobile devices in 2016) that their number has already exceeded the entire population of the world (around 7.19 billion). SMDs are small in size and constrained with respect to their processing power, battery life, network bandwidth and storage. In order to complement the limitations of SMDs, Mobile Cloud Computing (MCC) plays an important role by

helping to offload the complex computation and storage requirement to the cloud which has unlimited processing power and storage [17]. Various researchers have proposed different mechanisms of saving energy which are discussed in the following subsections.

2.1.1 Phone2Cloud

In Phone2Cloud, offloading happens which saves energy required for computation [17]. Phone2Cloud deals with offloading approach which leads to substantial energy savings on Smart Mobile Devices (SMD) in MCC. SMDs like smart phones, tablets, etc., have become very common amongst the users in the world as they can run variety of intelligent applications (in health care, military, gaming, etc.) either on the device itself or by offloading to the cloud.

The YoY (Year on Year) growth of mobile devices is greater than 50% and the total number of mobile devices has far exceeded the total world population. Besides this, mobile access is popular as compared to fixed internet access, and this is evident from the fact that the number of mobiles has already far exceeded the number of Desktop computers in 2014. There are four fundamental approaches related to SMD's for preserving energy consumption which leads to an extended battery life.

1. Use of smart battery models and energy cost models, wherein there are various battery models such as ideal model and other models which are present.
2. Avoiding energy wastage, can be achieved by putting the state of the component (either the entire system or the individual components such as the processor) to a sleep state such that it does not hamper the user's work.
3. Savings related to communication, wherein the radio interface used by the Wi-Fi is automatically shut down by the power management mechanism. Another mechanism known as Coolspots can be used wherein switching happens between Wi-Fi and Bluetooth (consumes lower energy).
4. Savings through offloading of computation, as computationally intensive tasks can be shifted to be executed in the cloud which leads to SMD energy saving. Also, techniques which partition an application and only partially offload the tasks to the cloud are used.

Figure 1 shows the basic arcitecture of Phone2Cloud and there are seven basic components which comprise the architecture of Phone2cloud [17].

- *Bandwidth monitor and Resource monitor*: The bandwidth monitor monitors the current bandwidth being used and helps to make the offloading decision. Resource monitor will monitor the SMD status.
- Execution time predictor, an important component which tells us the total execution time of the application on the SMD.
- Offloading decision engine, which actually decides whether to offload and the extent of application to be offloaded from the SMD to cloud.

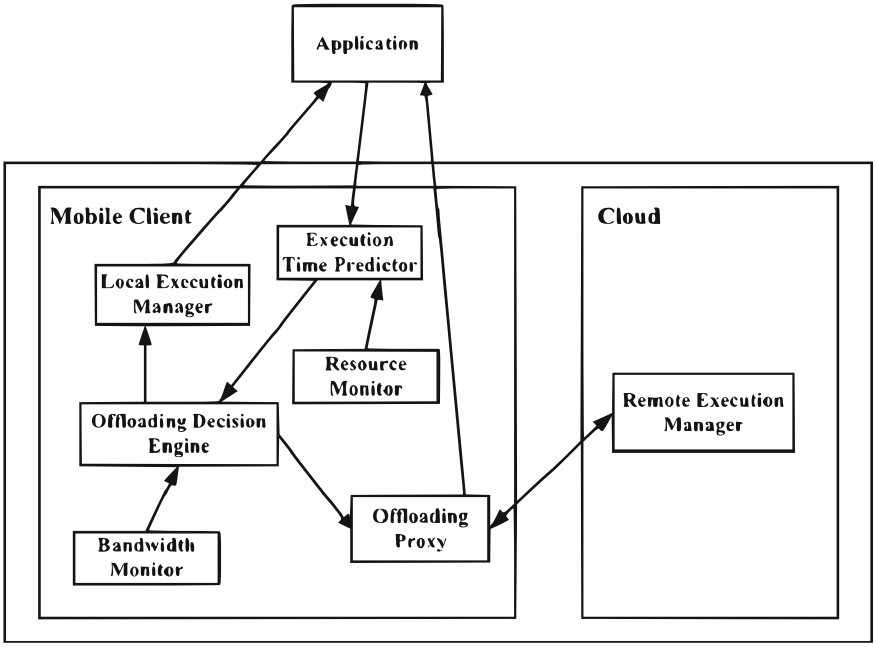


Fig. 1 Basic architecture of Phone2Cloud [17]

- Offloading proxy, which does the task of sending data to remote execution manager and the returned results after computation are sent to the application.
- Local execution manager and remote execution manager, both of these are responsible for managing the execution of application. As the name suggests, the local execution manager will execute application on the SMD (using OS like iOS or Andriod), and the remote execution manager will execute on the cloud.

Figure 2 shows the working flow for taking the offloading decision that means offload to the cloud or smartphone itself. Initially, an Execution Time Predictor (ETP) predicts the average time for execution denoted by T_{exec} and determines the users delay-tolerance threshold denoted by T_{delay} which is then compared to the T_{exec}. If $T_{delay} > T_{exec}$ then calculate the overall energy consumption for running an application on the smartphone denoted by E_{local} otherwise, the application is offloaded to the cloud. The energy consumption for running application on the cloud is denoted by E_{cloud} and by comparison with E_{local}, Phone2Cloud takes the decision for offloading. Figure 3 shows the application and scenario experiments used in case of Phone2Cloud.

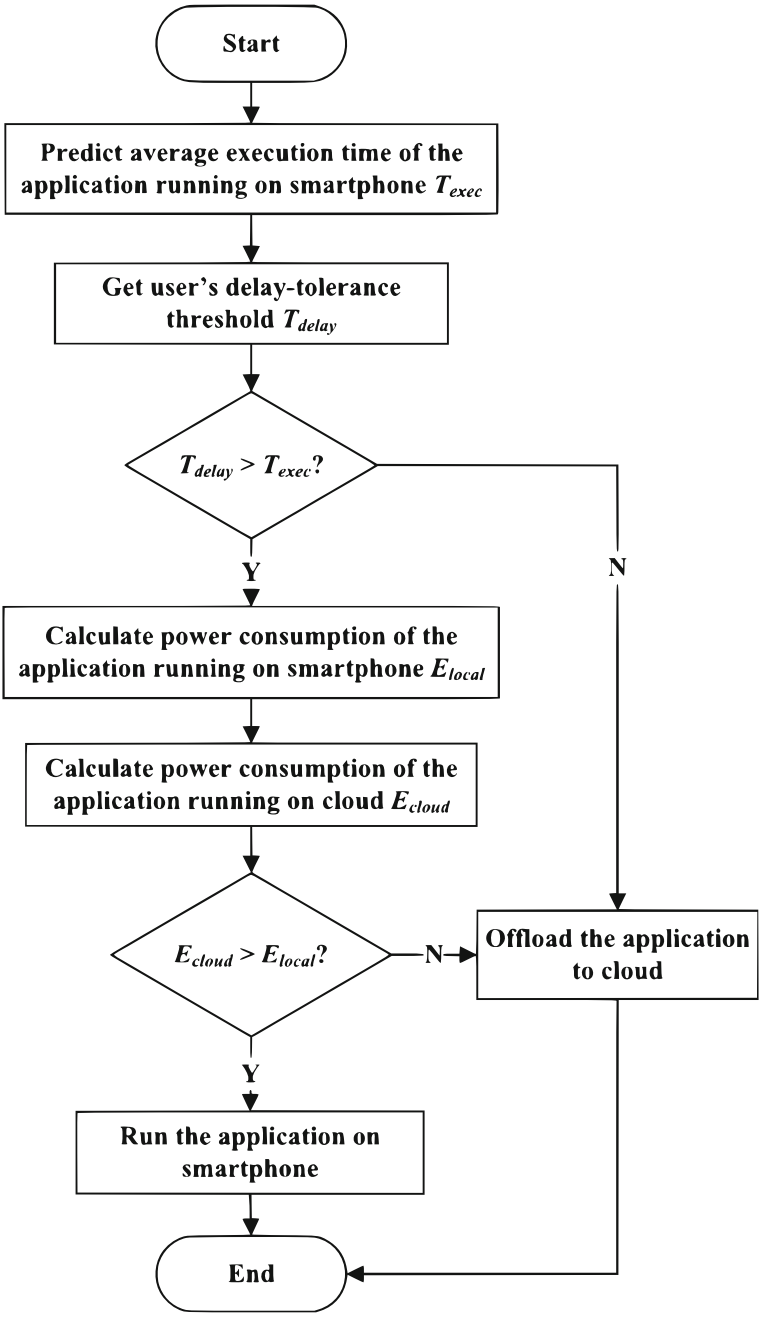


Fig. 2 Working flow for taking offloading decision

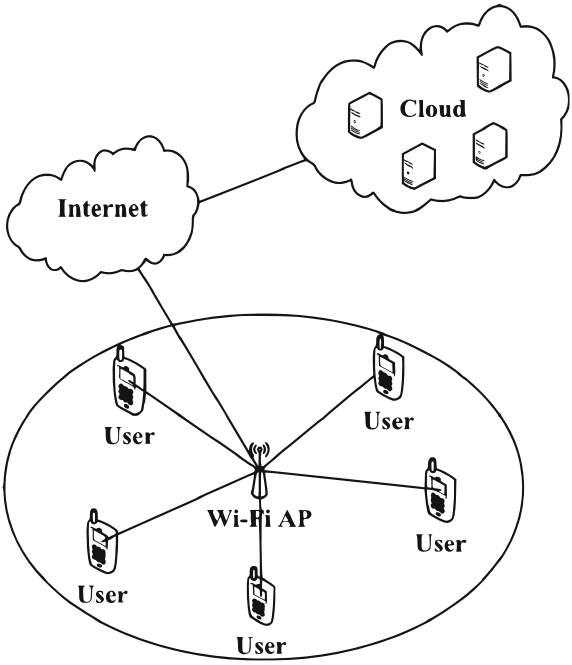


Fig. 3 Basic architecture of Phone2Cloud

2.1.2 eTime

eTime is about the energy efficient transmission between the cloud and mobile devices [13]. The need for offloading arises as the Smart Mobile Device (SMD) do not have enough resources and energy to perform complex computations and store huge data. The programs are offloaded from the mobile to the cloud such that they are efficiently executed on the cloud and the results are sent back to the SMD. It uses the Lyapunov technique and saves around one-third of the energy for different type of applications which are present on SMD's. Also, eTime has intelligence built into it such that it takes advantage of the presence of good network connectivity to perform the offloading process. If the bandwidth of current downlink is $\omega(t)$ and a transmission decision d(t) $\in \alpha = \{transmission\ of\ P_i(t), idle\}$. Where, in time slot t, $d(t) = transmission\ of\ P_i(t)$ when the data queuing transmit in $P_i(t)$ of an application i and $d(t) = idle$ when for saving the energy the data transmission deffered. Therefore, the data transmitted from mobile to cloud denoted as $\varphi(t)$ is expressed as follows:

$$\varphi(t) = \begin{cases} \omega(t)\beta, if\ d(t) = transmission\ of\ P_i(t); \\ \omega(t) = idle; \end{cases} \tag{1}$$

the time span of one time slot is represented by β eTime takes an intelligent decision which is expressed as follows [ ]:

$$\begin{aligned} &minimize\ Q = \lim_{t\to\infty} Sup \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}|D(t)| \\ &Subject\ to\ P < \infty, d(t) \in \alpha \end{aligned} \tag{2}$$

where, D(t) is the data transmission on a mobile device at the time slot t and T is the long time period and Q represents as the averaged energy consumption of the mobile device.

2.2 DREAM

DREAM stands for Dynamic REsource and task Allocation for energy minimization in Mobile cloud systems (DREAM). Offloading is a popular approach followed to save energy of Smart MobileDevices (SMD) in Mobile Cloud Computing (MCC) [18].

There are several factors that needs to be considered in real world like the offloading policy to the cloud, task allocation for execution on local mobile environment or to transmit to cloud and the CPU speed. The DREAM algorithm can save greater than one-third of the energy than the current algorithms. Figure 4 shows the architecture of DREAM and the different components in this architecture are as below:

- Mobile applications: can be of various types, such as mobile games, email, etc. These may or may not run in the mobile device.
- Mobile Device: consists of SMD.
- Application Interface: which is an interface from the mobile applications to the SMD. This interface is generally based on REST based API's.
- 3G/LTE: It is the network which is either 3G or Long Term Evaluation based and transmits the packets from the SMD to the cloud.

2.3 ThinkAir

In current world, Smart Mobile Devices (SMD) is becoming very popular and the number of SMDs is increasing every day. This popularity of SMDs is forcing developers to put their best of applications (like gaming applications, speech recognition, complex graphical applications, etc.) on the SMDs, so that it can be used by a large number of users and provide a good user experience in the hands of the user. These applications are quite complex in nature and also demand a lot or resources in terms of computational power and battery energy. This requires the support of offloading approach and "ThinkAir" is one such efficient framework used by the developers of these complex applications to offload their applications to the cloud in an efficient manner which enables the applications to run very easily and saves energy [12].

Figure 6 shows the architecture of ThinkAir and keeping the architectural viewpoints in mind about the easy use by the development team, wherein the interface is very simple and hence the developers find it very easy and convenient to use. Also, this framework takes care of not compromising on the performance part of the execution of any complex application and hence it is very easy for any new developer to use without causing any risk or delay for any cloud computation mechanism. This architecture also improves the performance of computation through offloading, and the speed of computation is increased and also less energy is used of the mobile device as the complex methods are offloaded and executed in the cloud and the end result is returned to the calling program. Also, ThinkAir facilitates Parallel execution for faster execution of complex applications, wherein the different methods are executed in parallel across the different Virtual machines (VM's), which helps to reduce the time required and also executes efficiently thereby delivering optimum performance of running complex applications.

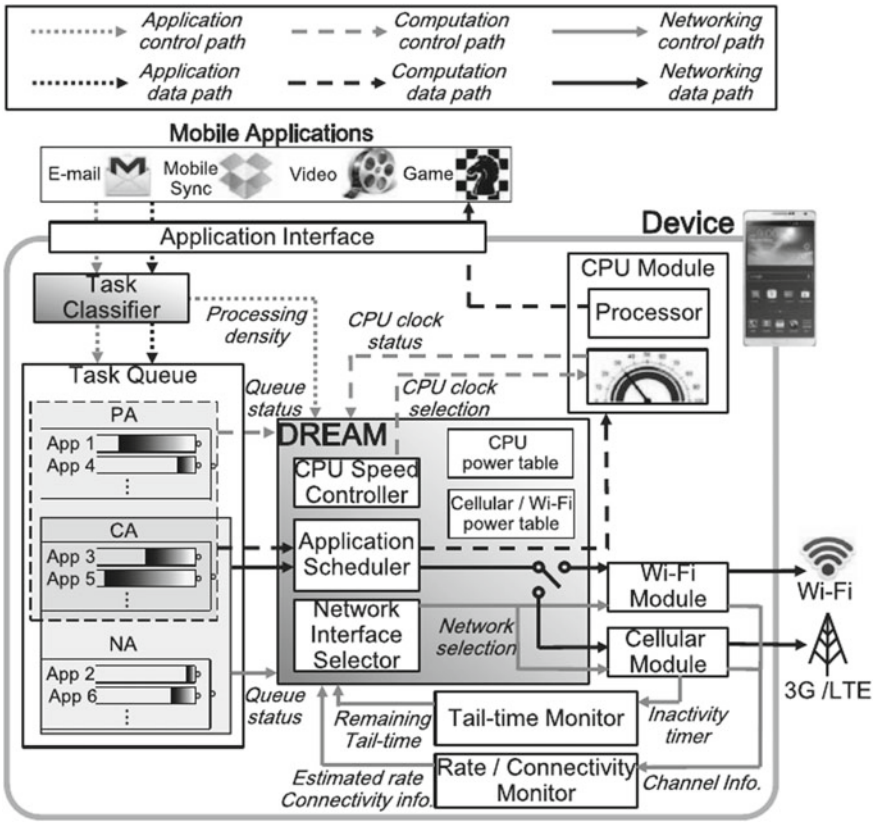


Fig. 4 Dream architecture

The different components of ThinkAir architecture are the Execution platform which is responsible for running of the complex methods in the cloud. It is mainly responsible for making the important decision of whether the particular method should be offloaded to the cloud or be run on the SMD itself. Intelligent algorithms are used which make the decision based on gathered data as well as the previous running of different methods of similar type. Also, the application server component is a very light weight component on the cloud which takes care of managing the code which is offloaded. This server is automatically started when the system is rebooted. The 3 main sub-components of this server includs the client handler, which is mainly responsible for the communication between the SMD and cloud. Besides, cloud infrastructure forms the part of the application server which is pretty flexible as it can be deployed on any of the virtualized cloud environment, whether public or private. Another component called the Profilers is also light weight component and is mainly of 3 different types like software, hardware and network. The Hardware profiler is responsible for inputting the hardware information into the energy model, and subsequently the decision to offload is taken. Different states of CPU, WiFi, Screen and others are monitored using the Hardware profiler. On the other hand, the software profiler is responsible for tracking the parameters connected to program execution like CPU details, time of execution, methods invocation number and others. The most complex of the profiler is the network one which is mainly responsible for collecting metrics like Round Trip Time (RTT), which in turn helps to get the network bandwidth which is very essential for taking decision about offloading (Fig. 5).

2.4 Energy Management Techniques in MCC with Big Data

There are a wide variety of applications available in mobiles ranging from human health care, military, environment, etc. [14]. The mobile applications are complex in nature, and need to process big data in order to come up with results. Under these circumstances, it is very necessary that we need to ensure that we are fully aware about energy related issues in case of mobile big data processing. Also, we need to ensure that we use different techniques and approaches in order to resolve this energy problem. The reason that these energy issues arise in case of mobile devices is because of the fact that the mobile is a small device and the battery is relatively small and hence cannot store huge energy [5]. Subsequently, battery life is less and we need to ensure that the energy is spent efficiently. The different sources of Big Data streams that come into the Mobile devices are shown in Table 1. Also, there are different formats of content that needs to be taken care.

The term Big data refers to data sets (or the combinations of data sets) whose size (volume), complexity (variability), and rate of growth (velocity) 3 Vs (Gartner) make them very difficult to be captured, managed, processed or analyzed by the present day conventional technologies and tools such as RDBMS (Relational Database Management Systems) and desktop statistics or visualization packages, within the time necessary to make them useful and advantageous for the user [19]. The size

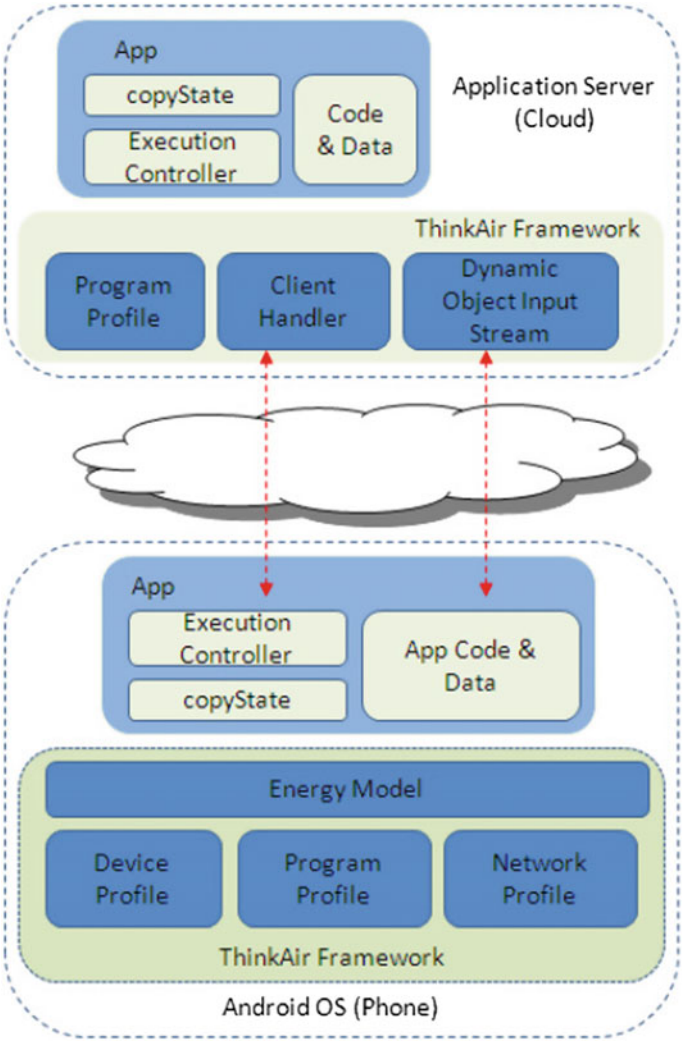


Fig. 5 Basic architecture of ThinkAir

that is fixed in order to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, however, most of the analysts and practitioners currently refer to data sets from 30–50 terabytes (10^{12} or 1000 gigabytes per terabyte) to multiple petabytes (10^{15} or 1000 terabytes per petabyte) as big data. Big data is used to process large amounts of structured, semi-structured or unstructured data like analyzing log files, etc. Also, when the processing can easily be made parallel like a sorting of an entire countries census data, big data becomes very handy. While running batch jobs, big data can prove to be of big help (for example

Table 1 Big data stream sources

Classifications	Descriptions
Social media	Social media helps for communication between the various communities and the social networks such as Twitter, Facebook etc.
Crowdsourcing	Where different kind of people are use SMDs for successfully complete their task
Internet of Things (IoT)	Here the thin devices (e.g., smartphones, tablets, RFIDs, PDAs) have IP addresses and communicate over the internet for enabling services of various types

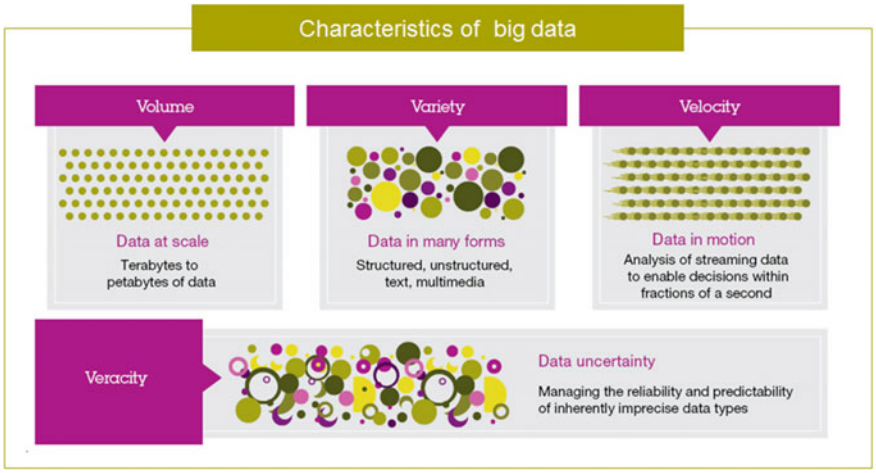


Fig. 6 Characteristics of big data

website crawling by search engines). As Big data involves lots of data it is specially advisable that you run these programs on cheap hardware (commodity hardware). It is not recommended that we use Big data to process GBs or few TBs of data [19].

The characteristics of big data are depicted in Fig. 6 [20].

The current data is growing big because of several factors such as advancements in Mobile Technology, better communication networks, cloud The current data is growing big because of several factors such as advancements in Mobile Technology, better communication networks, cloud computing availability, etc. The big data solution helps to solve various aspects like fraud detection, get insights into the call center records, discover customer sentiments from social media and better customer segmentation for improved sales and marketing.

There are a wide variety of applications available in mobiles ranging from human health care, military, environment, etc. The mobile applications are complex in nature, and need to process big data in order to come up with results. Under these cir-

cumstances, it is very necessary that we need to ensure that we are fully aware about energy related issues in case of mobile big data processing. Also, we need to ensure that we use different techniques and approaches in order to resolve this energy problem. The reason that these energy issues arise in case of mobile devices is because of the fact that the mobile is a small device and the battery is relatively small and hence cannot store huge energy. Subsequently, battery life is less and we need to ensure that the energy is spent efficiently. There are several Big Data techniques has been proposed which may works with MCC environment which are discussed as follows:

1. *Big Data Stream Mobile Computing (BDSMC)*: The term "Big data stream mobile computing" comprises of two concepts which are merged together. They comprise of the culmination of broadband based stream which provides a constant stream of Big data and the other one is mobile cloud computing which helps in processing this big data at real time [15]. Also, offloading is very important out here and plays an important role to process the big data which is collected by the mobile devices. Figure 7 shows the BDSMC architecture. There are three layers namely, the Radio Access Networks (RANs) or the User layer which is the source for big data, the Internet backbone layer and the remote networked data center layer. Offloading of data and compute intensive applications to the cloud takes place using the best possible manner wherein least energy is consumed and using the appropriate access technologies (e.g., either of WiFi, 3G/4G)
2. *StreamCloud Architecture*: StreamCloud Manager (SeCoM) design was inspired by the workload offered by real time big data streams and also that setup costs and energy wastage due to frequent ON/OFF transitions [15]. SeCoM supports

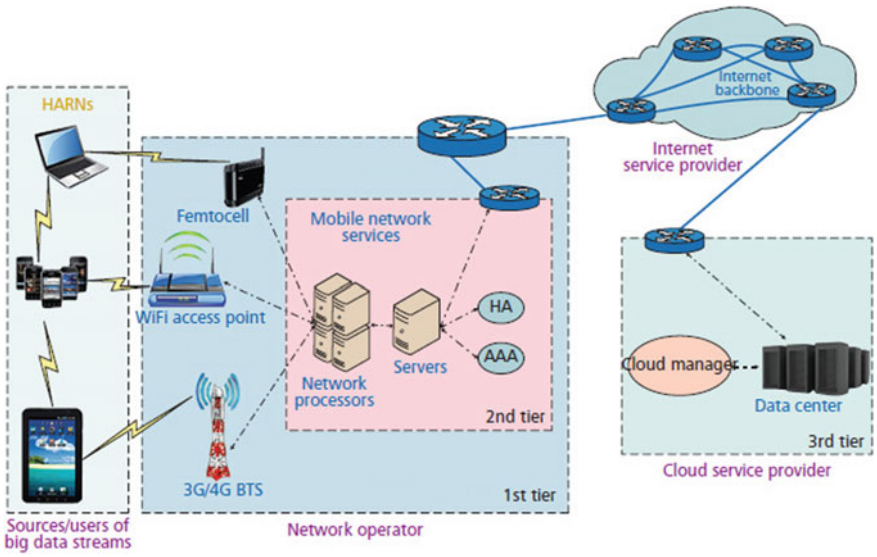


Fig. 7 BDSMC reference architecture [15]

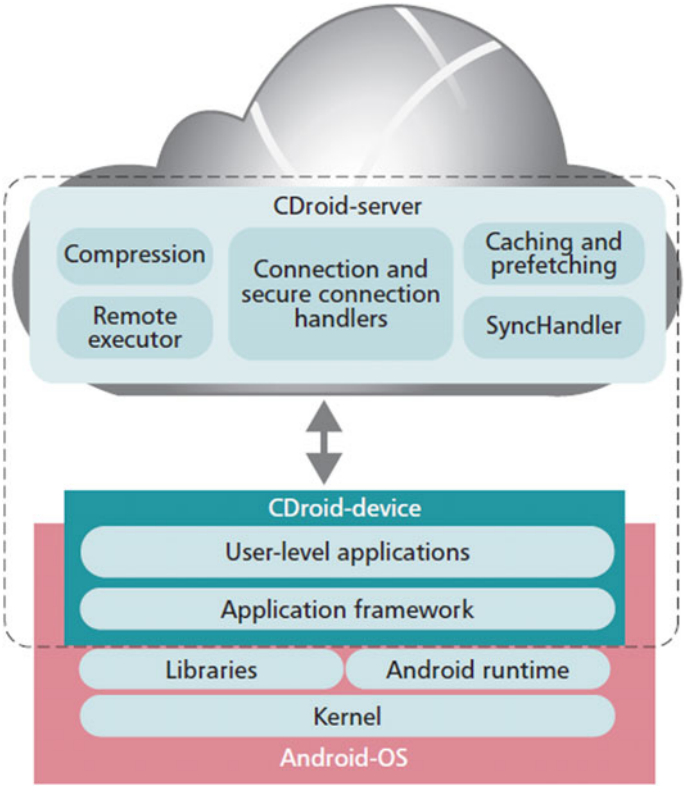


Fig. 8 VNetDC architecture based on StreamCloud platform

HIS (Hibernation stage) which helps to save energy. The main features of SeCoM module are the self-reconfiguration of the overall virtualized networked data centers (VNetDC) which helps to ensure minimum energy resource configuration. In the architecture, the following components are present. CDroid Server, which is responsible for doing the offloading of computation from mobile devices to the cloud. It has two sides, the device side and the cloud side. CDroid architecture has been shown in Fig. 8 and the components of the CDroid which help in energy saving offloading are:

(i) Communication handling module, responsible for management of all data traffic over mobile Internet connection and also performs http traffic tunnelling of the mobile device via the CDroid server
(ii) Caching and prefetching module.

Figure 9 shows the VNetDC architecture which is based on the StreamCloud platform which working as an ad-hoc-designed stochastic gradient algorithm and

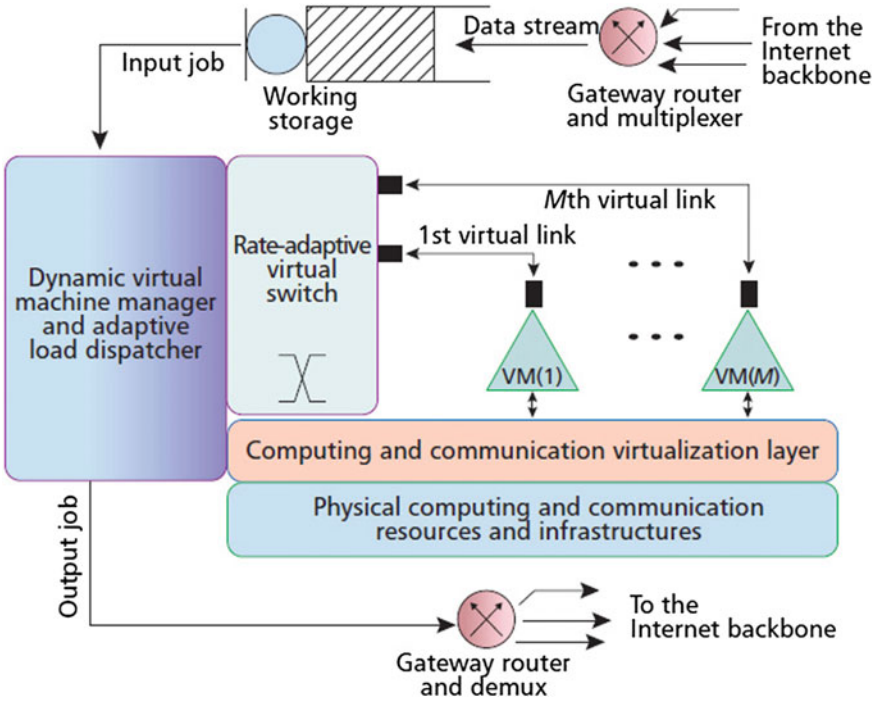


Fig. 9 BDSMC reference architecture

ensure that even if there are unpredictable dynamic fluctuation of the workload, it involves with the minimum-energy resource configuration.

3. *Mobile Cloud Sensing applications:* There are various type of applications which use MCC, Sensing capabilities together with Big data using 5G as the communication medium, which make life more meaningful and worthy [11]. These applications are spread over various areas like Health monitoring, environment monitoring and others. The taxonomy of the mobile sensing applications depends on the size of the applications which determines the volume of data, size of the application, execution speed and other parameters. We take a typical parameter of the size of the applications in mobile sensing area and it can be divided into the below sub-areas [11].

- Personal or Individual sensing, wherein the data is of personal users and used for their betterment, for example, the medical information of a person.
- Collective or Group sensing, and here the data belongs to the group and is aimed at their usage, like restaurant ranking and others.
- Community sensing, where the entire community data is collected and used to predict matters like global trends of warming and climate and other popular trends.

The typical architecture of mobile cloud sensing has been shown in Fig. 10.

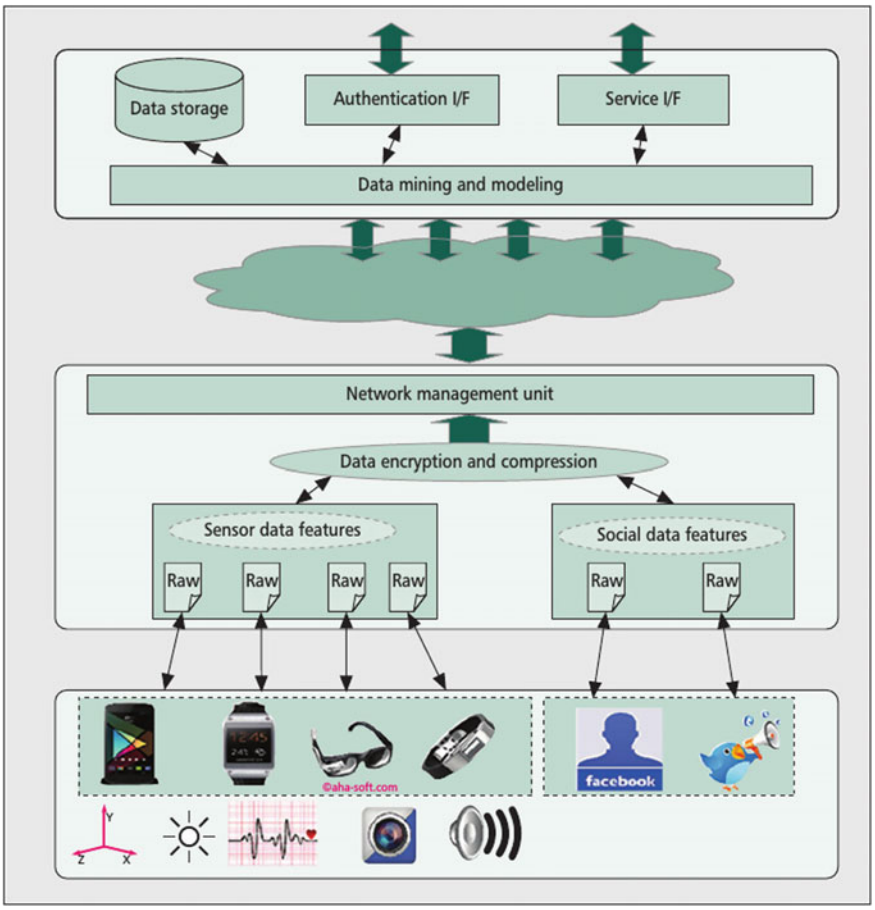


Fig. 10 Architecture of mobile cloud sensing

The different architecture components perform the functionality as explained.

- Data Sensing Unit
- Data Preprocessing Unit
- Network Management Unit
- Cloud Data Mining and Storage

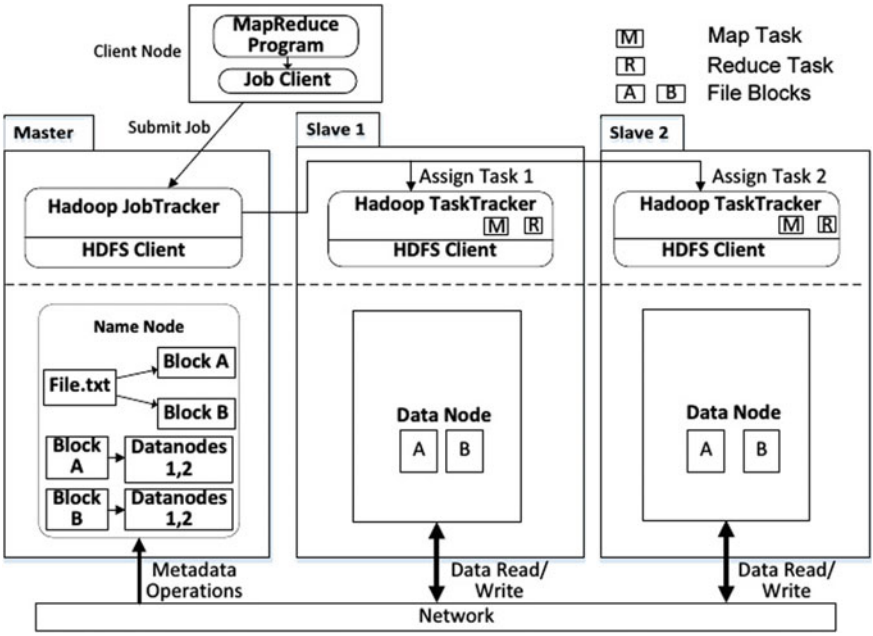


Fig. 11 Hadoop architecture [6]

However, there are also various limitations of Mobile Cloud Sensing, which are listed as follows.

- Currently, it is only in experimental stage and needs to be implemented in large scale
- The limitations of network resources
- Big Data usage

There are few works have been proposed which are based on the Hadoop platform. Figure 11 shows the Hodoop architecture and the various components of the Hadoop architecture are as follows [6]:

(i) Apache Hadoop system comprises of two basic components, mainly the Map Reduce (MR) framework and the Hadoop Distributed File System (HDFS). MR comprises of the Map and Reduce Task which is done in a parallel manner and runs on top of the HDFS. In the Map Task, the input data set is taken, and intermediate $< Key, Value >$ pair is produced and is then sorted and partitioned per reducer. In the Reduce Task, the reducer identifies the aggregate function to be applied to the output of the map task and then runs it. The number of times a reducer needs to be run will be set by the user. This mapped output is then sent across to the reducer for deriving the final output. Below is the general form of the Map and Reduce Task:

map: (K1, V1) → list(K2, V2)
reduce: (K2, list(V2)) → list(K3, V3)

In the MR framework, the computation is moved closer to the data node, which will minimize the congestion in the network and also maximize the throughput. Two important modules in MR are the JobTracker which accepts user jobs and does the splitting into multiple taks and the TaskTracker which are the nodes that execute the tasks in the cluster that run the tasks.

(ii) HDFS is a distributed file system which is very reliable and also fault tolerant. It stores huge datasets in petabytes and beyond and is load balanced to achieve efficiency.Each file is split into blocks where-in blocks are replicated across several devices in the cluster. It is designed to run on commodity hardware. It provides high throughput access to application data and is suitable for applications that have large data sets. The two modules in HDFS are NameNode which holds the metadata information about the different files storing the Inode records of files and directories, and the DataNode which is actually storing the file blocks and helps to complete the read/write requests coming from any client.

The Mobile Distributed File System (MDFS) helps to take care of the big data processing in mobile clouds. It is built on a k-out-of-n-framework which ensures reliability and security of data. Every file uses a secret key to encrypt and is fragmented into several file fragments using Reed Solomon algorithm. Also the secret key is split into fragments using Shamir's secret key sharing algorithm. MDFS helps to ensure high security as data cannot be decrypted until an unless the authorized user does not obtain all the different key fragments. Figure 12 shows the architecture of MDFS. MDFS incorporates a directory service which is distributed and synchronization of each node happens on a periodic basis which ensures that all device directories are updated at all times.

Figure 13 shows the distributed MDFS architecture. In case of distributed MDFS architecture, there is no central controlling authority to manage the cluster. Here, every node has a Name Server and Fragment Mapper. Also, every node will sync up with other nodes in a regular interval. In case of any new node, the broadcast messages are received for all new nodes. The best part is that in this architecture

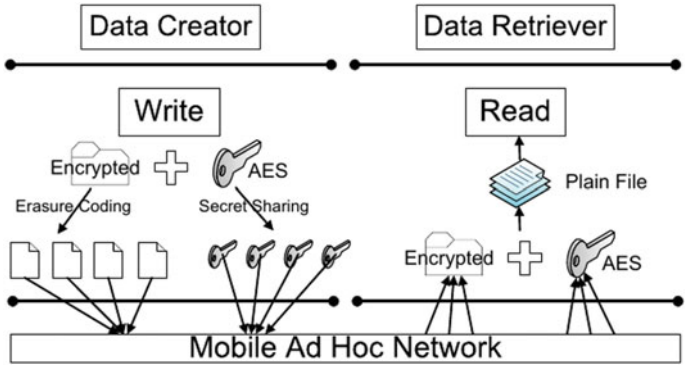


Fig. 12 MDFS architecture

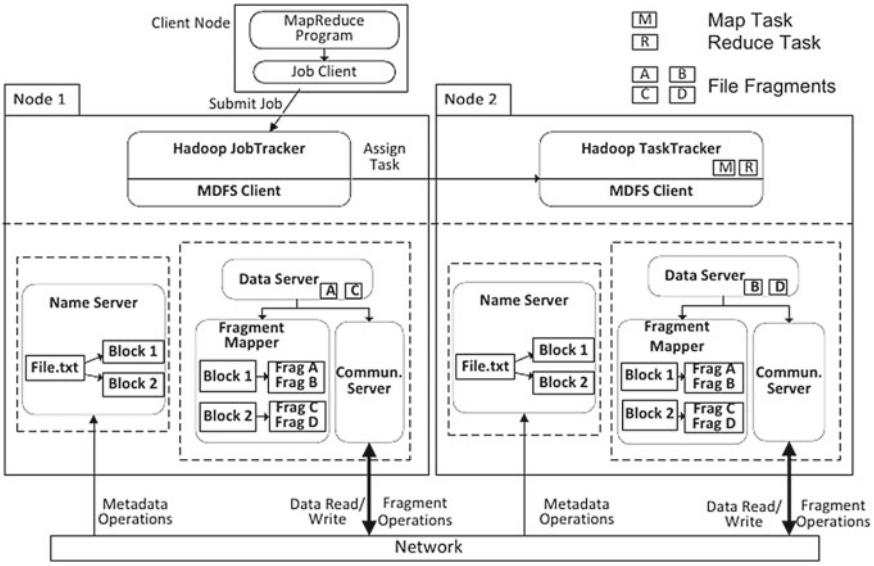


Fig. 13 Architecture of distributed MDFS

there is no single point of failure as each node can operate in an independent manner (since it stores a copy of the namespace and fragment mapping). This ensures that the load is evenly distributed across the cluster [6].

Figure 14 shows the centralized MDFS architecture. In case of centralized MDFS architecture, name server and Fragment Mapper are a single instance across the entire cluster. As daemons can be run in any node, the master node is the node which runs the daemon. It has an advantage as there is no device memory wastage because metadata is not stored across multiple nodes and also upon creation or change of file, it is not required to broadcast it to other nodes [6].

Figure 15 shows the data flow diagram for the read operation. The overall transmission cost during read operation will vary across nodes based on the different location of fragment and that of the reader. The different steps are summarized as follows [6].

1. User issues read request for file blocks of definite length have a certain byte offset.
2. MDFS client seeks information from Name Server to return all blocks of the file.
3. The data retrieval request from the data server is issues for each block.
4. Data Server returns the block.
5. Data server will then request Fragment mapper to furnish information regarding the key and file fragments, which is then returned.
6. Data Server will then request the communication server to fetch the different fragments from locations wich are previously returned by Fragment Mapper.

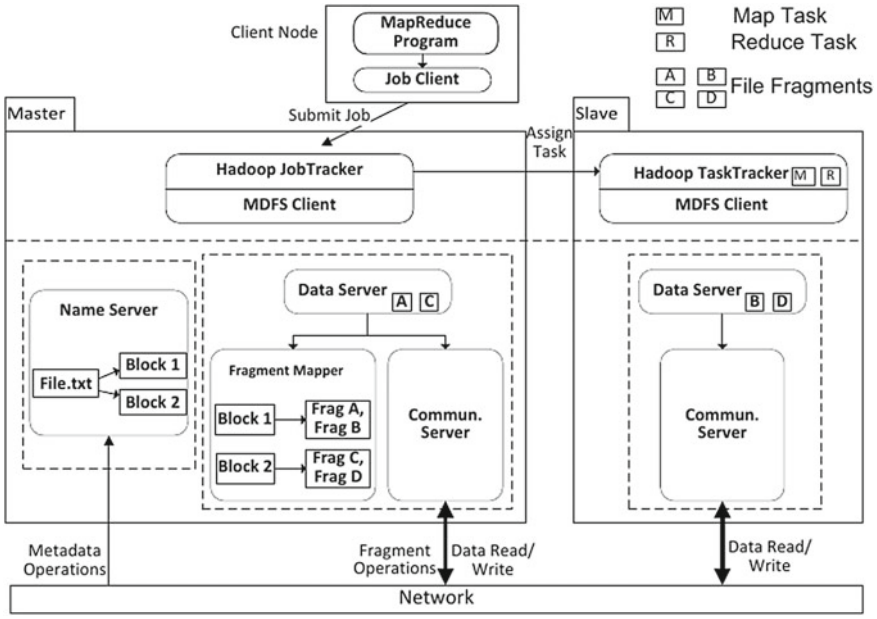


Fig. 14 Architecture of centralized MDFS

The fragments are fetched and stored in local file system of the client who has requested.

7. All the above operations are repeated for fetching the key fragments. Secret key will b created from key fragments.
8. Upon downloading completion of the file fragments, they are decoded and then decrypted to get original block.
9. Key and file fragments are deleted for security purposes.
10. Data Server will then acknowledge the client with the location of the block in the local file system.
11. MDFS client reads the requested number of bytes of the block and Steps 4 to 19 are repeated in case there are many other blocks to read. Once read operation is completed, the block is
12. deleted and original cluster state is restored.

Figure 16 shows the data flow diagram for a write operation. The HDFS write is not possible for MDFS unless the block is decrypted and decoded. The different steps are summarized as follows [6].

1. User issues a write to file request.
2. MDFS client will then request the Name Server for a new block Id.
3. Based on allocation algorithm, the Name Server returns a new block id.
4. MDFS client then issues a creation request to Data Server which then helps in block creation.

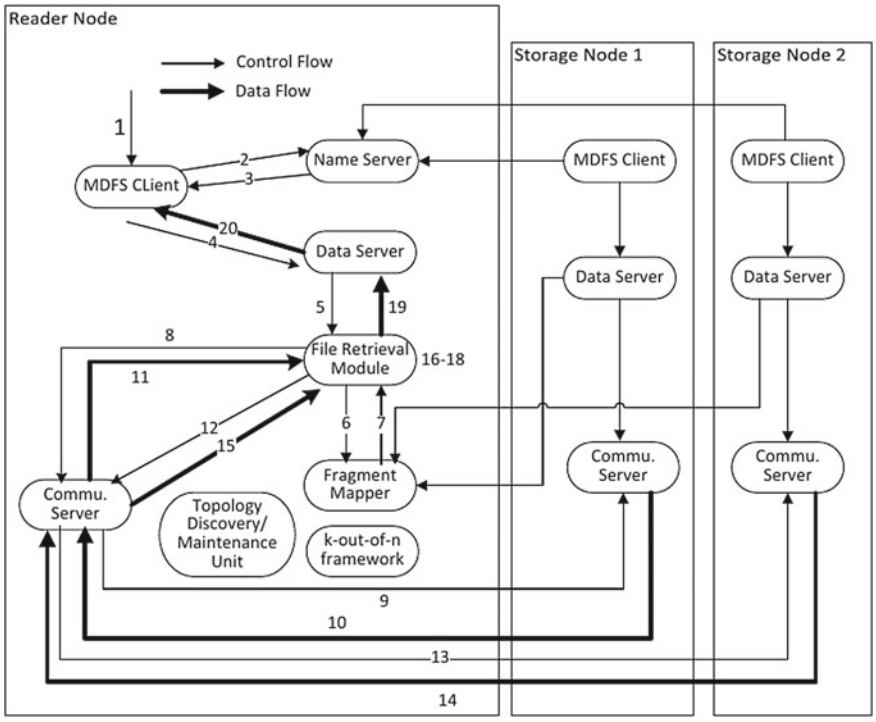


Fig. 15 Data flow diagram for read operation

5. Encryption using secret key takes place for the block in the local file system.
6. Using Shamir's secret key sharing algorithm, key is split.
7. K-out-of-n framework provides storage nodes.
8. Data Server makes a request to the Fragment mapper to add fragment information.
9. After file and key fragments are distributed in the cluster, Data Server will inform client about the file written successful status. The local file system of the writer is delete after write operation for security purposes.

3 Future Research Direction

There are several areas of research that enthusiastic researchers should delve into in the area of Mobile cloud sensing to make our planet *smart* and *intelligent*. The open research issues are discussed as follows, which will serve as a direction to the future researchers.

- Mobile cloud sensing has immense potential of research. If we take into account the scarce resources present in the mobile devices, there arises a need to be able

to optimize their usage in an efficient manner. This is a very intelligent area and lot of research is currently being pursued.

- Handling large amounts of data which is generated by the mobile sensing data is an area of research and a tough problem to solve.
- Also, 5G as a Infrastructure for Mobile Cloud Sensing is being used in research projects as it supports a peak download speed of 3.6 Gbps as compared to 100 Mbps in 4G. Also, we need to enhance capacity of data movement between the cloud and the mobile device which is expected to rise to 16 Exabytes by 2018 (ten times of the current data volume.
- BDSMC emphasizes on the real time offloading of code and/or data to the cloud through the mobile and Internet network and also real time configuration of the cloud. This area can be dealt with in detail as a research topic.
- As there is a need to provide real-time support of computation in BDSMC applications, the management systems are specially designed for the same. However, BDSMC does not provide a utomatic and dynamic adaptation to fluctuations of time of the input streams processing, and this problem needs to be researched.

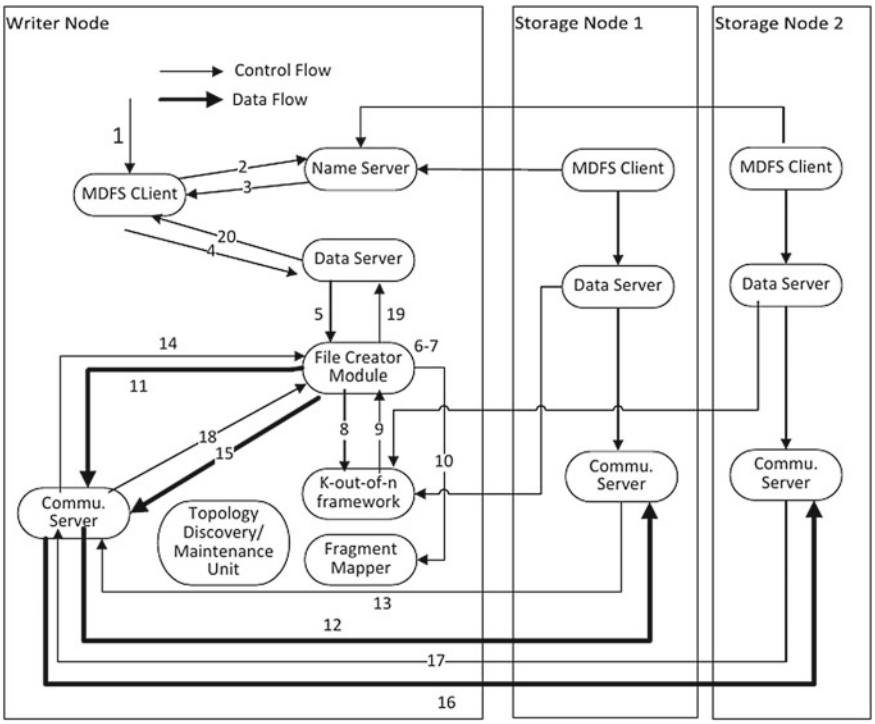


Fig. 16 BDData flow diagram for write operation

4 Conclusion

In this chapter, we have dealt with various techniques related to energy management of mobile devices using MCC like Phone2Cloud, eTime, DREAM and ThinkAir. Also, energy management techniques of MCC using big data have been discussed like BDSMC, StreamCloud and Mobile Cloud Sensing. Besides this, various flavors of MDFS architecture have been discussed as well. There are various limitations discussed here about the various architectures. These present architectures will find it difficult to handle huge volumes of data which are sent to the system at real time. Also, 5G is becoming popular and is necessary for faster data movement during offloading process. The research direction is to delve into the area of Mobile cloud sensing for handling large amount of data and to use 5G as Infrastructure.

References

1. http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/. Accessed 30 Dec 2016. Accessed 30 Dec 2016
2. http://www.bbc.com/news/business-26383058. Accessed 30 Dec 2016. Accessed 30 Dec 2016
3. Panigrahi, C.R., Sarkar, J.L., Pati, B., Das, H.: S2S: a novel approach for source to sink node communication in wireless sensor networks. In: Proceedings of 3rd International Conference on Mining Intelligence and Knowledge Exploration, pp. 406–414 (2015)
4. Kim, Y., Atchley, S., Valle, G.R., Lee, S., Shipman, G.M.: Optimizing end-to-end big data transfers over terabits network infrastructure. IEEE Trans. Parallel Distrib. Syst. **28**(1), 188–201 (2017)
5. Panigrahi, C.R., Pati, B., Tiwary, M., Sarkar, J.L.: EEOA: improving energy efficiency of mobile cloudlets using efficient Offloading Approach. In: Proceedings of IEEE International Conference on Advanced Networks and Telecommunications Systems, pp. 1–6 (2015)
6. George, J., Chen, C.-A., Stoleru, R., Xie, G.G.: Hadoop MapReduce for mobile clouds. IEEE Trans. Cloud Comput. **3**(1), 1–14 (2014)
7. Bowen, Z., Dastjerdi, A.V., Calheiros, R.N., Srirama, S.N., Buyya, R.: A context sensitive offloading scheme for mobile cloud computing service. In: Proceedings of the IEEE 8th International Conference on Cloud Computing, pp. 869–876 (2015)
8. Essa, Y.M., Attiya, G., El-Sayed, A.: Mobile agent based new framework for improving big data analysis. In: Proceedings of International Conference on Cloud Computing and Big Data, pp. 381–386 (2014)
9. Rong, P., Pedram, M.: Extending the lifetime of a network of battery powered mobile devices by remote processing: a Markovian decision based approach. In: Proceedings of 2003 Annual Design Automation Conference, pp. 906–911 (2013)
10. https://dupress.deloitte.com/dup-us-en/focus/tech-trends/2015/tech-trends-2015-what-is-api-economy.html. Accessed 31 Dec 2016. Accessed 31 Dec 2016
11. Han, Q., Liang, S., Zhang, H.: Mobile cloud sensing, big data, and 5G networks make an intelligent and smart world. IEEE Network **29**(2), 40–45 (2015)
12. Kosta, S., Aucinas, A., Hui, P., Mortier, R., Zhang, X.: Thinkair: dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In: Proceedings of 31st IEEE International Conference on Computer Communications, pp. 945–95 (2012)
13. Shu, P., Liu, F., Jin, H., Chen, M., Wen, F., Qu, y., and Li, b.: ETime: Energy-efficient transmission between cloud and mobile devices. IEEE Infocom, pp. 195–199 (2013)
14. Tawalbeh, L.A., Mehmood, R., Benkhlifa, E., Song, H.: Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access **4**, 6171–6180 (2016)

15. Baccarelli, E., Cordeschi, N., Mei, A., Panella, M., Shojafar, M., Stefa, J.: Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study. IEEE Netw. **30**(2), 54–61 (2016)
16. https://en.wikipedia.org/wiki/Mobile_cloud_computing. Accessed 31 Dec 2016 . Accessed 31 Dec 2016
17. Xia, F., DingJie, F., Xi, J., Kong, X., Yang, L.T., Ma, J.: Phone2Cloud: exploiting computation offloading for energy saving on smartphones in mobile cloud computing. Inf. Syst. Front. **16**(1), 95–111 (2014)
18. Kwak, J., Kim, Y., Lee, J., Chong, S.: DREAM: dynamic resource and task allocation for energy minimization in mobile cloud systems. IEEE J. Sel. Areas Commun. **33**(12), 2510–2523 (2015)
19. https://en.wikipedia.org/wiki/Big_data. Accessed 31 Dec 2016. Accessed 31 Dec 2016
20. http://www.dataintensity.com/characteristics-of-big-data-part-one/. Accessed 31 Dec 2016. Accessed 31 Dec 2016

Part IV
Industrial Practices of Mobile Big Data-Driven Models

Big Data—A New Technology Trend and Factors Affecting the Implementation of Big Data in Australian Industries

Bhavyadipsinh Jadeja and Tomayess Issa

Abstract Big data is new technology trend and it provides immense advantages. There are too many social networking websites people are using, these websites more than ever before. The data which has been created in the last 5 years is greater than the total size of data accumulated during the preceding. This indicates that people are producing big data knowingly or un-knowingly. In addition to that, every company receives an enormous amount of data in many ways. This data can be transformed into information and information can be converted into knowledge. This knowledge can be very helpful in product marketing. Australian industries can make use of big data. The main objective of this research is to provide more information about big data which is still in its infancy, and find the factors which may affect the Implementation of big data in Australian industries. Big data is still a relatively new field, especially in Australia. It is understandable that big data will become a significant player in Australian industries and that is why it is desirable for these industries to use big data. Big data holds the key to business intelligence and that is why it is important to undertake research on this specific topic. This research study has used a positivist approach and a combination of methodologies. Qualitative and quantitative methodologies have been used and the survey was used as the instrument for gathering data. The survey consists of five sections. The first section is intended to collect the participants' demographic data. The other four sections are each based on one of the four factors associated with big data: cost, technology, skills and maintenance. The survey has been designed to address these four factors. The Likert scale has been used in this research method. The research design has been explained in detail. The objectives of this research study are: (1) to find factors which determine the Implementation of big data in Australian industries, and (2) to understand the general trend of big data and the technology which can be used to analyze big data. This research provides very important information regarding how big data can be used by different organizations. The participants are

B. Jadeja · T. Issa (✉)
Curtin University, Perth, Western Australia
e-mail: Tomayess.Issa@cbs.curtin.edu.au

B. Jadeja
e-mail: bhavya029@gmail.com

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_11

employees in the IT department of various industries comprising retail, IT, education, oil-gas and healthcare. These participants' positions range from entry level to managerial level. The participants' responses, which constitute the data, were then entered in IBM's SPSS version 24. The data was entered and analyzed using the factor analysis method. In general, it was found that several factors can affect the Implementation of big data. These factors are: cost, maintenance, skills and technology. The analysis has indicated that statistical skills, IT skills, project management skills and communication skills are important for the people who work with big data. In addition, hardware and software cost, hardware and software maintenance, hardware and software technology also affect the Implementation of big data.

Keywords Big data · Factors · Implementation of big data
Australian industries

1 Introduction

This research topic concerns the recent technology trend which is big data. Big data is the newest technology norm in the tech industry. Big data is simply data that is very huge in terms of volume, velocity and variety. People are now using social networking sites more than ever before. According to a recent study, the data which has been created in the last 5 years is more than all the combined data of previous years. Moreover, it is assumed that in coming years this data will continue to grow and become exceedingly huge. The reader may well ask why it is necessary and useful to address this issue. It is useful because normal data possesses the information and this information can be converted to knowledge. Most companies are facing this situation. They have knowledge in raw data but they cannot use it because of the lack of availability of appropriate technology. That is why it is important to have appropriate technology which can work with big data. The main aim of this research study is to find out how big data can be implemented in Australian industries and the factors associated with big data that prevent Australian industries from using it. The researcher selected five different industries for this research: healthcare, IT, retail, education and oil-gas. Companies in these sectors can use big data to solve many internal problems. Moreover, technology is constantly evolving and it is important to implement big data to gain a competitive edge in the marketplace.

Big data is still new in the market and that is why it is very difficult to find appropriate technology which can work properly with companies' operations. It is difficult because big data cannot be stored in traditional data storages. As explained above, big data is very huge in terms of volume. So, it is not important to store such an amount of data in normal database storages. In addition to that, this data should be available in real time during the analysis process. So, it is important to find the storage facilities which can provide such advantages. The second issue with big

data concerns hardware. Legacy systems cannot perform high-end operations which big data require. A Simple computer cannot perform thousands of queries in real time and that is why a machine with high processing ability is required. Therefore, hardware is an important issue in big data. The third issue concerns the selection of appropriate software which can perform various operations on big data. Big data can provide business intelligence only when it is associated with proper software technology. There are many vendors out there which provide exceptionally good software packages. Hadoop from IBM is an example of such software. The last and most important issue is that at the end of the day, people work on big data. It is very difficult to find people with big data skills because they are scarce. It is the responsibility of organizations to create this talent. However, some of the technology companies have started to use training programs which can equip people with the necessary big data skills. Moreover, big data is more complex than regular data. Companies from the industry sectors mentioned previously can really benefit more from big data. It emerged in 2002 and the term was coined at that time. It has evolved very quickly in the last couple of years and has started to show its advantages. Most of the big organizations are now willing to benefit from this technology. If appropriate software is deployed in the company, it can provide quick analysis of large sets of data, enabling business managers to make quick decisions based on the reports rather than intuition. The next section outlines the main components of this research study.

2 Big Data

Big data is a widely-used technology term which has emerged relatively recently. People use social networking sites and mobile devices more than ever before, as a result of which enormous amounts of data are being generated constantly. These data possess very important information. Using appropriate technology, this information can be analyzed and knowledge can be derived. In simple words, big data includes the tools, processes and procedures that can create, delete, edit and use large amounts of data with strong storage facilities [17]. Enormous amounts of data are generated in each and every field. Big data has many advantages but needs numerous resources to accommodate it. Scientists are working on big data and on finding ways to reduce it so that it can be used to greater advantage. Mathematics is related to big data and there are unanswered questions such as algebraic, geometric and topological structures underpinning the mathematics of big data [22]. Hence, big data can be very advantageous in understanding such questions and providing solutions for the same.

However, some would argue about what big data really is. In simple words, big data can be combined as three Vs. The three Vs are volume, variety and velocity. These three can be explained as follows. Volume refers to the large amount of data which is being generated at a constant rate in different industries. So, it simply is large amounts of data. The second ‘V’, variety, means that the data can be of

different types such as documents, images, databases. The third 'V' is velocity. As stated previously, the amount of data being generated is very high. For example, the number of images uploaded on Facebook last year is more than the total images in previous years [23]. Therefore, it is important to consider this factor thoroughly while working on big data implementation. Apart from that, it is also important to see the speed of retrieval of data as well from these large data sets. However, big data should not be confused with 'lots of data'. They are different and have different meanings.

Big data is different from small data in many ways. The goals of big data are different from those of small data. The goals of small data are to just answer specific questions; whereas in the case of big data, the question paradigm changes and is flexible. Similarly, the location of big data and small data are also different. Small data is covered only for one institution while big data are covered between multiple servers and can be located anywhere in the world. Another obvious thing is that data contents and structure is also different. In the case of small data, the user prepares his/her own data whereas in big data, the data can come from multiple resources. Hence, big data is different from small data in its preparation. The longevity of small data and big data are also different. Small data are retained until the end of a particular project; on the other hand, big data are stored until the same amount of information is not available. Measurements of small data are done using one protocol only; however, big data can be measured using different protocols. Reproducibility is an important feature of data and the reproduction of small data is always repeatable while in the case of big data, it is seldom feasible. Another important difference between big data and small data is cost and stakes. For small data, the cost of a project is very limited and usually very low. However, big data can be very expensive. CEOs need to provide ample resources to implement big data and to research laboratories. Moreover, if a project fails, a half-developed solution is useless. However, big data still has many advantages. Introspection of small and big data is different as well. For small data, it can be done easily from the row and column address. However, in big data, it is not possible even if it is designed very well. Various techniques are used in big data which are called altogether as introspection. The most important difference between big data and small data is the analysis of such. In the case of small data, analysis is done together as this analysis consists of whole data while in the case of big data, its analyzed incrementally using different techniques [23]. Big data has many advantages. Everyday tens if not hundreds of thousands of people generate big data as they are using social networking websites more than ever. This data can be analyzed to produce information and knowledge which can be helpful to predict future business trends. The company does not need to do anything new to introduce big data into its operations but they need to just follow the same routine but with big data and by using appropriate techniques. For example, there are thousands of activities which are done in a particular bank. This bank then stores and analyzes this data in order to develop better methods to attract new customers and provide better service to its existing customers. In most cases, generated data are unstructured and big data technology is used to give structure to these unstructured data. The data come in

different sizes and different formats. If any company wants to take advantage of big data then it is important to provide a strong structure to the data, so that the company can generate sound knowledge and information from it. However, this is not as simple or straightforward as it seems.

There are some challenges associated with big data. Big data analytics goes through many different stages and each stage has its own challenges. These stages are explained below. The first stage is that of loading and ingestion. In this stage, data analysts add data in the database server by using various resources such as ERP and CRM software. In addition, this data can be in the format of documents and images. Extract, transform and load are the main three tasks which need to be done in this stage which are of course very difficult to do with this large amount of data. The second stage in big data management is manipulation and transformation. As explained previously, when the data is entered in the company's server, it is generally much unstructured but to obtain the most benefit from this data, it must be structured. Hence, manipulation and transformation of this data is required. Given this large set of data, the same protocol is time consuming; however, the small code can make this step very easy. However, employees who work on this step need the appropriate skills enabling them to work with such data. Finding employees with these skills is another challenge for the company.

The third stage of big data is to enable access to this structured data. After converting unstructured data to structured data, it is important for the company that tools like Hadoop and others provide access to this data so that applications of these tools can use big data and can generate useful information. This step is relatively easy and requires very basic edit, transform and load (ETL) skills. The challenge with this stage is to provide access to appropriate tools. The fourth stage in big data management is to create a model for the business applications. Business analytics is depended on big data and it cannot work without an appropriate model and modeling techniques. This stage is important as users actually pose queries here to find relevant answers. Also, users need an appropriate model so that they can fine-tune their search. This stage can be done easily by using simple coding. The challenge with this stage is to create a suitable model which covers all the areas of the business. It is also important to ensure that data is easily available from different units in order to create the model.

The fifth and final stage involves visualization and analysis. Visualization can be done using business intelligence tools or by writing code. There are tools available like dashboards and reports. This can help business managers to make informed decisions very quickly. Analysis plays a part in this stage as well. Analysis is the most important part of this stage because this is what generates the relevant information and converts this information to useful knowledge. The challenge with this stage is that small mistakes can generate entirely different results. In addition, visualization is not easy for such a big amount of data. So, it is important for the analyst to use suitable techniques and tools to visualize such a big amount of data [13]. By following the steps in these different stages, big data can be managed and the most can be made of the advantages it has to offer.

One may wonder why and how big data has evolved. The authors explain the reasons for this evolution and the factors which have affected it. The internet has played a major part in the evolution of big data. People are using the internet and mobile devices as never before. Moreover, social networking sites like Facebook and Twitter have also played a significant part in this movement. Another factor which has affected the evolution of big data is computation. In early days, computation power of mainframe computer was very low and it was not possible to work out something so big. But, the scenario has now changed. There are super computers which can respond to thousands of queries with a single click. This has provided a way to solve certain big problems and made way for big data evolution. Another factor is networking ability. In the early days, it was not possible to network a large area and local area network abilities were limited. But now, with new technology, it is possible to connect more machines to each other. In addition, the servers can be located anywhere in the world and one can execute queries on the other side of the globe. This has provided flexibility in carrying out big data tasks. Main activities such as edit, transform and load can be done easily because of high functioning networking technologies. Moreover, storage facilities have also evolved in the last decade. Previously, storage disks were room-size. Now, there are storage facilities the size of a human finger that can store more data. Also, these storage facilities can be located anywhere in the world. All of these factors have helped to produce the big data revolution. The next section discusses the sources of big data.

Authors previously explained how data are different how they are being generated. In this section, the authors explain the actual sources of big data. There are different types of sources which play an important role in the generation of big data. There are mainly three parts of data which contribute in generating data. These types can be defined as directed data, automated data and directed data. Directed data is a type of data which is generated in offices by employees who are usually directed by managers. This is also the type of data which is generated in laboratories. In addition, this type of data is stored automatically and in a suitable format. This type of data can be managed easily [21]. The second type of data is automated data of which there are various sub-types. This type of data is generated by automated devices. The first sub-type is automated surveillance. Each and every city is covered now by surveillance cameras. These cameras generate very large amounts of data which is automated. The second sub-type is generated by digital devices. The majority of the population now has access to smart devices and they carry their devices wherever they go. This also generates very large amounts of data. The third sub-type is sensed data. This data is generated by sensors and actuators. Most of the oil & gas companies use such devices. This helps to predict future demands and trends. Hence, these devices are called ‘contributors’ to sensed data. The fourth sub-type is scanned data. Similarly to those of sensed data, the data are being generated through scanners. This is called ‘scan data’. Most of this data are generated in the retail industry. The fifth sub-type is interaction data. Nowadays, most of companies in any industry use information technology on a large scale. Each and every interaction conducted within these companies contributes to the generation of

large sets of data. These are all types of automated data which constitute the largest part of total generation.

The third type of data is volunteered data which is entirely different from directed and automated data. The latter use certain devices to capture ultimately what is known as interaction which is not the case in volunteered data. Similarly to automated data, there are also several sub-types in directed data. The first sub-type is transactions. E-commerce has become one of the biggest businesses in industry right now. Every time the customer buys something, he or she enters personal information which contributes to generating large amounts of data. The second sub-type is social media. People upload images, documents and their communication messages with their peers, and this also generates large amounts of data. Therefore, social media also plays a very big role in the generation of volunteered data. The third sub-type in volunteered data is surveillance. This is different from surveillance because here a person or entity does surveillance of oneself. There may be different reasons for this such as health monitoring. The fourth sub-type of volunteered data is crowd sourcing and citizen science. In crowd sourcing, different people present their ideas and knowledge and share this online with others.

This also generates very large amounts of data. These are the three main types of data under the big data umbrella. If a company wants to use big data, then a strategy must be followed. Before using big data, there are various things which need to be followed as well. It can be categorized in eight different types. The first and foremost thing is big data requires a different culture than that of legacy culture. It is important that an organization have a center of all information and it is available easily for the people in IT to store and analyze. Other important thing for an organization is that it has skilled people and they have the ability to work on very large sets of data. Moreover, an organization needs to understand that data is every-where and it cannot be unseen. Organizations should follow the rule that everything which is in digital format is data. This data is the source of information. In addition to that, big data engineers are hard to find and this is the reason that an organization should find appropriate talent before a competitor does.

Moreover, big data means much information and much information means much knowledge. In the digital age, it is important that knowledge be kept safe. So, big data requires adequate security measures. An exit plan should also be in place. Also, big data raises privacy issues. Organizations will also need to ensure that there will be issues regarding privacy and they should be ready to address such issues. Moreover, apart from competitors, organizations should also keep in mind that governments all around the world are also making efforts regarding big data. The last thing that organizations should keep in mind is that big data is not just about big data or volume but it is about volume, velocity and variety [15]. Big data is the main source of business intelligence from which business managers can make informed decisions which are fast and reliable rather than just intuitive. Business analytics is something which converts raw material into useful information. Business analytics is done differently in big data compared to other analytics. Techniques and methods which are used in this theory are also different from the techniques used in simple analysis. There are special tools for big data analytics

which is explained later in this chapter. For big data analytics, the most important part is to build the perfect model for the individual business. Prediction depends on several variables. It is important for the analyst, who is working on big data analytics, that he or she identifies the critical variables associated with the stakeholder and decides which variables are independent and which variables are independent. Finding such variables is a difficult task but the most important one in big data analytics. Big data analytics can be done using powerful tools. In addition to that, business analytics provides the business intelligence. Business intelligence can be generated using appropriate tools and dashboards. Dashboards are useful for visualization. Visualization is as important as analytics. In addition to that, it is also important in order for business analytics to be successful that all the information from organization is available to the analyst at any time so that they can manipulate them and can give proper structure. Moreover, the storage should be done in such a manner that it is always available for business analysis and analysts have the appropriate skills to work on big data business analysis [31].

Australia is one of the leading technology-driven countries in the world and technology-driven innovation is an integral part of Australia's economy as most industries (such as retail, supply chain, accounting etc.) depend highly on cutting edge technology. Big data can play a very important role in Australia. This part presents the authors view on this topic; also, other authors' views have been critically reviewed here. Big data provides an advantage in predicting actual business demands. For example, Kollaras Group is one of the largest liquor manufacturers in Australia. The business of KollarasGroup was expanding and they needed something in IT to help it. Big data was introduced in the company and it helped. "Management is now able to rapidly analyze the business and respond to emerging trends using real time data, while eliminating significant manual processes in a way that was not possible before" [22]. Moreover, although analysis of large sets of data is very difficult, it provides more information ultimately and thus helps management to make informed decisions. Big data is certainly the next big thing in the IT industry and in every industry for that matter which uses IT. Oil & gas industries in Australia have started to benefit from this technology which provides extremely good business intelligence.

Big data can be regarded as an innovative approach to data management. Data management has been done before but in the case of big data, it is the management of large sets of data and that is the reason why it is difficult for a company to immediately integrate big data into its operations. However, in this competitive environment, a company cannot avoid big data when competitors are deploying it. The problem with big data is that if it cannot be applied properly in the business, then it will provide little to no assistance for decision-making in business. Also, it can lead to wrong decision-making. Another problem is that data which is collected for big data analysis is very complex and, for the software to be successful, it is important that the data which has been collected is reliable and in the correct form. Moreover, it is also important that this data has scope. So, it is obvious that when this large amount of data is gathered, there is the need for a large amount of space and good processing power. Importantly, this should not be limited to the private

sector only, but the public sector should also benefit from this. Fortunately, the Australian government has realized the potential of big data and the Australian Office of Information Management had released a service strategy.

The government is realizing that the information which is provided by big data can be helpful for the betterment of the community. According to the reports of the Gartner study [28], 64% of businesses will invest in big data in 2013. The main reason for most of the companies doing so was customer satisfaction and that is because the full potential of big data is yet to be fully understood. Yes, it is hard to manage and can be very damaging if one cannot understand it properly. Many companies in Australia are now adopting model which includes big data in it. However, schools and universities in Australia can play a major role in this transformation. These schools need to provide appropriate and regulated knowledge on big data. Australia's National School of Data has begun to do this. This was not possible 20 years ago but now educational economists have access to large n-sets of data. This data has information about students, education, schools and performance. This has provided a way to answer the questions about educational production which was not possible previously [23]. In Australia, there are three core international datasets in which students are represented. These three data sets are PIRLS, TIMSS and PISA. This has helped to make these data sets richer. It is still in the first stage and it will grow. This will help government to understand the education system and to make plans accordingly to benefit the Australian community at large.

Accounting firms such as KPMG play a very important role in the Australian economy… KPMG has 145,000 staff worldwide and the chief information officer of the Asia Pacific region believes that big data has very good potential to be uses as proprietary data by combining it with new data. This will give KPMG a heads up in the competitive market. The chief information officer of a multinational company also believed that the company is now looking at predictive analysis, data mining and how the company can add value to customers' data [13]. Similarly, Thought Web is one of the leading IT companies in the Australian market and it is the developer of an enterprise analytics studio for building a big data solution. The company argues that customers are able to create competitive advantage by leveraging the ability to create information from different large data sets, to retrieve useful knowledge and to enable enterprise collaboration. In addition to that, challenges which are related to big data are volume, variety, velocity and complexity of large sets of data. These are the big data issues. Collaborative business applications can be helpful to analyze text, images, video, transaction etc. so that enterprises can see the patterns and act upon it [21]. Australia is slowly realizing the potential of big data in various industries. However, big data has not yet expanded that much compared to other companies and industries in the United States. There are several issues regarding big data in Australia. First of all, this country is very small and big data is still new in this market. However, industries like mining, healthcare, IT, education and retail in Australia are now reaping the benefits which can be provided by big data technology. Technology vendors such as Microsoft and SAP provide technologies which can be useful for data mining and business analytics on big

data. Thus, big data, which is still new in the technology industry, needs to be introduced on a large scale in Australia.

3 Big Data Critical Factors

Similarly to all other technologies, big data also has certain characteristics which differentiate big data from other technologies. These characteristics are volume, velocity, variety and complexity. Apart from these characteristics, there are other factors which play a decisive role in whether or not a company can afford big data.

a. Skills

Big data technology is new and still in its early stages. It requires certain skills and personnel with those skill sets to work with big data. As explained previously, the volume of large data sets is very high and they comprise a large variety which includes image, documents, audio and video files. It is important for people to have the skills to work with such data. In addition, big data needs to be processed at high speed. They need to be transferred from storage to personal computers where they can be processed. Moreover, they are very complex and thus require certain skills. Big data skills are nothing new; they are similar to analytical skills but with large sets of data. Most of the technological companies are now looking for people with certain big data skills. Even IBM has launched a big data program with universities worldwide. Both lecturers and students can participate in the program. The sole purpose of this program is to create awareness about big data and develop students who have big data skills when they graduate. The U.S. Bureau of Labor predicts a 24% increase in the demand for professionals with data analytics skills during the next eight years. Many universities around the world are now designing courses which are related to big data analytics skills. For example, Dublin City University has created a master's Degree in Computer Science with big data, business analytics and smarter cities focus. According to reports, employers in every industry are seeking employees who can analyze the large sets of data and gain a competitive advantage [15]. Every day they create 2.5 quintillion bytes of data generated from various day-to-day transactions. Big data and analytics are a catalyst to help each business to grow and to compete. IBM has one of the best technologies in the world to deal with big data analytics but these technologies need people with a certain skill set. There are big data technologies available which help to create information which have a business value. These technologies are Hadoop, Map Reduce, Apache Pig, Hive and HBase. However, these big data technologies require a certain skill set. Only people with certain skills can work with such technologies [31]. In this manner, big data skills are high in demand because there is a shortage of such skills. Obviously then, people who have big data skills, knowledge and experience will be highly paid in future. The world has entered the digital age.

Each and every industry is using technology to reap the benefits in a competitive market. But as work has become digital, people with such skills are the ones who are sought after. The computational world of the 21st century will require a particular new skill set from workers [35]. A study conducted by the Apollo Research Institute identified that a massive increase in data will enable new possibilities but it will demand people with new skills. Big data comes with several challenges to security systems but there are advantages as well. But, this data-driven world needs new skills which is essential for Australian industries. There are various models for big data with which people will need to work and during this time, people need to understand that these models can predict from the data but they (these models) cannot recreate it. This will require skills and knowledge of languages from people who are working on it. Design mindset is one of the solutions for this. It is simply the creation of an environment for the people who are working in the industry to develop and to enhance workers' creativity. There are main four skills which need to be learned by IT professionals. These skills are computational thinking, transdisciplinarity, design mindset and cognitive load management. Professionals who are working in IT will require new tools and the skills to work with those tools. Hence, it is clear that big data will require new skills similar to other technologies [35]. Companies are now paying high dollars to people with big data skills and they are taking a chance that investing in data can play a major role in their competitive plans [2]. Hence, it is important that companies realize the potential of people with big data skills in Australia and should provide competitive salaries so that people will be encouraged to train in big data. The next section discusses the technology associated with big data.

b. Technology

Big data technology itself is new technology but it needs certain tools which can accommodate it. It is important for the reader to understand that it should not be confused with tools of big data. This technology can be software, hardware or any other for that matter. This topic has the covered latest technology which is related to big data. In addition, it is important to understand that this topic is still new and that is why it is evolving. It is important to understand the readiness of technology which can accommodate big data. As explained previously, big data is associated with very high volume, velocity, variety and complexity. It is also understandable that this data is being produced from both the private and public sectors. So, it is necessary that these both provide the necessary resources to find technology and approaches to find such technology. Science and engineering can play an important role in finding such technology. Technology readiness level is one of the techniques which can be used to predict the likely time when the technology will or will not be available. It is a qualitative approach. This readiness level depends on the readiness level which once was part of a system's lifecycle process [6]. Big data itself can predict future technology and where it will come from. There is lot of information in the form of scientific articles, patent application, news articles, corporate websites and press releases. These resources provide information on technology and it

can provide a way to harness the advantages of big data technology. It is also important to understand that quality is more important than quantity as it gives momentum to new technology. Data scientists and computational engineers should also keep in mind that this technology evolves in a non-linear, complex fashion [6]. Moreover, there are certain pathways to the new technology but they vary depending on the industry sector and technology domain. It should also be understood that information which is coming from different sources is not diffusion limited. Last, but not the least, technicians should use the basic elements of big data with care because these are the pillars of the big data technology. Serious consequences can ensue if these elements are not properly distinguished from one another. Now, one would argue that what the areas are actually in which big data technology has a scope to evolve. Information technology is now playing a major role in each and every industry and each industry requires the perfect combination of software, hardware and storage requirements. There are vendors out there who are providing software as a service. But, there are several technology areas in which big data technology can evolve. These are networking, data mining, information security and privacy protection. Networking is an important part as velocity is crucial in big data technology. Data must be available instantly for analysis. Data mining is similar to data mining of normal data but here technology is required in which data mining can be done with large sets of data. Information security and privacy protection are both similar to those of normal data techniques.

Software tools are available in the market which uses some of the big data technologies. It can be also seen that most of the big multinational companies are now experiencing the benefits of big data and now deploying information systems accordingly. In addition, some of them have even completed research and development, design and promotion of new products and provided new services. These new services provide greater operational efficiency when compared to traditional database [7]. There are several technologies which are useful in big data analysis and processing. Map Reduce and parallel database technology are such technology. The research studies have shown that a combination of both these technologies can provide greater benefits than standalone technologies. So, it is important for scientists and technicians to find ways in which big data technology can be used more easily. Business Intelligence (BI) is part of big data. It includes the applications and processes used to analyze the big data using various analysis techniques [4]. An organization's performance and profitability can be increased through good analysis and decision-making. Business intelligence helps to achieve that through big data. Predictive analysis technology is another important term in big data technology. Analysts find certain patterns in the data and predict certain outcomes. This also helps management of organisations with decision-making. Hence, software tools have been developed for big data technology. Hardware equipment also plays an important role in big data technology. There are super computers which can respond to thousands of queries in seconds. However, it is not possible to provide a super computer to each person in the office. So, it is the role of technicians to find the technology which can accommodate big data practices on their computers. The third important part is storage technology. It is not easy to store such large amounts

of data in traditional storage systems. More importantly, data should be available for analysis from this storage point when and where required. In this manner, this part describes on big data technology which is available and how it can be more improvised. The next part describes about the maintenance of the big data technology.

c. Maintenance

As explained previously, IT is now everywhere and it is an integral part if not the most important part of any organisation. As IT has become an integral part of organisations, the job of IT professionals has become more important and it is all-encompassing. Almost every corner of any company or organisation is touched by information technology. If IT stops for a day, the whole company comes to a halt. So, it is important that the risk management department of any firm has a disaster recovery plan ready for such situations. This topic is important in this research study as big data is a very complex technology and if anything goes wrong in such widespread technology, then it can have a very adverse effect on business as management will be making very important business decisions based on the analysis of this data. Multinational companies can afford to hire talents who specifically specialize in maintenance and recovery plan only. The problem is with small and medium-sized businesses which cannot afford to hire such talents. So, the main problem with these organizations is that they simply cannot afford to commit time and resources to the maintenance which ultimately costs in big loss [26].There are two main points associated with maintenance which should not be confused with each other. These are the business continuity and the disaster recovery plan. These are entirely different. Business continuity planning is simply a methodology which is useful to create and to validate the plan for disaster recovery. It is useful to maintain business continuation.

However, disaster recovery is a part of business continuity. This term can be defined as a quick response to disaster. So, whenever there is any breach or fault in IT, the immediate next step will be to follow the instructions according to disaster recovery plans [26]. So, both the disaster recovery plan and the business continuity plan fall under the maintenance umbrella. There are main three components of any business. These are people, process and technology as IT is being maintained in this research study. Technology is being used by people who follow specific processes. These three components are further explained below. People are the entities who decide the processes and implement the technology to be used in the organisation. People are the most important element in maintenance as they create the processes and make the plans to fight a disaster and to recover from it. According to a recent study by IBM, 80% of total IT faults and data loss is caused by human error [30]. Hence, there are some problems which are usually caused by people themselves. So, if the IT maintenance department wants to make a start towards instituting a better maintenance procedure, then people should be covered first in that plan. Participation of such people is also important in such plans. The most important factor when deciding this is that this plan should not be designed in a vacuum, but

should cover all the areas of the organisation. The second element in this is the processes.

There is a simple rule that the better the processes, the better the technology implementation. As explained previously, people create these processes to implement technology. So, it is up to people to create better processes. So, processes should be defined accurately. These processes are important because they run the day-to-day business activities of an organisation [36]. So, when there is a disaster or any emergency situation, then these processes will be the first things to be terminated. It is part of maintenance to create such plans which will make it easier to return to previous processes and start working. It is also important to identify the most important processes. The third entity is technology in this research study. IT professionals have an idea of what happens to the technology in dire situations such as security breaches or any kind of storage failure or anything for that matter [24]. So, it is important to identify the technology that might be highly vulnerable to such attacks and if there is an attack on the technology, then what the plans are to tackle this situation. Moreover, it is important to work with people of other departments as IT is now each department and it is equally necessary to understand the demands of people of other departments. Their needs will be different from those of employees who are in other departments. This will help to clarify the situation better and the maintenance department can create processes which can meet the requirements of all departments. During this time, it is not only important to know the processes which are needed to make up and run the department but also it is important to define the processes which will be helpful in managing the crisis [33]. This should be assessed and addressed by the plan. Most of the middle-level companies do not design a maintenance plan until they are hit by the cost of a failure. It is important to understand that the cost of a failure is always bigger than the cost of planning. Yes, it is understandable that it requires resources and time to develop proper maintenance plan, but in the end it is worthwhile. Nowadays, most companies are moving towards big data. Big data technology requires very costly hardware, software and storage devices. Failures in such assets through natural disasters or for any other reason can cost a company a small fortune. Data will be added regularly even if the machines are not working at the other end. This can result in false predictions and ultimately wrong decision-making [14]. So, it is important that there are processes in the workplace which are there solely for big data technology. Maintenance can be resource-consuming in the case of big data but it will save dollars in the event of disasters. More importantly, for businesses to perform day- to-day activities successfully, it is important to have designed processes and plans for maintenance of costly assets.

d. Cost

As big data is a new technology, and as has happened with every technology in the past, the cost of implementing big data in a firm is really high. There is a certain belief in the IT industry that storage is comparatively very cheap. This comparison is in context with hardware prices. However, the real-case scenario is different. The

reason is that most of the time, various data operating expenses are ignored [29]. It is imperative for all organizations to implement strategies that maintain a balance between value creation and risk exposure, ultimately unlocking competitive advantage and maximizing value from the application of big data. Total cost of ownership gives a better idea of the total cost for a particular company. It is important to know here that this cost is related to hardware cost only. Other costs such as software and implementation costs have not been included in the total cost of ownership of big data at the moment. According to research studies, research cost is approximately five to seven times higher than hardware acquisition costs [3]. Moreover, users and application owners do not tend to realize this cost and believe that the cost of hardware is very low. Apart from total ownership, the maintenance of such storage also requires tremendous work and resources. It is important to understand that storage is not quite a problem at the moment but as data has transitioned into big data, the problem is becoming very serious. Some medium-sized organizations are investing so much in storage that they have to compromise on IT strategic investing. According to research studies, it was found that there are basically three costs with which any organisation has to deal. These three costs are hardware cost, non-hardware cost and combined costs. These costs are believed to be increased by 50% with this data growth rate [18]. Almost half of the investment in IT investment is spent on existing IT infrastructure. So, it is obvious that an organisation is not going to invest more in storage, and ultimately, will not benefit from improved storage conditions [29]. It is also believed that the costs of managing big data and the evolution of data analytics programs will increase. This will force organizations to spend more on big data related technologies. More importantly, people with big data skills will be in high demand and it is also important to attract the right talent for the organisation. In the last century, storage was not given proper resources as data generation was not a big problem at that time and thus CEOs spent most of their resources on software and hardware [20]. However, the world has changed so much afterwards and now, the data which has been generated in last five years is greater than the combined data generated before. So, now it is important to invest wisely in storage facilities as big data is not only about storing and locating data in one place, but it is also important that when big data has to be accessed, the data is available for data analytics. So, it is important that companies invest in storage regardless of the cost. There are various options available on cost-effective ways for storage of big data. One of the simplest forms is to use storage tires which span different areas of the information life cycle [29]. The second important factor related to cost is hardware cost. The legacy systems cannot keep up with real-time data analytics and thus really fast processing device is needed, if not super computers. The third cost factor is associated with software and the people who, of course, work with big data technology. It is really important to understand and find the technology which best suits the culture and needs of the organisation [37]. Different types of organizations need different types of technologies. For example, the technology which suits healthcare organization might not be of any help to an education organisation and vice versa. So, it is

important to understand the real value of technology and select suitable hardware and software which complement each other.

Hadoop provided one of the low cost technologies in 2010 but today it can be seen that there are several inherent weaknesses that need to be overcome when implementing big data analytics [34]. There are various challenges which are related to the implementation of big data. Customers want easier and simpler technology and thus it becomes costly to develop such technology. In addition, it is also difficult for an organisation to address these challenges and the acquisitions of such skills. Hence, it can be concluded that there are four main factors which must be considered and examined before companies take any steps to implement big data. There is a wide range of technologies available in the market and organisations can choose the technology which best suits a particular organization's values and vision goals. For example, the oil & gas industry can select big data technology which helps to find natural resources. Another factor is skills. People with big data skills will be highly in demand in coming years because of the increased data rate. So, it is an important task of the human resource department to hire personnel skilled in the management of big data. In addition, it is important to provide training to existing staff. The third factor is maintenance. Research studies have shown that most organizations spend half of their IT resources on maintenance and ongoing activities. So, in the case of big data, it is important to take precautions for maintenance so that even if there is any security breach or failure, it does not affect the ongoing projects of the company or affects minimal. The fourth and last factor is cost. As it happens with all technology, there are three main types of costs associated with big data technology. These are hardware costs, non-hardware costs and storage costs. The following section explains the type of tools which can be used in organizations to manage big data.

e. Tools

Big data cannot be operated on its own. That is why it is important to have suitable facilities which can accommodate the needs of big data. There are three most important things which should be available in order to execute tasks of big data. These three things are software, hardware and storage. This is not possible without the appropriate tools. There are various tools available from different technology vendors which can be used in different industries. These tools make analytics easier. The thing with big data is that it is indeed very big and storage is therefore a problem. In addition, data must be available at the time of analysis. There is a need of analysis tools because organizations increasingly want to move beyond offline analysis of extracted or summarized data to incorporate all relevant data in their business processes in real time [5]. There are five main tasks that any tool should be able to perform: reporting, analysis, visualization, integration and development. The tool which supports these functionalities should be considered. Big data is more complicated only because its scale is very much higher compared to old data scale and that is why there is a need for new tools and technology

because old technology and tools cannot keep up with new generated data [19]. Some of these tools are described below:

- *Pentaho tool:* Pentaho was released first as a report generating engine which was then turned into a successful business analytics tool. One of the advantages of Pentaho is that it can be used with NoSQL databases such as MongoDB. Afterwards, analysts just have to drag and drop the columns for visualization. Hence, it is a simple-to-use business analytics tool. In addition to that, sorting and shifting of tables in Pentaho is quite simple and useful. Moreover, it can be used between different clusters.
- *Jasper soft BI suite:* This business intelligence tool was developed on the JAVA platform. The unique selling point of this tool is that it was one of the open source tools available in the market and it is the best of the open source tools when considering its functionalities. Avery attractive functionality is that it can turn SQL tables and information into pdf format files. Moreover, there are functionalities which can fetch the data from MongoDB, Cassandra and other databases. Jasper soft can convert this data into interactive graphs and tablets. However, there are no innovative ways to look at data using this suite.
- *Karma sphere studio analytics:* Karma sphere originally was not designed to provide analytics functionalities and it was just a plug-in. However, it has several specialties which can help to run one or more Hadoop jobs. There is also a tool named Karmasphere Analyst which helps to simplify the process of going through all of the data in a Hadoop cluster. In addition, there are various tools available for programming jobs. So, overall, it provides good support for programs built in Hadoop although it does not support all databases. Thus, it is a good big data tool but is not the full package.
- *Talend open studio:* is similar to Karma sphere in that it was also designed as a plug-in for Eclipse and it also offers the IDE-based data processing jobs of Hadoop. Apart from that, it provided functionalities like data integration, data quality and data management. There is a canvas-like structure in Talend which allows analysts to drag and drop icons on this canvas. Talend has also a rich source of extensions which makes it easier to work with the company's other products. The integration of other projects is made fairly simple with this tool as well.
- *Sky tree server:* Sky tree differs from other tools in a unique way. Sky tree is designed in such a way that it is easier for analysts to string code together with visual mechanisms. Also, it offers a functionality which can perform many of the advanced sophisticated machine-learning algorithms. Analysts can do this by simply typing commands in the command line. The developers argue that this server is optimized to run a number of advanced machine learning algorithms on data which is ten thousand times faster than other packages. There is a free version of software available which has, of course limitation: it can accept only 100,000 rows. However, any company can check this free version before implementing it and thus 100,000 rows of data will provide a good idea about whether or not this software is for the one best suited to the company.

- *Tableau desktop and server:* As discussed above, most of the aforementioned tools do not enable the data to be examined in an innovative way. However, Tableau offers this. It is basically a visualization tool which makes it simple to look at data in a different way. In addition, this data can be broken down into portions and again each portion can be visualized in different ways. The option of mixing data is also available. Dozens of graphical templates are provided. Also, this software is well structured and its GUI is outstanding as well. So, overall, this tool is one of the best big data tools currently available.
- *Spelunk:* This big data tool is similar to traditional databases and it is completely different from its peers in every manner. Unlike all other similar tools available in the market, this tool does not offer any functionality for visualization and graph generation; on the other hand, this tool provides an index to data files like a data column in a book. This is just like traditional databases, but indexing here is very flexible. It is being sold in a number of application packages. Co-relation becomes easy with the help of the indexing feature provided by this tool.

These are some of the big data tools available in the market at the moment. There are other tools for big data as well and it is assumed that in coming years, there will be more tools available as big data is still in its infancy. Each and every tool offers functionalities. It is foolish to select any tool randomly. So, it is advisable that, IT department of any firm should spend considerable amount of time to choose the perfect big data tool which best suits the company. There are free versions available of different tools which can be used for testing purposes before investing in them. Also, the type of tool(s) chosen also depends on the type of data being generated in the firm and future growth rate of the data. Hence, big data tools are an important part of any company's big data strategies.

4 Research Gap

As explained previously, big data is still in the early stages and that is why people in different industries find it difficult to understand. There are perfect practices related to big data which involve best practices and techniques. In Australia, some of the big companies have started to use big data in their IT operations. Others who have not yet included it in their operations are preparing and making changes accordingly. However, companies are not taking the best approaches or applying the most appropriate techniques to big data. Big data, if used with suitable techniques, can provide maximum benefits. It is because of very little exposure to real use of bi data in such industries [12]. Big data depends on highly sophisticated software, hardware and storage services. Most companies have transformed their legacy systems to more advanced big data systems. This is why most employees are not aware of the best practices associated with big data. Ultimately, this leads to big data not being used to full advantage.

There are various gaps with regards to big data in Australian industries. This has been explained in this literature by the authors. Four main factors have been discussed in this literature. These factors are software, hardware, technology and storage. Software is at the heart of any new technology. This technology should be supported by appropriate techniques and moreover, these techniques should be the best possible approach [27]. There are various types of software available in the market and it is important that the IT person choose appropriate software best suited to the requirements of the company. For an example, the oil & gas industry should choose software which includes geological information systems so that the companies in this industry can fully benefit from big data. There are different types of software platforms available and CEOs of companies should play an important role in choosing suitable software. However, in Australia companies, it can be seen that there are not appropriate techniques to choose software and that is what makes this a very big gap. One of the most important factors which differentiate legacy systems from advanced systems is hardware. The selection of suitable hardware is very important in big data technology as data size is of very high volume, variety and complexity. That is why it is important that people in IT companies should select hardware according to the companies' needs. One of the deciding factors in big data is availability of technology which best fits the operations of the organisation. The thing with information systems is that they change and change quickly. Upper management still does not see much value in IT. That is why it seems unrealistic to the upper management to invest such amounts and resources every two years. Now, because of this, they will not have inadequate infrastructure which can accommodate big data very well. So, it can be said that it is the responsibility of top level management of any organisation to see that big data is being deployed flawlessly. So, there are basically three main points which contribute in creating gaps in big data management. These points are as follows: Absence of ability to store and access data easily; Absence of technology which has high computational power and highly sophisticated software which can manipulate the data; Talented employees who can analyze the data [11].

These three play important roles in the successful use of big data. In addition to that, it is also the responsibility of IT personnel of individual organizations to create a strategy which seems to be problems as well. Storage of data is one of the most important things in big data case. In addition to that, this stored data needs to be made available swiftly during the analysis process. It does not matter if the storage facility is located on company premises, although it is still difficult to find an appropriate storage facility which can meet the needs of big data. This is the first gap problems most of the companies have never previously experienced storage problems. The second reason for the gap is the absence of technology. The computers in big data need to perform thousands of queries per second. But, computers with such computational powers are very costly so it is not possible to give every analyst such high computational machines. Moreover, these high computational machines need software to perform these operations. This adds to the gap as well. As explained in the previous chapter, there are amazing tools available which are being provided by IBM and other technological companies. Hardware resources are

as important as software. It is understandable that legacy systems cannot work with big data and that is why it is important to have systems which can produce results in real time. This is a large gap as these systems are very costly and it is important that top management sees the value in this. In this manner, this gap has been created. Last but not least, at the end of the day, people work with these machines and analyze the data. So, if people are not capable of doing such things, then big data will not matter at all. So, it is important for organizations to find the right balance between talented employees and big data. Employees should know what they are working with and how they can make the most of it. Hence, these are the three main reasons which contribute to the big data technology gaps. The top level management can play crucial role and it is also responsibility of IT departments.

5 The Chosen Research Approach and Research Question

This study aims to answer the following research question: What are the essential critical factors for big data implementation in Australian Organizations? To examine and answer this research question, the authors employed an online survey in the Australian industries via Qualtrics. Surveys are one of the easiest and simplest research methods of all. Designing a survey is a difficult task but once it is done, the rest of the procedure is very easy. Participants can understand concepts of the study very well from the survey. Also, because of the Likert scale, they can answer the questions without becoming confused. The survey research method has several other advantages. Once the authors have received all the responses from participants, he or she can insert this data into data analysis tool and can generate accurate information from that. The survey research method has been chosen for this research study since it is faster, cheaper, more accurate quick to analyze and easy to use for participants and researchers, flexible and results will be presented in various formats from SPSS, word, excel, PowerPoint and PDF, although online surveys are outstanding tool for collecting data, however, hacking, and internet infrastructure can be a problem for using online survey in a research [8–10, 25]. The survey consists of five parts based on the current literature review. The first part asks for the demographic information about the participants. In this part, each participant provides his or her information on job role, company, age and gender. This helps authors to analyze the data from various perspectives. The next four parts of this survey focus on the questions which are related to the big data and information. These questions will help to convey the main purpose and requirements of this research study to participants. A Likert scale has been used in this survey design as it is easy to use and to understand. There are basically five options: strongly disagree, disagree, neutral, agree and strongly agree from which the respondents can choose their answer. The target population for this study is the employees in different industries who have an IT background. People with an IT background have been chosen because this research study focuses more on big data technology, and these people's views on big data are most important. These people

Table 1 Participants—prepared by the authors

Answer	Count	Percentage (%)
Industry		
Educational	17	19.77
Retail	15	17.44
Oil & gas	15	17.44
IT	18	20.93
Health	16	18.60
Other	5	5.81
Total	86	
Age		
Male	42	50
Female	42	50
Total	84	
Job designation		
Entry-level	21	25
Mid-level	29	34.52
Manager	22	26.19
Other	12	14.29
Total	84	

have IT backgrounds and e-mails were collected from social networking websites such as LinkedIn, Facebook and twitter. Other resources like yellow pages also helped to find these resources. After those e-mails were collected, they were sent to the participants of the study. The participants of this study have provided their opinions willingly. LinkelIn, Facebook and Twitter applications are used to distribute the online survey link. Table 1 shows the online survey participants for this study; in total 86 completed the survey. The majority of participants are from IT, followed by educational, and health with 20.93, 19.77 and 18.60% respectively. As for the gender, 50% for both Male and Female. The highest percentage of Job Designation are 34.52% from Mid-level while, 26.19% from the Manger level.

As this study focuses on five different industries, fifty members each from healthcare, education, oil & gas, retail and IT have been chosen. The survey questions were designed accordingly. The online survey tool 'Qualtrics' has been chosen to distribute the survey and collect the responses. The survey was sent to participants via email. The survey link was provided in the email so that participants could easily respond. The collected data was then stored for analysis. Moreover, the literature related to big data has been reviewed. The data was collected according to the guidelines of the ethics committee. The participants in this research study completed this survey voluntarily. The quantitative research method produces data which the authors have analyzed using SPSS (version 24). SPSS provides a rich set of options to perform analysis on input data. There are too many statistic techniques to choose from in SPSS and that is why it becomes more important to choose the most appropriate technique to analyze data. It has also been used to categorize data

according to measurement scales. Coding errors and input errors can have an undesirable effect on the output, so testing of data is done before final input. Finally, the survey questions were compared with those of previous studies. Literature has been reviewed to ensure that this survey meets the validity requirements. Moreover, the research results have been generated using appropriate statistical techniques and thus they are reliable. All the guidelines have been followed to ensure that the research study throughout is reliable and has validity.

6 Results

In this section the authors will discuss the outcomes from the online survey, this discussion will discuss based on the factors namely: skills, technology, maintenance, and cost. The KMO for the skills, technology maintenance and cost are Middling and Mediocre according to Beavers et al. [1]. The Bartlett's test of sphericity is highly significant for skills, technology maintenance and cost, $\chi 2 = 94.897$, 80.525,169.850 and 64.870 df = 45, 15, 36 and 21 respectively and $p < 0.000$, indicating that the items of the scale are sufficiently correlated to factors to be found [32, 16] in Table 2.

Skills are a very important part of any particular strategy. People ultimately work with different technologies and that is why it is important to have people with appropriate technology skill sets. In the first phase of the survey, people have been asked to choose their answer from given five options on various skills. These skills are collaboration skills, information technology skills, communication skills, hardware skills, software skills, database skills, people skills, research skills, big data skills and other skills. These skills have been considered to work with big data. Participants have provided their views on these skills and the level at which they would be required in order to work with big data. Table 3 shows three new skills which are required for big data implementation especially in Australia; technical, analytical, debate, applying, analyzing, synthesizing and IT skills.

Technology changes in every decade. Technology is important in terms of both hardware and software. Moreover, the database is one of the most important parts of big data. In addition to that, companies already have legacy systems which they cannot write off from the organization just like that. So, it is important to consider

Table 2 KMO and Bartlett's test—prepared by the authors

KMO and Bartlett's test		Sills	Technology	Maintenance	Cost
Kaiser-Meyer-Olkin measure of sampling adequacy		0.602	0.673	0.773	0.672
Bartlett's test of sphericity	Approx. Chi-Square	94.897	80.525	169.850	64.870
	df	45	15	36	21
	Sig.	0.000	0.000	0.000	0.000

Table 3 Skills—factor analysis—prepared by the authors

Rotated component matrix[a]				
		Component		
	Skills: working with big data required	1	2	3
Communication skills	Technical skills	0.751	0.213	0.102
Statistical skills		0.724	0.294	−0.135
Project management skills		0.662	−0.203	0.265
Problem solving skills	Analytical and debate skills		0.743	
Debate skills			0.700	0.117
Collaboration skills		0.329	0.561	
Critical thinking skills	Applying, analyzing, synthesizing and IT skills	−0.215	0.214	0.785
Information technology skills (New & Latest technology)		0.189	−0.173	0.615
Research skills		0.147	0.115	0.519
Search skills			0.340	0.373

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
[a]Rotation converged in 6 iterations

participants' views on recent technologies. In this action of the survey, there are various technologies types which have been considered. The participants selected their opinions from five options on a Likert scale. Table 4 generated two new technologies based on the survey outcomes namely: Legacy existing technology and using New Technology and Storage and web technology. The new outcomes from the survey should raise the awareness among the Australian industry that using big data needs new technology.

Table 4 Technology–factor analysis—prepared by the authors

Rotated component matrix[a]			
	Technology: big data required latest	Component	
		1	2
Legacy systems improvement (i.e. existing information systems)	Legacy existing technology and using new technology	0.787	
Software (i.e. IBM Hadoop)		0.723	0.103
Technology (i.e. Cloud computing)		0.688	0.291
Data center	Storage and web technology	0.166	0.834
Web Technology (i.e. web 3.0)		0.433	0.620
Hardware (i.e. super computers)			0.607

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
[a]Rotation converged in 3 iterations

Table 5 Maintenance—factor analysis—prepared by the authors

Rotated component matrix[a]			
	Maintenance: Big Data required maintenance for:	Component	
		1	2
Data center	Ongoing storage and projects	0.779	0.169
Ongoing projects		0.750	
Software		0.662	
Asset (i.e. Hardware and knowledge management)		0.639	0.444
Data safety		0.440	0.414
Hardware		0.374	0.324
Review (i.e. Monthly, quarterly or yearly)	Ongoing review maintenance		0.825
New technology		0.246	0.712
Service (i.e. IT service)		0.157	0.678

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
[a]Rotation converged in 3 iterations

Maintenance is important because it has been said that most of the organizations spend at least 47% of their TI resources on maintenance to ensure that all processes run smoothly. So, it is important for participants to consider how big data implementation in the current situation can affect the organisation. The authors have included various factors of maintenance which can have an effect on big data implementation. These factors include IT maintenance, hardware maintenance, legacy systems maintenance, software maintenance, service maintenance, ongoing project maintenance, review maintenance, people maintenance and general maintenance. Table 5 created two new concepts from the Australian's industry perspective namely ongoing storage and projects and ongoing review maintenance. The following result should be taken into consideration by the Australian industry to ensure that big data working successfully and productively.

Finally, cost of new technology is high. Moreover, the hardware and software which can be used with big data are costly. Also, people who work with big data need comparatively high pay as well. Apart from hardware and software, database plays an important role as well in the designing of big data. The authors have included several factors which can have an effect on the final cost of big data implementation. These are hardware, software, database and other cost implementation issues. Table 6 confirmed that big data cost is mainly focus on maintenance and IT management, cutting edge research and tools. Therefore, Australian industry should consider and assign some type of budget for big data especially for maintenance and research.

Table 6 Cost—factor analysis—prepared by the authors

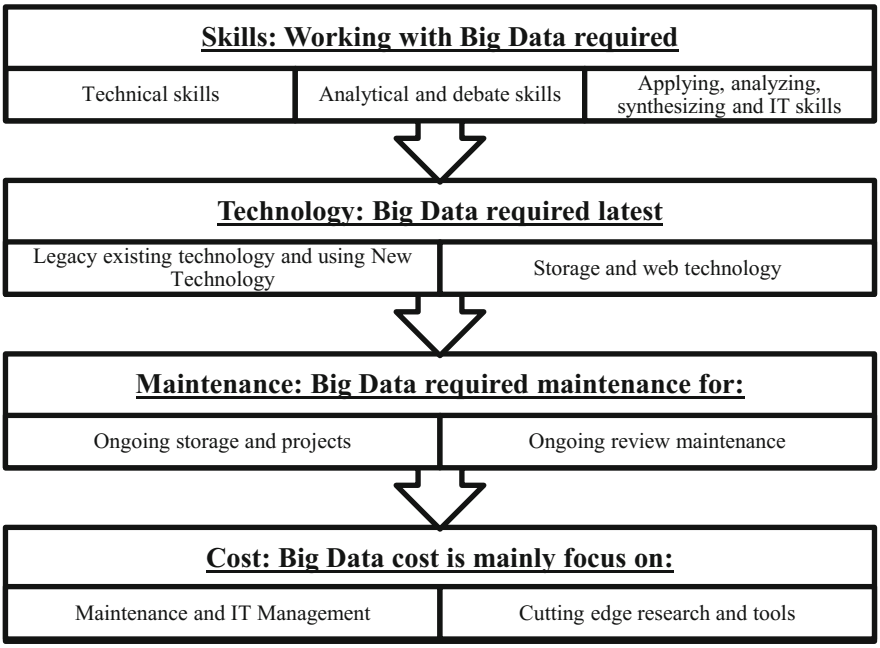
Rotated component matrix[a]			
		Component	
	Cost: big data cost is mainly focus on:	1	2
Talent retention	Maintenance and IT Management	0.676	0.218
IT management		0.644	0.226
Maintenance		0.604	0.125
Research	Cutting edge research and tools	0.389	0.636
Software		0.186	0.610
Hardware		−0.512	0.605
Database		0.227	0.585

Extraction Method: Principal Component Analysis
Rotation Method: Varimax with Kaiser Normalization
[a]Rotation converged in 3 iterations

7 Discussion, New Findings, and Limitations

The world has entered the digital age. It is well known that information technology is an integral part of every organisation now regardless of the background of the company. For example, education organizations have nothing to do with IT with it being a subject as an exception. However, these industries still need information systems in their business operations. Every small business is heavily dependent on information technology. Social networking websites have played their role in this change as well. In addition to that, organizations' possess additional resources of information which keep generating information. All this data can be very helpful in developing business by using this information. Big data comes into the picture at this point. This large set of data carries very valuable information that can be converted into knowledge and this knowledge can be used to further develop the business. Therefore, big data helps to generate knowledge from un-used data. Big data is now a new technology trend and thus it needs very advanced hardware and software devices. Big data is very large in terms of variety, volume and velocity. Big data has 'variety' because the data comprises images, text, documents and every other category. These elements create a very broad range of variety and thus it contains vital information. Big data, as its name suggests, has a very large volume. The third and interesting part is its velocity. Storage plays a very important role in the case of big data. Big data cannot be stored in traditional storage systems and that is why it is important to have dedicated storage systems for this type of data. Another important factor of storage is that it should allow the prompt retrieval of data for analysis as needed. This means that analysts need to have access to the data in real time. Where the data is stored is not important as long as it can be accessed for analysis purposes quickly in real time. Another important characteristic of big data is its complexity, making it more difficult to understand, it provides only meaningful information only if it is managed and analyzed appropriately manner.

Table 7 Factors for implementation of big data in Australian industries—prepared by the authors



The next section will discuss the new results from the online survey from an Australian perspective toward big data implementation.

The results which have been analyzed provide information on the factors which can have an effect on the Implementation of big data in Australian industries (see Table 7). It is understandable that these four phases have different factors and factor analysis was performed to find specific factors which are responsible for the implementation of big data. From the results, it was found that, in case of skills, there are three factors which can influence the Implementation of big data. These factors are shows three new skills which are required for big data implementation especially in Australia; technical, analytical, debate, applying, analyzing, synthesizing and IT skills. In the case of technology, there are two factors. These factors are Legacy existing technology and using New Technology and Storage and web technology. In the case of maintenance, there are two factors as well. These factors are ongoing storage and projects and ongoing review maintenance. The fourth and last component is cost. There are two factors which can have an effect on the Implementation of big data in Australian industries.

These two factors are maintenance and IT management, cutting edge research and tools. From the abovementioned analysis, it can be seen that all these four factors have an effect on the Implementation of big data. This shows that the research question is answered through this study; by identifying the essential critical factors for big data implementation especially in Australian organizations. Here, in the above study, it has been mentioned that there are various factors of big data which have an effect on the Implementation of big data. From the research, it was also found that industries like IT and retail tend to be softer on big data-related improvements. In addition to that, employees, who are at higher level in IT, are optimistic about big data related technologies. Hence, this research study answered the research questions and identifies the factors of big data from the Australian perspective. Finally, is study was limited to Australia, therefore, in the future, authors will conduct the same survey in various countries, including developed and developing, with larger and more diverse groups of organizations is required to strengthen the research findings and aims.

8 Conclusion

This chapter examined the new technology trend which is big data. Big data is very unique in terms of variety, volume and velocity. This type of data cannot be analyzed using traditional legacy systems. Hence, traditional software and hardware cannot be applied to big data. The second part of this research study focuses on Australian industries. Five industry sectors have been chosen for this research: oil-gas, IT, education, healthcare and retail. Several factors have been identified which can have an effect on the big data, especially from the Australian industries from the skills, technology maintenance and cost. There are several methods available, but survey method has several advantages. The chapter results confirmed that big data implementation needs specific and certain skills from technical; analytical and IT skills, technology, including the adoption of web technology i.e. super data centers, web 3.0 and cloud computing. Maintenance is crucial in big data implementation, especially the storage and technology, while the cost is vital for employing and applying big data in the Australian industries especially in the maintenance and research. In the future, authors will employ the same survey in numerous countries with larger and more sundry groups of organizations is required to strengthen the research findings and aims. Finally, the authors believe that this study will be helpful for future research on big data.

References

1. Beavers, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J., Esquivel, S.L.: Practical considerations for using exploratory factor analysis in educational research. Pract. Asses. Res. Evaluat. **18**(6), 1–13 (2013)
2. Bednarz A: Big Data Skills Pay Top Dollar: Mastering Big Data Languages, Databases and Skills Could be the Ticket to a Bigger Paycheck. https://www.scribd.com/document/250490075/Big-Data-Skills-Pay-Top-Dollar-Network-World(2014). Last accessed 3 Jan 2017
3. Boyle, J.: Biology must develop its own big-data systems. Nature **499**(7456), 7 (2013)
4. Brands, K.: Big data and business intelligene for management accountants. Strateg. Finan. **95** (12), 64–65 (2014)
5. Carlos, S.: MarkLogic® Executive to discuss tools for big data at TDWI BI executive summit. http://www.marklogic.com/press-releases/marklogic-executive-to-discuss-tools-for-big-data-at-tdwi-bi-executive-summit/(2012). Last accessed 3 Jan 2017
6. Cunningham, S.: Big data and technology readiness levels. IEEE Eng. Manag. Rev. **42**(1), 8–9 (2014)
7. Dawei, X., Lei, A.: Exploration on big data oriented data analyzing and processing technology. Int. J. Comput. Sci. Issues **10**(1), 13–18 (2013)
8. Dillahunt, T., Mankoff, J., Forlizzi, J.A.: Proposed Framework for Assessing Environmental Sustainability in the HCI Community. In: CHI 2010, Atlanta, GA, USA, 2010. pp. 1–3
9. Dillman, D., Glenn, P., Tortora, R., Swift, K., Kohrell, J., Berck, J., Messer, B.: Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Soc. Sci. Res. **38**, 1–18 (2009)
10. Dillman, D., Reipus, U., Matzat, U.: Advice in Surveying the general public over the Internet International Journal of Internet. Science **5**(1), 1–4 (2010)
11. Eddy, N.: Big Gaps in Small-Business Data Backup Plans: Carbonite. http://www.eweek.com/c/a/Data-Storage/Big-Gaps-in-Small-Business-Data-Backup-Plans-Carbonite-562817 (2012). Last accessed 3 Jan 2017
12. Evans, M.: Fragmented care: big gaps remain in sharing patient data between retail clinics and health systems. Mod. Health Care **44**(7), 14 (2014)
13. Foreshew, J.: Big data high on KPMG agenda. http://www.theaustralian.com.au/business/technology/big-data-high-on-kpmg-agenda/news-story/944ec6c23d45e3e4fc94f5fa9134e991 (2013). Last accessed 3 Jan 2017
14. Gijzen, H.: Development: big data for a sustainable future. Nature **502**(7469), 38 (2013)
15. IBM (2013) IBM Narrows Big Data Skills Gap By Partnering With More Than 1,000 Global Universities. https://www-03.ibm.com/press/us/en/pressrelease/41733.wss. Last accessed 3 Jan 2017 2017
16. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 53 (1):59–68. (2010) https://dx.doi.org/10.1016/j.bushor.2009.09.003
17. Knapp, M.M.: Big Data. J. Electr. Res. Med. Libr. **10**(4), 215–222 (2013)
18. Mahon, D.: Cost-effective big data retention enables better analytics (2011). http://www.dbta.com/Editorial/Trends-and-Applications/Cost-Effective-Big-Data-Retention-Enables-Better-Analytics-75457.aspx. Last accessed 3 Jan 2017
19. Mattmann, C.A.: Computing: a vision for data science. Nature **493**(7433), 473–475 (2013)
20. Perry, P.M.: Analyzing Big Data: A Finance Chief's Guide. http://ww2.cfo.com/big-data-technology/2015/10/analyzing-big-data-finance-chiefs-guide/(2015). Last accessed 3 Jan 2017
21. PrNewswire, ThoughtWeb—Australian BIG DATA Player Poised For Global Growth. http://www.prnewswire.com/news-releases/thoughtweb—australian-big-data-player-poised-for-global-growth-182280751.html(2012). Last accessed 3 Jan 2017
22. PRNewswire, Actian, Data Transformed and Yellowfin BI Mashup Helps Kollaras Group Reap Big Data Rewards PRNewswire. http://www.prnewswire.com/news-releases/actian-

data-transformed-and-yellowfin-bi-mashup-helps-kollaras-group-reap-big-data-rewards-271223601.html (2014). Last accessed 3 Jan 2017
23. Pugh, K., Foster, G.: Australia's national school data and the 'big data' revolution in education economics. Aus. Econom. Rev. **47**(2), 258–268 (2014)
24. Shneiderman, B.: The big picture for big data: Visualization. Science **343**(6172), 730 (2014)
25. Smyth, J., Dillman, D., Christian, L., O'Neill, A.: Using the internet to survey small towns and communities: Limitations and Possibilities in the early 21st Century. Am. Behav. Sci. **53**, 1423–1448 (2010)
26. Snedaker, S.: Business continuity and disaster recovery planning for IT professionals. Newnes (2013)
27. Snijders, C., Matzat, U., Reips, U.-D.: Big Data: big gaps of knowledge in the field of internet science. Int. J. Inter. Sci. **7**(1), 1–5 (2012)
28. Stamford, C.: Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years. http://www.gartner.com/newsroom/id/2848718 (2014). Last accessed 3 Jan 2017
29. Tallon, P.P.: Corporate governance of big data: Perspectives on value, risk, and cost. Computer **46**(6), 32–38 (2013)
30. Talmon, R., Coifman, R.R.: Empirical intrinsic geometry for nonlinear modeling and time series filtering. Proc. Natl. Acad. Sci. **110**(31), 12535–12540 (2013)
31. Tambe, P.: Big data investment, skills, and firm value. Manag. Sci. **60**(6), 1452–1469 (2014)
32. Tobias, S., Carlson, J.E.: Brief report: Bartlett's test of sphericity and chance findings in factor analysis. Multivar. Behav. Res. **4**(3), 375–377 (1969)
33. Turk-Browne, N.B.: Functional interactions as big data in the human brain. Science **342** (6158), 580–584 (2013)
34. Unisphere Research, The New World of Business Intelligence and Analytics—Big Data Analytics—Tipping the Needle from Cost to Value by Oracle. http://www.unisphereresearch.com/Issues/4697-The-New-World-of-Business-Intelligence-and-Analytics.htm (2014). Last accessed 3 Jan 2017
35. Wilen-Daugenti, T.: Big data requires new skills. Trade J. **23**(6), 1 (2012)
36. Wogan, D.: Big Data. Big Energy. Scien. Am. **310**(1), 22 (2014)
37. Zhang, L., Wu, C., Li, Z., Guo, C., Chen, M., Lau, F.C.: Moving big data to the cloud: An online cost-minimizing approach. IEEE J. Sel. Areas Commun. **31**(12), 2710–2721 (2013)

Extending the Sana Mobile Healthcare Platform with Features Providing ECG Analysis

Katerina Tsampi, Spyros Panagiotakis, Elias Hatzakis, Emmanouil Lakiotakis, Georgia Atsali, Kostas Vassilakis, George Mastorakis, Constandinos X. Mavromoustakis and Athanasios Malamos

Abstract The great development of technology recently provides innovations that improve everyday life. The major benefit of it is that medicine is also affected, so better healthcare can be provided. In that context, it can be critical for patients who suffer from chronic heart diseases to have in their availability a system that can monitor and analyse their electrocardiogram (ECG) displaying either normal or abnormal findings. The current chapter describes such a system that uploads, stores, processes and displays an ECG, calculating certain ECG findings necessary for doctors to make a diagnosis. To this end, the SANA mobile healthcare platform,

K. Tsampi · S. Panagiotakis (✉) · E. Hatzakis · G. Atsali · K. Vassilakis
G. Mastorakis · A. Malamos
Department of Informatics Engineering, Technological Educational Institute of Crete, Heraklion, Crete, Greece
e-mail: spanag@ie.teicrete.gr

K. Tsampi
e-mail: katerinatsampi19@gmail.com

E. Hatzakis
e-mail: hatzakis@cs.teicrete.gr

G. Atsali
e-mail: gogoatsali@gmail.com

K. Vassilakis
e-mail: K.Vassilakis@teicrete.gr

G. Mastorakis
e-mail: mastorakis@gmail.com

A. Malamos
e-mail: amalamos@ie.teicrete.gr

E. Lakiotakis
Foundation for Research and Technology—Hellas, Institute of Computer Science, Heraklion, Crete, Greece
e-mail: manoslak@ics.forth.gr

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_12

with its OpenMRS open source enterprise electronic medical record system, has been chosen and extended in this work for storing, processing and displaying the ECG data. OpenMRS provides a user-friendly interface and a database for collecting medical big data. Analysis of ECG signals is leveraged by the Physionet toolkit. Physionet contains many ECG databases and the WFDB software for processing ECG signals. According to the scenario we have processed, an ECG is uploaded onto OpenMRS platform using a mobile device or any other Internet-enabled device and is stored in the database that OpenMRS uses. Then, ECG signal is filtered using a finite impulse response (FIR) filter to remove noise and using WFDB functions it is processed so certain intervals are determined. Finally, with the appropriate algorithms specific ECG findings are calculated. When the procedure completes, the results are stored into the database using SQL Queries. Using an HTML Form results and graphs are integrated into the OpenMRS website highlighting abnormal values with red color. Authorized users can have access to this information through any web browser.

Keywords Healthcare applications · Big data · Electrocardiogram
OpenMRS platform · ECG signal processing

1 Introduction

The development of technology provides us innovations to improve our everyday life. Information Technology (IT) describes the procedure of creation, manipulation, storage and secure exchange of different types of electronic data [1]. The major benefit is that the area of medicine has been affected as well in order to provide better healthcare. Although the initial target of mobile healthcare systems was to help patients in developing countries, they are regularly used in worldwide scale. Such systems are mainly destined to support health in village areas. The use of such a unified system increases the quality of health care. Additionally, the exponential progress of recent mobile devices with remarkable computational resources (CPU, RAM, etc.) combined with internet connectivity that Internet Service Providers guarantee allow quick and instant access to medical data using a simple device (smartphone or tablet).

Health Information Technology (HIT) is a sub-category of information technology that refers to healthcare data storage and processing [2]. It enables health data management and secure medical information exchanges. The Electronic Medical Record (EMR) or Electronic Health Record (EHR) is the major component of HIT. The use of EMR or EHR has significantly improved the healthcare quality.

C.X. Mavromoustakis
Department of Computer Science, University of Nicosia, Nicosia, Cyprus
e-mail: mavromoustakis.c@unic.ac.cy

The EMR systems describe systems that store medical information in digital format providing the potential to read clinical notes using a device [3]. Medical records may include a variety of data, including demographics, medical history, medication, vital signs and personal information such as weight and age. EMR systems use also data that express crucial fundamental health indicators such as blood pressure, respiratory rate, oxygen saturation, temperature etc. These indicators react as warning for instant doctor treatment in case of abnormal operations. One major form of such health data is electrocardiogram (ECG).

An electrocardiogram (ECG) signal describes the electrical activity of the heart using electrodes on the body surface over a time space usually ten seconds. The electrical activity describes the amount of impulses in a heartbeat and provides information about the heart rate, rhythm, and morphology [4]. Heart produces electrical current that we can record. Connecting electrodes between the body and the electrocardiograph creates an electrical circuit. Each pair of attachments is called "lead".

According to cardiologists, in order to evaluate an electrocardiogram, they have to extract empirically the required metrics. The ECG signal is represented on a graph paper, which makes easier the calculation of those metrics by the doctor. Our system provides the appropriate processing and produces the ECG findings, helping doctors to make a critical decision more quickly. Those ECG findings indicate:

- heart rate e.g. beats per minute
- amplitude of the ECG waves e.g. R-wave amplitude in lead I
- duration of the ECG waves e.g. R-R interval
- the heart axes e.g. QRS axis
- ratio between the waves amplitude e.g. V1 Ratio.

In this research work, we implemented a system in which users, who have authorized access, upload an ECG signal file through their mobile phone, or a web browser, to the EMR system. After uploading the file, ECG signal processing takes place. During the processing stage, we initially remove noise by using several filters (such as FIR). After this step, our system evaluates the amplitude and duration of the signal waveforms and this information is stored into the database. Our application also illustrates patient's heart performance into graphs providing representative visualization to the medical staff. Authorised users can access electrocardiograms via this medical platform and detect possible abnormal operation. Abnormal findings are highlighted with red colour to alarm the user.

The EMR system that is used for the current implementation is the OpenMRS platform. OpenMRS is compatible with mobile devices and provides connectivity with Sana, which is a well-known mobile Health system. The ECG signals come from Physionet PTB Diagnostic ECG Database. The PTB Database contains 15-lead ECGs. Twelve of them are the standard leads and the remaining three are the Frank XYZ leads. Physionet provides the PhysioToolkit software, which contains free applications and functions used for ECG signal processing.

The remainder of the chapter is organized as follows. Section 2 presents the related work in this domain and Sect. 3 introduces to our architecture. Section 4 discusses several implementation issues and illustrates the outcomes of our work.

Finally, Sect. 5 concludes our work and proposes possible future extensions of our system functionality.

2 Related Work

2.1 Overview

Healthcare applications contain a variety of technological achievements. Many researchers work hand in order to improve healthcare conditions. Recent innovative approaches often contribute towards this direction. For this reason, healthcare systems can be approached from many different perspectives. In the following sections, we analyze generally healthcare applications, the role of Big Data in this area and additionally the use of Internet of Things (IoT) technology and mobile devices for high quality healthcare services. Based on the above, we focus on Electronic Medical Record (EMR) systems by introducing similar projects and we conclude by describing the background of EMR systems that analyze electrocardiogram (ECG) and the required information for this topic, which is the main objective of the current research work.

2.2 HIT—Health Care Applications

Health information technology (HIT) improves the health of individuals and the performance of providers. It also strengthens the quality and effectiveness of health care, making it healthcare efficient and productive. A major factor is that HIT reduces the medical expense. Exchanging health information between doctors shortens the diagnosis-waiting period and also prevents from medical errors. HIT enables big data manipulation and sharing healthcare information in worldwide scale using Internet in order to make effective decisions and help medical research area.

2.2.1 Big Data and HIT

Big Data is a term referred to a large amount of data that expand rapidly. Big Data in health includes data originating from different types of sources related with health condition such as sensors and mobile devices. A subcategory of Big Data is the Medical Body Area Networks (MBAN), which provide continuous monitoring of patient's health by transmitting measurements from heart rate, blood pressure, respiratory rate, body temperature, and electrocardiogram (ECG). The role of the Big Data is crucial for clinical decision and health information systems [5].

The main features of Big Data in health are the volume, variety, velocity, veracity, validity and volatility. These characteristics refer to the amount of data, the sources that data stems from, the continuous data collection in real time and the availability over time [6]. The usage of Big Data is critical for taking instantly important decisions about patient's treatment. Big Data provides a proper and complete overview of a patient's condition, offering solutions against chronic diseases and tries to eliminate time that patients spend in hospitals [7].

In developing countries, quality of health service has to be high, so the applications of data analysis can contribute towards this direction. A platform of analyzing Health Data is OpenMRS, which enables the graph creation and allows quick access to databases that are useful in over 220 clinics [8]. Big Data are personal data and should be guaranteed and protected from harmful users. Finally, the Big Data analysis helps the improvement of healthcare services in developing countries [9], especially in cases of dangerous diseases for human life.

2.2.2 IoT-Based Healthcare Applications

Remarkable technology development during last decades has improved the health information technology. Different types of health care services are provided, preventing patients from staying in hospitals, improving their quality of life. The Internet of Things (IoT) area offers new technologies using sensor networks, providing a system that offers doctors the capability of remote diagnosis. In industry, there are sensors for movement detection such as gyroscope and some others for vital signs monitoring such as temperature [10].

IoT development in recent decades results in innovative applications that are useful in biomedical and healthcare research. Due to easy use and little size, portable devices are widely used in IoT-based applications. Vital signs are monitored via these devices and are sent to the doctors via IoT and Android Application Framework collects the required data in order to have an accurate diagnosis.

The Body Sensor Network (BSN) system consists of various types of wireless wearable sensors that collect measurements from patient's body and transmit them to a data collector unit (e.g. a database). Based on the metrics that they monitor, sensors are separated into two categories. The first category includes sensors that should monitor continuously and collect a large amount of data such as electromyogram (EMG) sensors and electrocardiogram (ECG) sensors. The second category includes sensors that collect smaller amount of data such as temperature sensors and blood pressure sensors. BSNs are also divided into categories based on the way of data transmission. Wireless sensors use wireless communication technologies such as Zigbee and Bluetooth, Radio Frequency Identification Devices (RFID) and Ultra-Wide Band (UWB) for communication with other sensors or devices [11].

Healthcare applications provide high quality services using sensors that measure biometric information. A critical decision that should be taken for reliable health monitoring describes the symptoms of the disease that should be examined. For

instance, some symptoms for bulimia appear when a person eats in secret or if food disappears in a short period. Furthermore, patients with bulimia have more complex health problems such as hypertension and fever. According to the symptoms described above, applications for bulimia detection should include sensors for movement detection in the kitchen area and sensors for measuring vital signs such as body temperature and blood pressure. Another example is the Alzheimer's disease that is a serious illness than can kill a patient if he is lost. Memory loss and depression are some fundamental symptoms. In this case, geolocation sensors are required because memory loss causes difficulties of orientation. Depression can be detected using a sensor that examines the rhythm of heart rate. Moreover, the intersection with use of blood pressure sensors will provide more information about patient's health that can lead in proactive diagnosis and successful treatment [12].

Biometric sensors can be used to detect stress or aggression in patient's behavior informing people who visit a patient in purpose to avoid violent attacks. IoT can also inform medical staff in cases when equipment of the hospital needs refill such as medication or oxygen tanks. RFIDs scan barcodes of items and warn the staff in case of trackable items is finished such as dressings. RFIDs can also be used in patients' home in order to inform him if a drug needs replacement [13]. It is obvious that IoT technology can provide significant help in doctoral and patient's daily life, leading to a high-quality healthcare system.

2.2.3 Mobile Health Care Applications

The development of the mobile devices has affected the area of medicine significantly. The medical software applications for mobile devices (m-Health systems) have become very important today. Clinicians have access to medical information through tablets and smartphones. Mobile devices have to satisfy some basic requirements in order to be useful in health area. Such operations are easy constant Internet connectivity, quick access in electronic medical records, capability of providing information about treatment or a disease and also can support interactive teleconference conversations between doctors located in different places for exchanging opinions for emergency cases.

The benefits of using mobile devices in health are many, including the improvement of the knowledge level and instant access to medical data. Rapid decisions are made with reducing risk and the quality of data accessibility and management is increased [14]. Patients can have direct communication better doctors during treatment and doctors can help instantly in cases that patients are located far away from hospital via a simple mobile device.

Mobile devices in combination with body sensor networks provide effective disease prediction in real time and prevent patients from spending money and time travelling or waiting for treatment in hospitals. In developing countries, many people that live in inaccessible regions can be treated by doctors easily which is valuable. According to the above, the health care system should incorporate mobile applications, which are used for storing medical data and accessing medical data via

Internet. The patient can use the application to store his medical information including symptoms and the doctor can use this application to view the medical history and write a diagnosis and medication. The patient is informed about the recording of the diagnosis by the application [15].

The architecture of the m-Health systems includes three layers, which are the data collection, data storage, and data processing. The data collection layer describes the procedure of collecting data via mobile device. The data storage layer is responsible for storing it in Big-data form and the data processing layer analyses the data and display the report or the diagnosis [16].

Many mobile healthcare platforms have been created to provide help and better patient care. A well-known mobile healthcare platform is eMOCHA (electronic Mobile Open-source Comprehensive Health Application), which is a free open-source application, developed by the Johns Hopkins Center for Clinical Global Health Education and was designed to be applied in developing countries. Moreover, the National Library of Medicine (NLM) Mobile Resources contains a collection of mobile-friendly websites and applications. Magpi (formerly EpiSurveyor) is a free mobile phone and Web-based data collection system for global health. FrontlineSMS is open-source software that allows laptop and mobile phone to communicate each other, exchanging messages. Additionally, RapidSMS is free and open-source framework for data collection and communication by sending and receiving messages [17].

One additional example of m-Health systems is called Sana Technology Platform, which was developed by a team at MIT and simplifies the procedure of data collection. Via a mobile device, a user can send medical data that the doctors can view, make a diagnosis and add a medication. The results from the doctors are accessible also via a mobile device. Sana can be integrated into OpenMRS [18] or other medical records systems.

Sana technology platform is an open source application that provides expandability and variety on storage and processing methods [19]. It is also ease-of-use and comprehensible. It provides a database in which texts, images, videos and whole folders can be stored. It ensures safe data transfer without loss or corruption even in areas without reliable internet connectivity. Using Sana utilities allows users to upload medical data, which then are stored into OpenMRS. Appropriate algorithms process the uploaded data and after this step, the results are sent from OpenMRS back to the Sana application. Data upload process requires WiFi, USB connection, GPRS and SMS. Finally, the Mobile Dispatch Server (MDS) is a program that is responsible for the communication between OpenMRS and Sana, receives and synchronizes the data [20]. The Fig. 1 shows the operation of the Sana technology.

The development of mHealth systems results in preventing health problems in case of patients who have difficulties in mobility. The way of life is a major factor that the developers took into consideration to build these systems. Mobile applications in combination with sensors have increased the quality of life in people in developing countries [21] and generally contribute in high quality medical services.

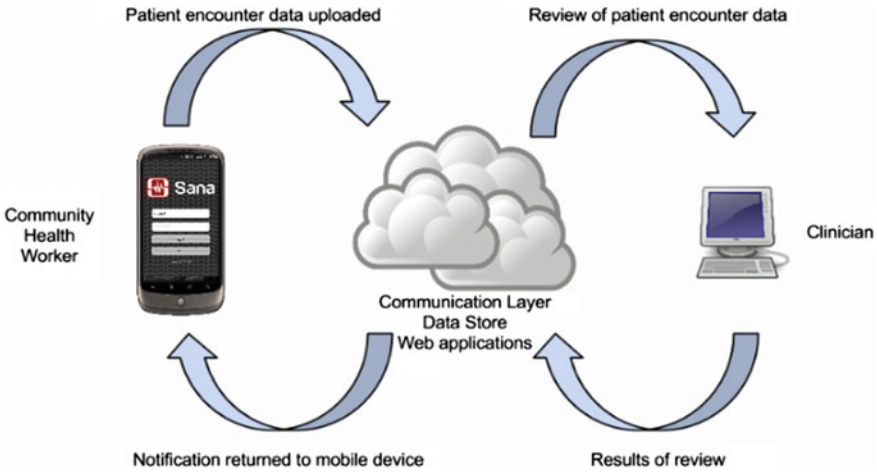


Fig. 1 Sana architecture

2.3 EMR Systems

2.3.1 Overview

Electronic Medical Record Systems (EMR) target to abolish the established "paper" in hospitals not only in the United States but also in developing countries. The quality of health care has increased using a system that contains all required information. The low cost of the installation and the wide use of those systems are the major benefit.

EMR systems store different types of medical data for many patients. They contain a database whose tables are flexible and in synchronization with each other and are also scalable in case of large-scale systems. They use widely known language and can export data in standard formats. In addition, these systems do not require complex operations and are characterized as secure in terms of privacy [22]. On the network side, EMR systems allow simultaneous data access and insertion from different places using parallelism. Furthermore, data synchronization is a major issue that EMRs provide when there is no internet connectivity. Due to the lack of reliability of the network in developing countries, a local database is used and when the internet recovers, the system synchronizes the medical data [23].

The systems that are preferred in developing countries have not to be expensive for the implementation so the proprietary medical record systems are not recommended. Over time, free proprietary software was developed such as Google Health, Microsoft and Healthvault etc. In order to avoid separated systems satisfying different hospital needs, systems should offer expandability among hospitals. Those systems provide open source software and are available for further processing. The Veterans Health Information Systems and Technology Architecture

(VistA) was the first EMR system that supported medical care for the soldiers in U.S. and used a language that was not widely used. Consequently, the use of VistA did not expand. Care2x is another well-known web-based system, which due to disorganization and the lack of structure, was not reliable. OpenMRS is a descendant of these systems, which needs no programming knowledge and provides adaptability, expandability and free installation. Java is used for OpenMRS programming and runs on the Apache Tomcat web server with a MySQL database. Using the concept dictionary that includes all diagnosis, drugs and all essential information related to medical care, make the analysis easier and accurate. Many developing countries are using OpenMRS such as Kenya, South Africa, Uganda, Tanzania and Zimbabwe [18].

Additional EMR systems are Mosoriot Medical Record System (MMRS), which during data storage process, mistakes happened. In Kenya, Partners In Health (PIH) web system should be reliable to ensure there is no data loss when there was no internet connection. After some years, the problem of the unreliable internet connectivity was solved and the system was working locally. In Uganda, the Careware system is a stand-alone database that uses Microsoft Access. In Malawi, the EMR that is used provides insertion data using a touch screen, which is difficult in case of long data. Similar active and open-source approaches are FreeMED, GNUmed, GNU Health, Hospital OS, HOSxP, OpenEMR, OSCAR, THIRRA, ZEPRS, ClearHealth, and MedinTux [24].

2.3.2 EMR and ECG

IoT evolution has offered great progress in health domain. Healthcare improves the quality of hospital services and dramatically decreases the required time for disease detection. Many IoT applications are used to contribute in healthcare such as blood glucose monitor and electrocardiogram (ECG) monitor. Via smartphones with Internet connection, heart condition is feasible independently space and time. Moreover, warnings inform doctors about emergency cases. iCarMa is a heart monitoring system that takes as input a photoplethysmogram (PPG) signal from sensors and detects heart diseases such as tachycardia and bradycardia. PPG signal uses infrared light at different body parts and detects changes in light absorption [25].

Cardiovascular diseases (CVDs) are diseases that involve heart such as myocardial infarction. CVDs may cause sudden deaths for example heart attack if there is no early diagnosis. Using some parameters such as age, cholesterol and blood pressure can prevent from CVDs. A mobile system can use these parameters and export results related to the estimation of cardiovascular risk preventing patients from heart failure. An ECG monitor is used for collecting ECG signal and the evaluation starts. When processing is terminated, ECG signal is displayed using a mobile platform [26].

The above analysis shows that EMR have critical role is medical area due to the capabilities they offer related to portability, early detection and accuracy. These

factors are crucial in many dangerous deceases that require constant monitoring and instant treatment. Heart diseases have also similar requirements and for this reason, EMR systems that use ECG are well known in medical industry.

2.4 Electrocardiogram

2.4.1 Overview

Electrocardiogram (ECG) describes the process of monitoring heart activity using electrodes over patient's body. A standard 12-lead ECG needs 10 electrodes. Leads are divided into three sets: Bipolar Limb Leads, Unipolar Limb Leads and Precordial Leads. The 12-lead ECG has three bipolar limb leads (I, II, III), three unipolar limb leads (AVR, AVL, AVF) and six Precordial leads (V1, V2, V3, V4, V5, V6) [4].

A typical normal heartbeat ECG consists of a P-wave, a QRS complex, and a T-wave. P-wave corresponds to depolarization of the right and left atrium. QRS complex reflects the depolarization of the right and left ventricles. T-wave corresponds to repolarization of the ventricles. Based on these values, researchers divide ECG into parts that support accurate diagnosis. PR Segment represents the delay in atrioventricular node and is the interval between the ends of P-wave and the start of the QRS complex. PR Interval begins from the start of P-wave till the start of QRS complex. QT Interval contains the duration of the QRS complex and T-wave duration and is the interval from the beginning of the QRS complex till the end of T-wave. ST Segment is a period of inertia and is the interval between the end of QRS and T-wave start. There is also the R–R interval that is the time period between 2 consecutive R waveforms. R-R is useful because it indicates the heart rate [27]. Figure 2 depicts the previously described terms used in ECG bibliography.

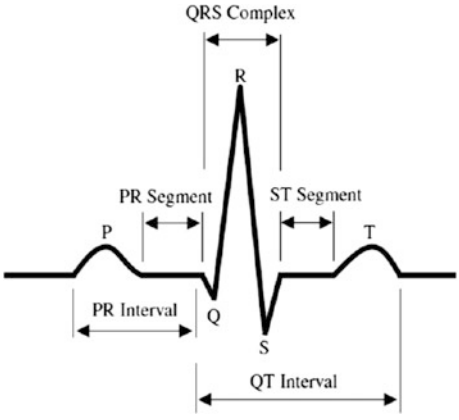


Fig. 2 An ECG heartbeat signal

2.4.2 ECG Standard Formats

According to the International Organization of Standards, the stored electronic medical data should be accessible by many authorized users and also secure transmission is required [24]. Many digital ECG formats have been implemented so far but all require political commitment and international cooperation.

Digital ECG formats and standards are divided into seven groups [28]. Formats that are supported by the Standard Development Organizations (SDOs) belong to the first group, independently of the nature of the data format. Well- known formats in this category are the Standard Communications Protocol for computer-assisted electrocardiography (SCP-ECG, European standard), the Health Level 7 annotated ECG (HL7, American standard), the Digital Imaging and Communication in Medicine (DICOM) and the Medical waveform Format Encoding Rules (MFER, Japanese standard). These formats are similar with each other and can be either binary or XML-based.

SCP-ECG is one of the most widespread standards for exchanging digital ECGs medical informatics standardization and is supported by the European Committee for Standardization (CEN). It is a binary encoded representation and intended for short term diagnostic ECGs. Many proposals for promoting the SCP-ECG standard had been submitted such as release of open source SCP-ECG tools under GNU General Public License, the development of implementation guides and programming contests and tools related to SCP-ECG. The need of the Food and Drug Administration (FDA) for digitalization a great number of varieties of formats annotated ECGs created an XML-based format for digital ECGs, which was the HL7 aECG. DICOM standard was developed for exchanging medical images. A DICOM extension is the handling of biomedical signals such as the ECG. Despite the advantages of DICOM for viewing, interchange, and archiving the ECG signals some users claim that the use of DICOM is limited. The last standard in this category is the MFER, which is a preliminary complementary standard, specialized in medical waveforms. Moreover, the MFER standard is expected to integrate into the 11073 group, which is referred below, and some improvement point are required such as a specification for standard 12-lead ECG.

The X73 Family of Standards in Digital ECG contains the less known formats that are supported by SDO. The Vital Signs Information Representation (VSIR) format was used in cardiology included an object-oriented domain information and service model. The File Exchange Format for vital signs (FEF) is a format that leveraged VSIR and biomedical measurements. The next format is the X73-Point (X73PoC) of Care specialization IEEE P11073-10306 for ECG devices and describes the data transfer between ECG Virtual Medical Devices. The X73–Personal Health Devices (X73PHD) standard refers to transmission of ECG data, 1–3 leads, between personal ECG devices.

The second group contains the existing binary encoded formats. The Holter applications contain a large amount of data and have different requirements that are covered by the International Society for Holter and Noninvasive Electrocardiology (ISHNE). The large amount of numerical data requires a different file format and

storage model that the Hierarchical Data Format (HDF) can cover. The last standard in that group is the improved version of SCP-ECG protocol that is called enhance SCP-ECG (e-SCP-ECG+) which overrides the limitations that exist in the SCP-ECG by creating new sections. This format is compatible with the SCP-ECG and accepts additional vital signs but also demographic data.

The third group uses eXtensible Markup Language (XML) format. This group is divided into two proposals, which are the general-purpose proposals such as PhilipsXML, ecgML, XML-ECG and I-Med and specific-use case proposals such as mECGML, ECGaware, Unisens and BSPM-XML. Philips created the PhilipsXML format that was used by its own electrocardiographs, was written in the W3C XML Schema Language and it was available including documentation and software tools for easy access. The ECG data are compressed using an algorithm without loss and encoded into ASCII using a base 64-encoding scheme. Philips XML format uses Scalable Vector Graphics (SVG) as display format and allows connection with other standards like HL7 aECG. The electrocardiography Markup Language (ecgML) format is recommended as a solution for integrating ECG data into electronic medical records that provide applications for easier use such as an ecgML generator and an ecgML browser. Its creators refer to the XML-ECG format as a simpler structure, which increases readability. I-Med standard provides the capability to exchange several types of medical data but also basic features such as QRS duration. The Mobile ElectroCardioGraphy Markup Language (mECGML) format is used for exchanging and storing ECG data between mobile devices. The ECGaware extends ECG standards to support patient's heart tele monitoring during daily activities. The UNIversal data format for multiSENSor data (UNISENS) provides recording and storing data processes from different types of sensors such as ECG, blood pressure and respiration rate. The XML-Body Surface Potential Map (BSPM) supports less prominent methods.

The fourth group contains formats intended for neurophysiology but is also used for ECG signals. Given the fact that the ECG signal has similar structure with the neurophysiological signals, such as electromyogram and electroencephalogram, the standards that belong to this group provides manipulation for these signals. The formats in this group are separated into two categories, which is the Data Format family such as EDF, EDF+, GDF, BDF and OpenXDF and the formats initially intended for neurophysiology such as E1467-92, SIGIF, EBS, SignalML and IFFPHYS. The European Data Format (EDF) is a 16-bit format suitable for exchanging time series supporting multiple sampling rates. EDF+ is an improvement of EDF, which provides interrupted and time-stamped recordings. The General Data Format (GDF) is designed to provide some extra modifications to cover some EDF limitations, such as coding scheme for events. The BioSemi Data Format (BDF) is a 24-bit version of the 16-bit EDF format, which was designed for electroencephalography applications but is applied in similar signals. The Open eXchange Data Format (OpenXDF) describes an XML-based extension of the EDF format. The E1467-92 format is a format that was designed to transfer digital

neurophysiological data but later it extended its operation allowing transferring multiple types of digital data such as ECG. The SIGnal Interchange Format (SIGIF) is highlighted due to its versatility and adaptability according to its designers and can store both raw and processed data. The Extensible BioSignal EBS file format is a binary file format for storing multichannel time-series recordings. The Signal Markup Language (SignalML) was used to avoid compatibility problems in case of different formats of digital data. The Interleaved File Format for PHYSiological data format (IFFPHYS) is an extension of the IFF format and provides storing ECG signals and other physiological signals.

The fifth group refers to main existing database formats. Many organizations have created their databases for research and experimental purposes. Open data format, software and reference materials are available by those organizations to handle their formats. Moreover, the databases include ECG data in a form compatible with the Standard formats such as EDF, BDF and SCP-ECG. Some of the existing databases are the Massachusetts Institute of Technology, Beth Israel Hospital (MIT-BIH) database, the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database and the Physikalisch-Technische Bundesanstalt (PTB) database, which are supported by the Physionet component and its Waveform Database (WFDB) software. Additional databases are the American Heart Association (AHA) database, and the Common Standards for Electrocardiography (CSE) database.

The sixth group contains the Integrating the Healthcare Enterprise (IHE) profiles that aim to ECG domain. IHE Cardiology is related to information sharing, workflow, and patient care in cardiology and its profiles contain stable documents, such as the Retrieve ECG for display, draft for trial implementation, such as the Resting ECG Workflow (REWF), and draft for public comment such as the Waveform Communication Management (WCM). IHE Cardiology Framework integrates existing standards, usually HL7 and DICOM.

The seventh group contains existing and ongoing works on ECG ontologies. Ontologies are embraced to define controlled vocabularies for shared use among different medical domains. Such ontologies are the SCP-ECG Ontology (SEO), the National Center for Biomedical Ontology (NCBO) and the NEMO.

There is one extra group, which contains the standards created by manufacturers to use them in their own ECG devices. Such standards is the Siemens Interchange Format for medical records (SIFOR) by Siemens, the Unipro by Mortara and the ECG-9x by Nihon Kohden. Many manufacturers have declared that their standards are compatible with the SCP-ECG.

In the current implementation, the PTB database is used. PTB database contains a compilation of digitized ECGs supporting the SCP-ECG format and is provided by the National Metrology Institute of Germany. It is supported by the Physionet and the PhysioToolkit software and is widely used for research purposes. Healthy control ECG signals are provided but also pathological ECG signals such as myocardial infarction and heart failure for experimental setup.

3 Proposed System Architecture

3.1 Overview

The EMR system that we use in the current implementation is the OpenMRS platform. OpenMRS is an open source EMR system with common framework and functionality [22]. The ECG signal comes from the PTB Diagnostic ECG Database in Physionet, which is supported by the National Metrology Institute of Germany. Our EMR system uses the PhysioToolkit, which is open source software used for displaying, analysing and simulating physiologic signals and contains the WaveForm DataBase (WFDB) software that consists of applications and functions that evaluate signal amplitudes and durations.

The implementation consists of three parts. The first part describes the ECG signal uploading process. The ECG uploading procedure exploits the OpenMRS functionality for uploading files. The second part refers to the ECG signal processing, signal attitudes evaluation and processing output transmission to the database. The third part includes the process of results visualization. The ECG file is stored into the server. The WFDB software and algorithms are implemented in the server-side and are triggered by ECG file upload. When ECG file is uploaded, filtering is applied in order to remove noise. A FIR (Finite Impulse Response) filter is used in this step. Signal processing results are sent to the database using SQL stored procedures. Data kept in OpenMRS database are accessible via HTML forms. An HTML form is developed to display the ECG findings and the graphs into the platform. The access to values into the database is achieved by using JQueries.

The processing stage that our system uses and the data transmission to the database are implemented using Python 2.7 as a programming language.

Figure 3 depicts an abstract view of our architecture. A patient, who has access in the OpenMRS platform, uploads his Electrocardiogram file or other health information on the platform. Apart from the patient, medical staff such as nurses can upload ECG files using a computer or a mobile device. Finally, ECG file can be transmitted directly to the platform by the electrocardiograph using Internet connectivity. The platform is installed into a server and there is access via mobile phones through an application and computers via a web browser. In server side, WFDB applications and algorithms are implemented to analyze the ECG file and export the necessary results that a doctor must check. When processing is accomplished, the doctor or the nurse can view the results whenever is needed.

For ECG upload, the user should store the ECG file on his personal device. The ECG file is a binary 12—lead WFDB signal file, in SCP-ECG format. It contains a header file, which contains information about the signal and is necessary for the functionality of the WFDB applications. User can also upload a compressed folder that includes both files. The compressed folder is checked for non-existence of files. In case missed files or a file compatibility problems, the system halts and

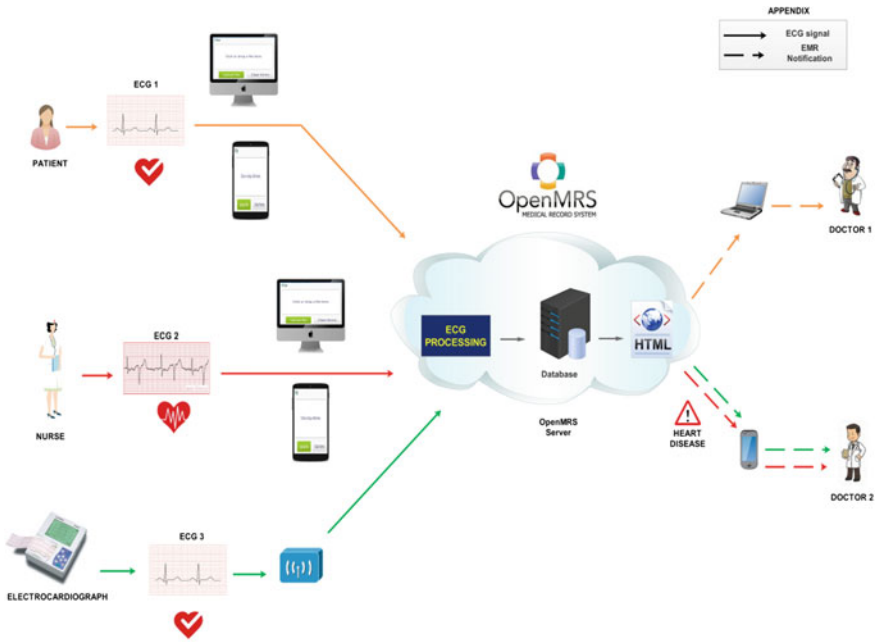


Fig. 3 System architecture

warns the user for improper input data. User can also upload different files, such as images, that are stored in the database and can be viewed.

A module that allows upload file upload is installed into the platform. User navigates to the specific field and selects the compressed folder for uploading. Once the folder is loaded, WFDB applications and algorithms are executed. The results are stored in the database and can be displayed by other users.

Different technologies are used to upload the ECG signal file, produce the ECG results and display them into the platform. Through an OpenMRS module called Visit Document Module, a user can upload a file that is stored into the database. Using Java, the system checks whether the file is a WFDB signal file. Then, Python scripts are executed via Java for signal file processing using WFDB applications. When the execution of the applications ends, we use functions in python to find the rest values. When the operation is completed, we transmit the results to the database. For this purpose, we use SQL Queries and save them into table form in the database. After this process, doctors can study health information via a web browser. We implemented appropriate HTML scripts that visualize this information for better understanding. The graphical environment of our application will be introduced in the next session.

The OpenMRS platform provides the opportunity to create an HTML Form. This form can be written in HTML and can use capabilities that HTML uses such as CSS, JavaScript and JQuery [29].

3.2 System Design

In the current implementation, OpenMRS is the main component of our system. In conventional research approaches, OpenMRS is used only for storing medical data into the platform in image format. The main steps of previous research are depicted in Fig. 4. Medical data such as electrocardiogram are exported via health monitoring devices (electrocardiograph). This information is uploaded by a doctor or medical staff to the OpenMRS platform as images by using a computer or a mobile device. This platform collects medical data for all patients and doctors can access this information using also a computer or a smartphone in order to form the diagnosis. The diagnosis relies on empirical findings by observing the visualized results that OpenMRS provides. OpenMRS supports all types of files but it provides visualization capability only in cases of image uploaded data. In case that the diagnosis indicates a health misoperation, doctor should provide the appropriate treatment to the patient.

This approach allows doctors to study medical exams remotely using hardware but the diagnosis is based on observing the results of the tests. In this case, constant connection to the OpenMRS platform is required for the doctors. Also, doctors should periodically check the recent upload files by the patients and this is not practical in regular basis due to the workload that is assigned to the doctor.

In the current implemented system, we try to extend OpenMRS functionality by expanding its duties. Apart from the previously described operation, OpenMRS analyses the input signals (ECG waveforms) that are uploaded by the medical staff. The analysis process is based on algorithms that use ECG background. The outcomes from this process are used as indicators by our system in order to create the diagnosis for the patient. In case that our system detects health anomalies, it uses appropriate notation to inform the doctor for instant treatment. Additionally, our

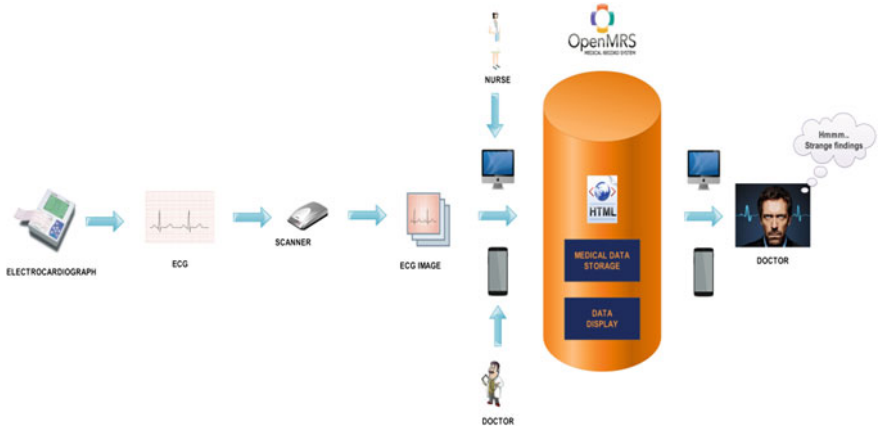


Fig. 4 Conventional OpenMRS functionality

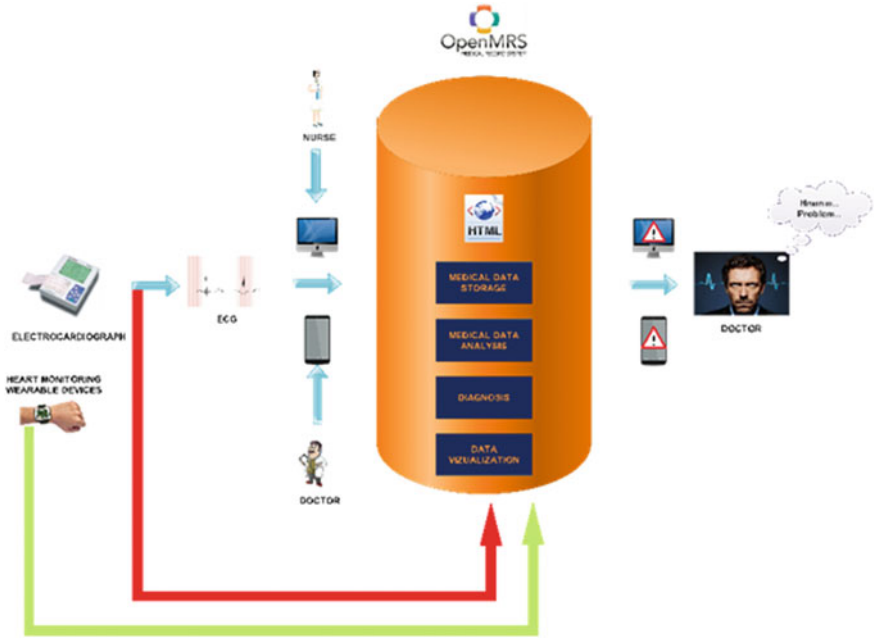


Fig. 5 Proposed schema

system allows direct connectivity with the health monitoring devices without requiring any medical staff. This approach tries to embed the workload for the diagnosis process form the doctor to OpenMRS and generally decrease the degree of human intervention in the whole operation. The overview of the proposed schema is contained in Fig. 5. Also, the system allows real time measurements transmission from wearable health monitoring devices using TCP/IP protocol. Finally, the system receives measurements in binary form for the analysis, diagnosis and visualization process on contrast with previous implementations that display only uploaded data images.

OpenMRS is the EMR system that is used for the implementation. OpenMRS is free, expandable and all his components are open source. It extends faster and it is used in many countries because it has positive effect. Furthermore, OpenMRS satisfies functionalities such as patient registration and retrieval, clinical notes in the system and secure drug prescription. In addition, there are no specific hardware requirements [23] and a dedicated implementers Wiki where the members direct anyone new developer or implementer [30].

OpenMRS contains a database for big data storage and a user-friendly interface [22]. Furthermore, it is compatible with mobile devices such as the Sana

Technology Platform, which is a mobile Health system, integrated into OpenMRS. Sana supports the connectivity with mobile device, which allows users to send medical data that the doctors can view, make a diagnosis and add a medication [19] using a simple mobile phone.

The ECG signal is extracted from the PTB Diagnostic ECG Database in Physionet. PTB database contains a compilation of 15 measured digitized ECG signals, measured by typical 12 leads together with the 3 Frank lead ECG signals. Healthy control ECG signals are available but also pathological ECG signals such as myocardial infarction and heart failure [31].

4 Implementation

System operation consists of three steps. The first step describes the ECG file upload. After this step, the analysis of the ECG signal takes place and the connection to database for result transmission and the final step contains the result visualization to the user.

4.1 Uploading ECG File

For uploading the ECG file, a module that is called Visit Documents Module is used. A user is navigated into the platform and uploads a compressed folder to the corresponding field. The compressed folder contains the ECG signal in SCP-ECG standard format. Figure 6 depicts a screenshot of the ECG upload process in OpenMRS platform.

The ECG upload process in OpenMRS platform can be provided via a mobile device as Fig. 7 shows.

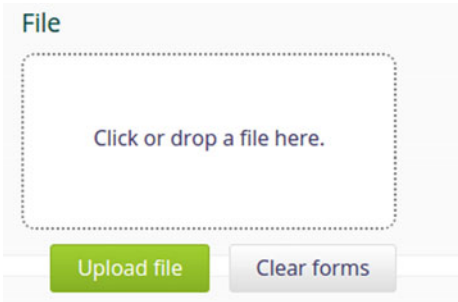


Fig. 6 Upload dialog box

Fig. 7 ECG upload process via mobile device

4.2 ECG Analysis

4.2.1 ECG Measurements

The electrocardiogram is imprinted on a graph paper, which makes it easier to measure the height of each lead and the duration of the waves. The heights are the distances between amplitudes of the peaks, positive or negative, and a baseline. The graph paper is divided into dark lines that have 5 mm distance between them and lighter lines that have 1 mm distance. The horizontal axis displays the time that is 1 mm = 0.04 s = 40 m sec. The vertical axis displays the amplitude of the waves that is 1 mm = 100 μ V = 0.1 mV [32].

Based on Fig. 8, the variables are defined as a = P-wave amplitude, b = Q-wave amplitude, c = R-wave amplitude, d = S-wave amplitude and e = T-wave amplitude.

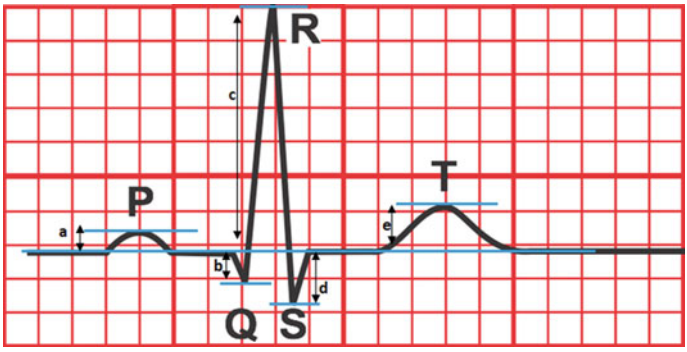


Fig. 8 Amplitudes in a heartbeat

For successful beat measurement, complete P-wave, QRS complex and T-wave are required. Otherwise, next beat is selected for ECG.

A typical heartbeat contains one P-wave, one QRS complex and one T-wave. The excitation wave may be change by some heart disease and that affects the QRS complex. A disease may cause the absence of the Q-wave or the amplitude of the Q-wave is greater than S-wave. Moreover, an abnormality appears when the R-wave is missing. In addition, the S-wave may be not displayed. These are the major indicators of abnormal heart operation.

4.2.2 ECG Findings

Examining an electrocardiogram provides some findings that are necessary for the human health. First, we calculate the heart rate using R-R interval. The formula to calculate the heart rate is given by the following equation

$$\text{bpm} = \frac{60}{\text{R} - \text{R interval}} \tag{1}$$

where constant 60 is used for minute-to-seconds conversion and R-R interval represents values in seconds. The range of the heart rate is between 60 beats per minute (bpm) till 100 bpm. If the heart rate is below 60 bpm, this is a bradycardia symptom. If the heart rate is above 100, possibly tachycardia [33] is decided.

The intervals in a heartbeat are also crucial. P-wave's duration should be less than 120 ms. QRS complex in a normal heartbeat lasts for about 70 ms up to 110 ms. T-wave holds for about 300 ms after the QRS complex and lasts about 160 ms. P-R interval lasts about 120 ms till 200 ms and Q-T interval lasts from 350 to 440 ms [34]. Using Q-T interval, we calculate the QTc, which is a corrected Q-T evaluation. QTc depends on the R-R interval and we calculate it using the Bazett's formula [35]

$$\text{QTc} = \frac{QT}{\sqrt{R - R\,interval}} \tag{2}$$

Some extra values that should be estimated are the amplitude of the waves in the ECG. The amplitude values are calculated as the distance between the top of each wave till the baseline. The amplitude of P wave is up to 2.5 mm in Limp Leads (I, II, III, AVR, AVL and AVF) and 1.5 mm in Precordial Leads (V1, V2, V3, V4, V5 and V6) [36]. The amplitude of the QRS complex is the largest of the other waves and contains the Q-wave, the R-wave and the S-wave. Table 1 shows the values that each waveform can take [37, 38].

Using the evaluated amplitude values of the QRS waveforms, we calculate the R/S ratio. Given the R amplitude values in leads V1, V2 and V6 and the S amplitudes in the V1, V2 and V6, we estimate the appropriate ratios [39, 40] shown in Table 2.

Table 1 Amplitudes of ECG waves

P height II	≤ 2.5 mm
R amplitude I	15 mm < R ≤ 20 mm
R amplitude II	≤ 20 mm
R amplitude III	≤ 20 mm
R amplitude AVF	≤ 20 mm
R amplitude AVL	≤ 13 mm
R amplitude AVR	≤ 3 mm
R amplitude V1	≤ 26 mm
R amplitude V2	≤ 26 mm
R amplitude V5	≤ 27 mm
R amplitude V6	≤ 27 mm
S amplitude I	≤ 8 mm
S amplitude V1	≤ 30 mm
S amplitude V2	≤ 30 mm
S amplitude V5	≤ 17 mm
S amplitude V6	≤ 4 mm

Table 2 R/S ratio

V1 Ratio: R-V1/S-V1	≥ 1
V2 Ratio: R-V2/S-V2	≥ 1.5
V6 Ratio: R-V6/S-V6	≤ 3

Finally, the axes of the heart that describe the direction of the electrical waves play critical role in accurate diagnosis. The most important axis is the QRS Axis, which shows the direction of electrical current propagation through the myocardium. The normal QRS Axis ranges from -30^{o} to $+90^{o}$.

There is also the P Axis, which represents the direction of P wave and the T Axis that represents the T wave direction. The P Axis ranges from 0^{o} to 75^{o}. Using QRS Axis and T Axis, we estimate the QRS-Tangle. QRS Tangle is the angle between QRS axis and T axis and is between 20^{o} and 116^{o} for females and between 30^{o} and 130^{o} for males [41].

4.2.3 ECG Signal Processing

In this section we describe the ECG analysis process that takes place in our system. For this reason, WFDB Applications along with some our algorithms are used for calculating the findings described in the previous section. The processing of the ECG file starts after the upload phase is completed. Figure 9 depicts the steps of ECG processing. The ECG signal passes through filtering phase and then ECG analysis takes place. The results of this analysis are stored into the database. Lastly, the doctor can view the ECG signal in the form of a web page via the OpenMRS platform and make a diagnosis.

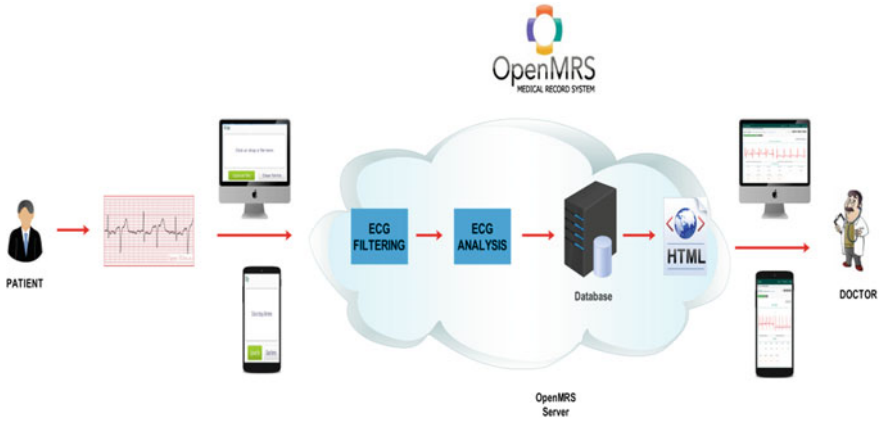


Fig. 9 ECG processing

The uploaded file is a compressed folder containing a binary file. Initially the binary file has to be read. For this purpose, we use the rdsamp application. Rdsamp converts the binary file into decimal numbers. The text file contains thirteen columns separated by tabs. The first column contains the elapsed time of the ECG and the remaining twelve columns contain the ECG leads (I, II, III, AVR, AVL, AVF, V1, V2, V3, V4, V5 and V6) as decimal numbers denoting voltage in mVolts.

Subsequently, FIR application is enabled. FIR is used to apply a finite impulse response filter. The filter that is used in the current work is a low-pass boxcar filter or rectangular window filter, whose impulse response has non-zero values and has finite duration in a "window" containing N samples. This filter attenuates the high frequencies and reduces the noise that an ECG signal may have [42]. The data that are used and visualized in the platform pass through this filtering process.

After signal filtering, sqrs application is used to determine QRS complex values in the ECG signal. After this step ann2rr application determines the R-R intervals in the ECG signal. Also, Ihr uses the same information for instantaneous heart rate calculation. Finally, averaging the latter the average ECG heart rate is estimated.

The last processing step is the ECG wave definition in the ECG signal. By using ecgpuwave application in each lead, we evaluate the appropriate interval spaces per lead. For this process, rdann and ecgpuwave are incorporated. The results of these WFDB applications help us to calculate the ECG findings. Firstly, we should determine the baseline. The samples of the ECG signals are not evenly distributed around the axes. Some of them are above the baseline and the waveforms that should be below the baseline, are above, like Q-peaks and S-peaks. To avoid that issue we determine the area defined as zero potential in an ECG signal [43]. That area is between the end of T-wave and the start of P-wave so we calculate the average of the amplitudes in that interval and subtract it from each sample from the entire signal.

Using the intervals of the wave we have determined, we calculate the corresponding heights by approximating the Q, R, S peaks (maximum and minimum values) in each interval. After high estimation we proceed to axes definition by classic geometrical methods. Next section illustrates the outcomes of this processing.

4.3 Resulted HTML Pages in OpenMRS with ECG Data

OpenMRS provides a web application where a doctor has access and can view every patient's dashboard as Fig. 10 depicts. Patient's dashboard includes personal information about patients such as their address and medical data, such as information about vitals and allergies that are appropriate for doctor to make a diagnosis.

We can view Patient's dashboard also via a mobile device as Fig. 11 shows.

Before medical records are inserted in a patient's folder, a doctor should initiate a virtual "Visit". A Visit event takes place when a patient shortly interacts with the system in a specific location. This represents the conventional visit by the doctor via

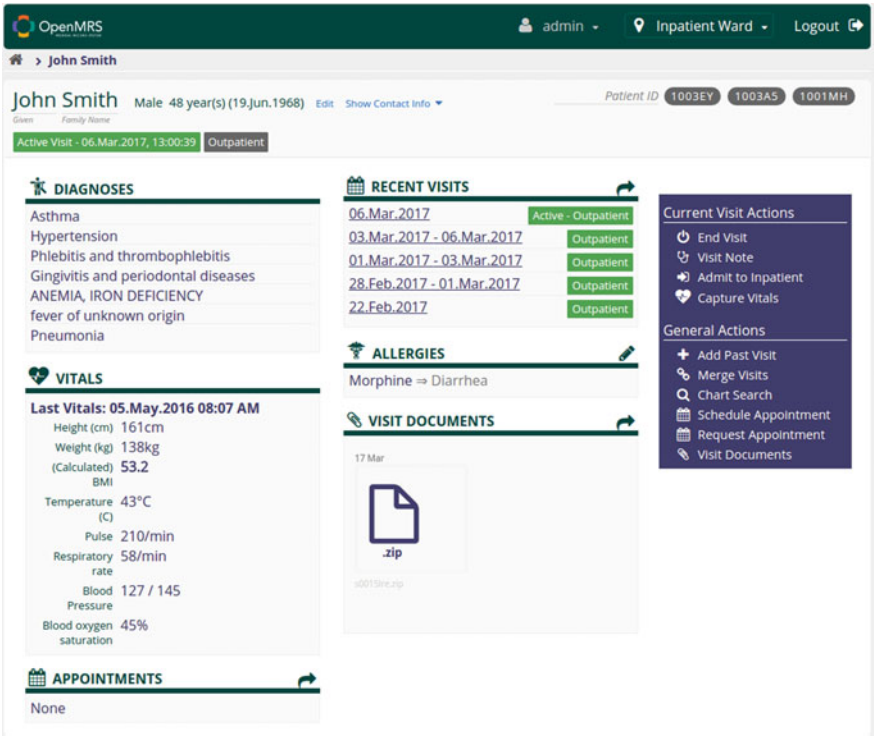


Fig. 10 Patient's dashboard

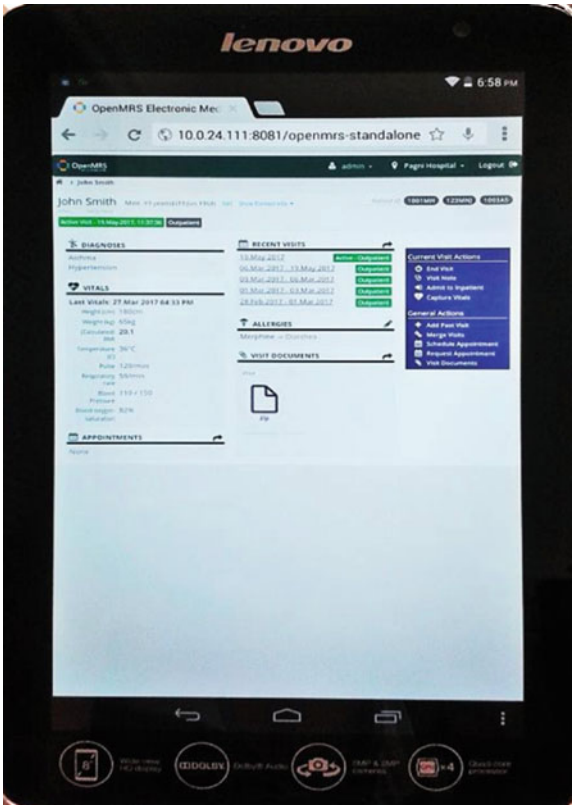


Fig. 11 Patient's dashboard in a mobile device

physical presence in a hospital. One patient checks in at the clinic and starts his Visit until the checkout time which ends the Visit. The Visit contains updated encounters that are added to the patient's record from each doctor the patient visits and are categorized according to the date. The system also keeps a log of previous Visits in the patients' dashboard.

The data storage process in our system follows the hierarchy applied in OpenMRS platform The OpenMRS model is divided in levels. The first level contains the Visit. Each visit is described by a unique visit_id. Data that are collected every time a patient visits a doctor via a Visit are called "encounters" and are part of visits. A Visit can contain multiple encounters with different encounter_ids. Encounters consist of observations. Observation is regarded as a single unit of clinical information that is imported for a patient characterized by an obs_id. Each Observation includes Concepts that are any data we want to store for a patient. Concepts have person_id as key value. Figure 12 depicts the information structure in OpenMRS.

For the current implementation, we created an encounter called Electrocardiogram (Fig. 13). A visit may include more than once the same encounter and the number of occurrences each encounter is met depends on the frequency that the

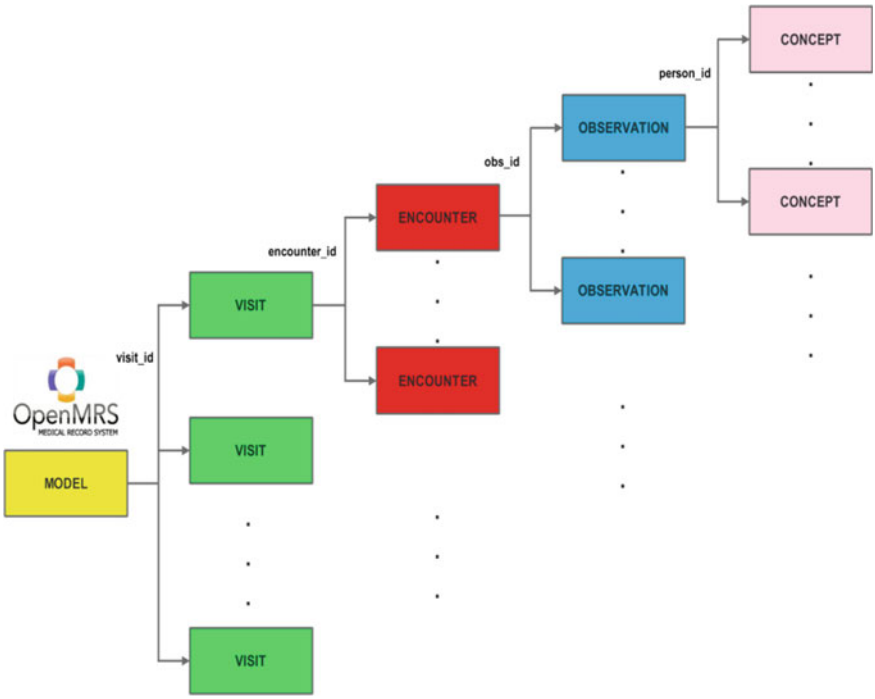


Fig. 12 Medical information structure

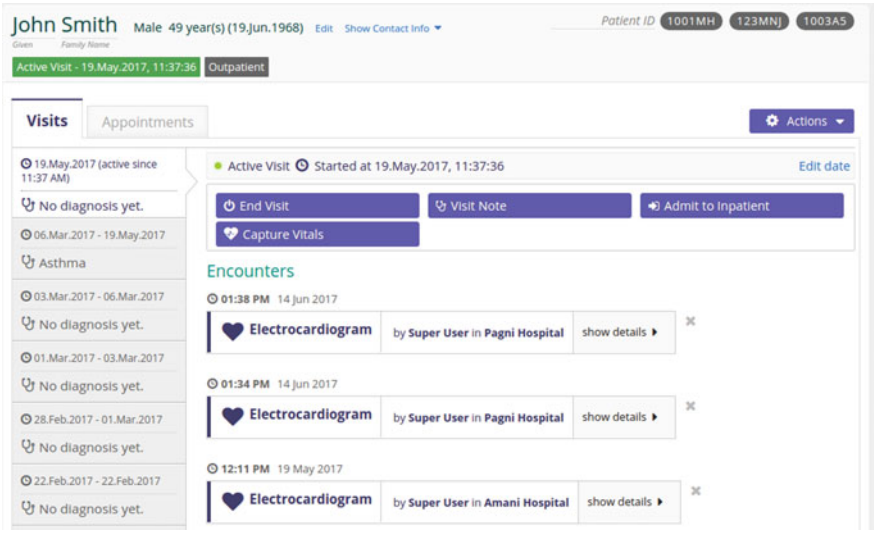


Fig. 13 Patient's encounters

Fig. 14 ECG visualization

encounter is repeated during the Visit. The Electrocardiogram encounter contains several Observations. Each Observation is a Concept that has been created for storing each medical processed value in the database. To be more specific, Observations and Concepts have been created for all values in Tables 1 and 2. The encounters can be displayed or deleted if the doctor selects the corresponding fields.

Data are visualized using a form that OpenMRS provides. After the data collection, we use HTML Form Entry module to display them. The HTML Form Entry module allows users to create HTML forms and insert data into the platform. We can use HTML for developing but also JavaScript and CSS. Each value is stored into the database with different identified numbers (ids). Combining id and jQuery allows displaying these values. Before visualization, in the HTML we check if the

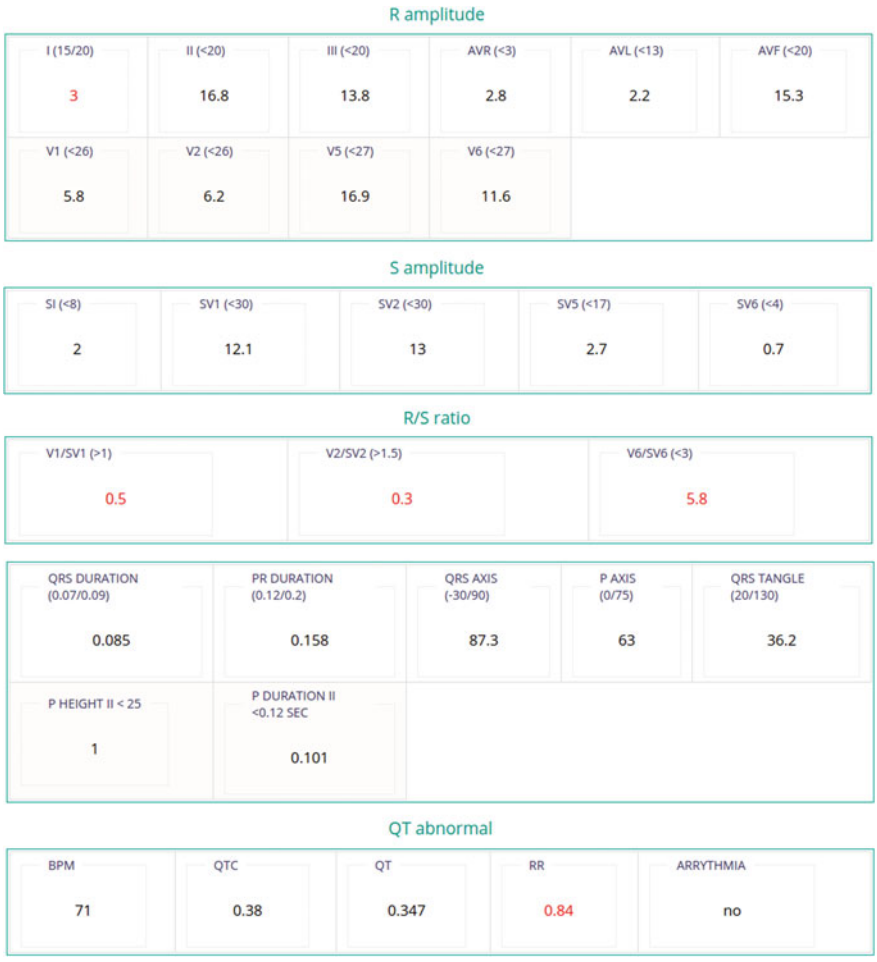


Fig. 15 ECG metrics and warnings

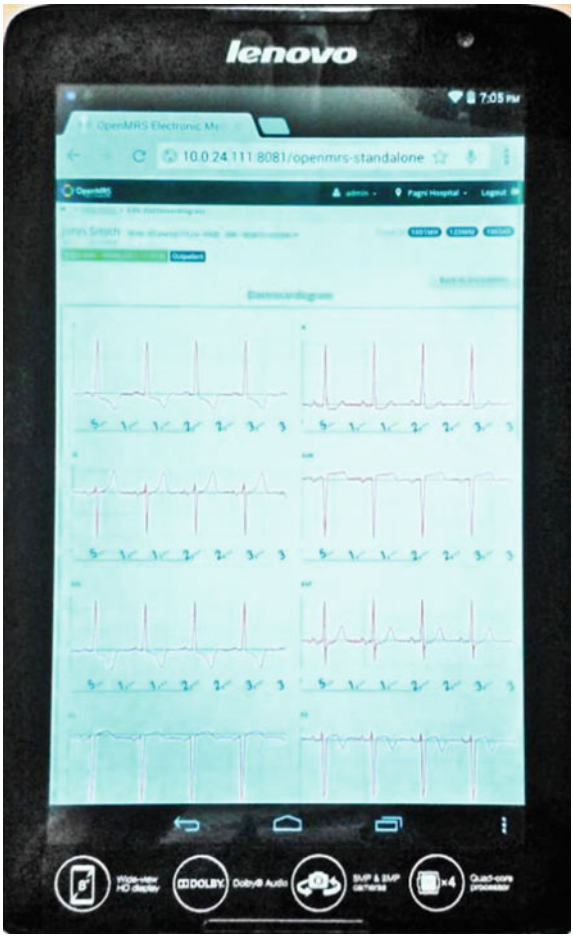


Fig. 16 ECG visualization in a mobile device

values are normal or not. We compare them with the corresponding limits. If the values are abnormal, appropriate warnings in red color inform doctors.

Figure 14 shows an example of ECG visualization results exported into OpenMRS platform. Each plot describes the ECG signal in 12-leads (I, II, III, AVR, AVL, AVF, V1, V2, V3, V4, V5 and V6).

In Fig. 15, we study an example of abnormal heart operation. The indicators in red color (R1 amplitude, V1 ratio, V2 ratio, V6 ratio and RR interval) are the warnings shown to the doctor indicating that patient needs treatment.

ECG visualization is also feasible via a mobile device as Fig. 16 shows.

5 Conclusions—Future Work

Due to the technological development during recent decades, new achievements are enabled into medical research area. IoT and computationally powerful mobile devices are assigned to execute operations that initially belonged to doctor's responsibilities. This evolution amplifies the role of healthcare in recent medical system. Using healthcare application allows doctors to save time and also have access to patients' health data without requiring physical presence into hospitals.

In this current work, we focus on a healthcare system related with heart diseases. In more details, we describe a system that uses electrocardiogram (ECG) from a patient as input. Our system analyzes this signal and evaluates some metrics that help for an accurate diagnosis against heart diseases. Apart from the processing part, our system transmits this information to a centralized database using an open source platform called OpenMRS and makes it accessible to the doctor. He can have instant access to the data and examine the history of each patient. This model provides dynamic access to health data using a single laptop or a mobile device improving diagnosis conditions. Additionally, our implemented architecture virtualizes the data and also in emergency cases it informs the doctor with appropriate warning messages for instant treatment.

The implemented system uses open source software so this allows the research community to extend its capabilities and also to adjust to any possible packages. Furthermore, open source software has zero cost so our application amplifies the costless profile of the health system. Based on the financial circumstances that exist in many countries around the world, such applications can prove to be valuable in human life prevent people from dangerous health problems due to lack of hospital staff or equipment. It aims to the direction of a shared health system that supports all humanity.

Our system consists of open source software. For this reason, it can incorporate any updated package distributions that will extend its functionalities. We use OpenMRS platform, which allows the storage and manipulation of the medical data and it does not have any constraints against other software. Moreover, OpenMRS has modular structure, which means that it can increase its functionalities by embedding the proper module.

Apart from the additional packages that can be incorporated in OpenMRS platform, our system can accept any possible biomedical signal as input for storage or processing. This means that except from electrocardiogram signal it accepts other signals representing health indicators such as temperature, pressure, oxygen consumption etc. This information can be kept in the database and be accessible to the doctor instantly. Following this direction, our system can be used for monitoring patients for multiple health indicators simultaneously reducing health cost and time spent in hospitals.

Additionally, the exponential growth of IoT industry and the evolution of wearable devices or Body Sensor Networks (BSNs) have the potential to lead to a next generation of health system using wireless technology. This is practical for

older people that stay home for a long period without a specialist for treatment. It also limits the need for hospitalization for medical tests. This is practical for counties that lack of medical staff or equipment. The wearable sensors can transmit the measurements in real time because they contain transceivers (Bluetooth or ZigBee protocol [44, 45]) and this helps people that are not familiar with technological achievements and mobile devices. Moreover, based on the computational power that recent mobile devices, computers, wireless sensors and healthcare devices have, a possible future extension of our system is to move the workload for ECG processing from our centralized OpenMRS server to the mobile devices that patients use in order to upload the ECG signal. Each patient's device can extract the required information and then transmit it the server. This reduces dramatically the requirements that the server should have and the only duties that it keeps is monitoring this information and informing doctor in emergency cases. This results in a distributed architecture that is depicted in Fig. 17. This approach increases the security and the redundancy of the system because patients can keep locally a copy of their ECG signals before uploading it to the server. In case of a DOS attack to the server, following this method, no data loss will occur.

All the above possible extensions have the potential to increase the flexibility and the effectiveness of the recent health system in order to respect the patients and

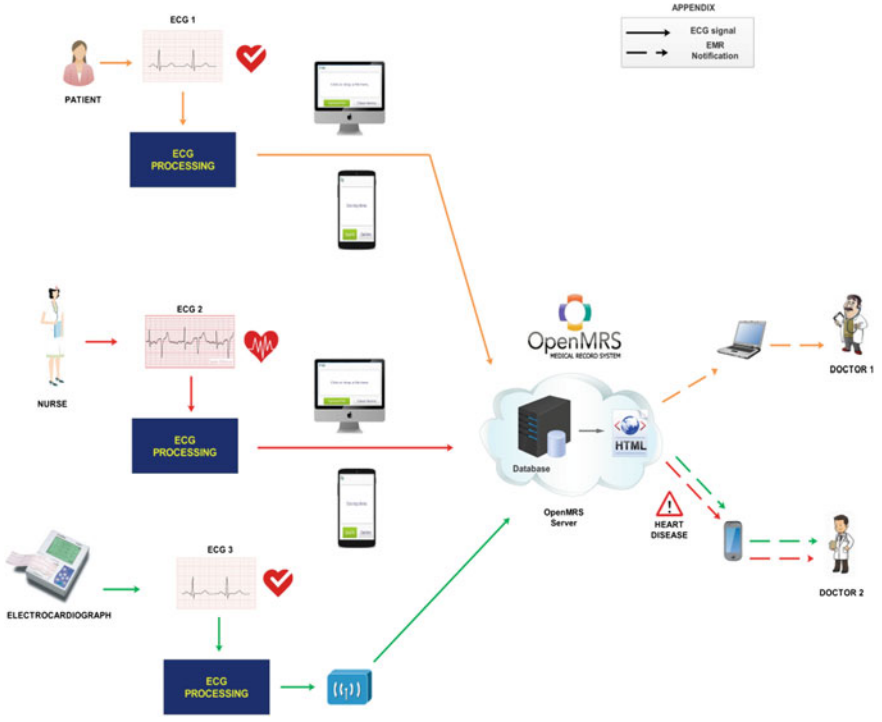


Fig. 17 Distributed architecture

reduce the obstacles that creates in their daily life. It offers the capability of connecting all hospitals by creating a global digital database that uses information by the local databases in each hospital. This is crucial for studying some exceptional cases that are met in medical science and exporting useful conclusions, supporting research in this area. Finally, this idea allows opinion exchange between doctors located in different hospitals around the world using a laptop or a mobile device. This approach saves time and eliminates distance, which are crucial factors that must overcome in health branch.

References

1. Wikipedia: Information technology. https://en.wikipedia.org/wiki/Information_technology (2017)
2. Wikipedia: Health information technology. https://en.wikipedia.org/wiki/Health_information_technology (2017)
3. Wikipedia: Electronic health record. https://en.wikipedia.org/wiki/Electronic_health_record (2017)
4. Wikipedia: Electrocardiography. https://en.wikipedia.org/wiki/Electrocardiography (2017)
5. Ranjan, R., Kołodziej, J., Zomaya, A., Alem, L., Wang, L.: Software tools and techniques for big data computing in healthcare clouds. Future Generation Comp. Syst. **43**, 38–39 (2015)
6. Sahay, S.: Big data and public health: challenges and opportunities for low and middle income countries. Commun. Assoc. Inf. Syst. **39**(20) (2016)
7. Ma, Y., Song, J., Lai, C.F., Hu, B., Chen, M.: Smart clothing: connecting human with cloud and big data for sustainable health monitoring. Mobile Netw. Appl. **21**(5), 825–845 (2016)
8. Warwick-Clark, B., Obeysekare, E., Mehta, K., Bram, J.T.: Utilization and monetization of healthcare data in developing countries. Big Data **3**(2), 59–66 (2015)
9. Madhukant, R., Prabhakaran, V.M., Gokul Kruba Shanker, R., Balamurugan, S.: Internet of health: applying IoT and big data to manage healthcare systems. Int. Res. J. Eng. Technol. (IRJET) **3**(10) (2016)
10. Coronato, A., Amato, A.: An IoT-aware architecture for smart healthcare coaching systems. In: 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), pp. 1027–1034. IEEE (2017)
11. Geetha, G., Sundara Velrani, K.: Sensor based healthcare information system. In: Technological Innovations in ICT for Agriculture and Rural Development (TIAR) 2016, pp. 86–92. IEEE (2016)
12. Laplante, N.L., Laplante, P.A.: A structured approach for describing healthcare applications for the internet of things. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 621–625 (2015)
13. Laplante, N., Laplante, P.A.: The internet of things in healthcare: Potential applications and challenges. IT Prof. **18**(3), 2–4 (2016)
14. Lee Ventola, C.: Mobile devices and apps for health care professionals: uses and benefits. **39** (5) (2014)
15. Saleh, A., Mansour, M.M., Zarka, N.: Mobile healthcare system (2016)
16. Khan, M.A., AlGhamdi, M.A., AlMotiri, S.H.: Mobile health (m-health) system in the context of IoT. In: 2016 4th International Conference on Future Internet of Things and Cloud Workshops, pp. 39–42. Aug 2016
17. Knowledge for Health: mHealthKnowledge. http://www.mhealthknowledge.org/resource-type/applications-platforms (2017)

18. Nimunkar, A.J., Webster, J.G., Kalogriopoulos, N.A., Baran, J.: Electronic medical record systems for developing countries: review. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1730–1733 (2009)
19. Sana: http://dev.sanamobile.org/
20. Sarmenta, L., Rotberg, J., Marcelo, A., Clifford, G., Celi, L.A.: Mobile care (Moca) for remote diagnosis and screening. J. Health Inf. Dev. Countries **3**(1), 17–21 (2009)
21. Vereijken, B., Becker, C., Todd, C., Taraldsen, K., Pijnappels, M., Aminian, K., Mellone, S., Helbostad, J.L.: Mobile health applications to promote active and healthy ageing. Sensors **17** (3), 622 (2017)
22. King, A., Lee, I., MacDonald, A., Fernando, A., Hatcliff, J.: Rationale and architecture principles for medical application platforms. In: ACM/EEE Third International Conference on Cyber-Physical Systems (ICCPS 2012), pp. 3–12. April 2012
23. Bru, J., Berger, C.A., Millard, P.S.: Open-source point-of-care electronic medical records for use in resource-limited settings: systematic review and questionnaire surveys. BMJ Open. **2** (4), e000690 (2012)
24. Haiqi, A., Zaidan, B.B., Zaidan, A.A., Kiah, M.L.M.: Open source EMR software: profiling, insights and hands-on analysis. Comput. Methods Program. Biomed. **117**(2), 360–382 (2014)
25. Ukil, A., Bandyopadhyay, S., Singh, R., Pal, A., Mandana, K., Puri, C.: iCarMa: inexpensive cardiac arrhythmia management–an IoT healthcare analytics solution. In: Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems, pp. 3–8 (2016)
26. Landínez, S.F., López, D.M., Blobel, B., Villamil, C.A.: A mobile ECG system for the evaluation of cardiovascular risk. In: MIE, pp. 210–214. Sep 2016
27. Agrafioti, F., Hatzinakos, D., Plataniotis, K.N., Wang, Y.: Analysis of human electrocardiogram for biometric recognition. EURASIP J. Adv. Signal Process. (2007)
28. Alesanco, A., Martinez, I., Garcia, J., Trigo, J.D.: A review on digital ECG formats and the relationships between them. IEEE Trans. Inf. Technol. Biomed. **16**(3), 432–444 (2012)
29. Arbaugh, J.: HTML form entry JavaScript reference. https://wiki.openmrs.org/display/docs/HTML+Form+Entry+JavaScript+Reference (2014)
30. Mamlin, B.W., Biondich, P.G., Fraser, H.S., Wolfe, B.A., Jazayeri, D., Allen, C., Miranda, J., Baker, E., Musinguzi, N., Kayiwa, D., Fourie, C., Lesh, N., Kanter, A., Yiannoutsos, C.T., Bailey, C., Seebregts, C.J.: The OpenMRS implementers network. Int. J. Med. Inf. **78**(11), 711–720 (2009)
31. The National Institute of Biomedical Imaging and Bioengineering (NIBIB) National Institute of General Medical Sciences (NIGMS): PhysioNet the research resource for complex physiologic signals. https://physionet.org/
32. Hudson, K.B., Naples, R., Sudhir, A., Mitchell, S.H., Ferguson, J.D., Reiser, R.C., Brady, W. J.: The ECG in Prehospital Emergency Care. Wiley (2012)
33. Papazaxos, G.: The Electrocardiogram in Clinical Practice. Medical Publications of Litsas (2000)
34. van Herpen, G., Bots, M.L., Verweij, N., Rijnbeek, P.R.: Normal values of the electrocardiogram for ages 16–90 years. J. Electrocardiol. **47**(6), 914–921 (2014)
35. Wikipedia: QT interval. https://en.wikipedia.org/wiki/QT_interval (2017)
36. ECGpedia: P wave morphology. http://en.ecgpedia.org/wiki/P_Wave_Morphology (2011)
37. Bove, D.W., Norris, K.E., Conyers, R.J., Conradi, E., Rowlands, S., Scott, D.T., Romhilt, R. C.: A critical appraisal of the electrocardiographic criteria for the diagnosis of left ventricular hypertrophy. Circulation **40**(2), 185–196 (1969)
38. Keys, A., Simonson, E., Rautaharju, P., Punsar, S., Blackburn, H.: The electrocardiogram in population studies. Circulation **21**(6), 1160–1175 (1960)
39. Zhang, Z.M., Crow, R.S., Prineas, R.J.: The Minnesota Code Manual of Electrocardiographic Findings. Springer Science and Business Media (2010)
40. Park, R.E., Marchlinski, F.E., Hutchinson, M.D., Garcia, F.C., Dixit, S., Callans, D.J., Cooper, J.M., Bala, R., Lin, D., Riley, M.P., Gerstenfeld, E.P., Betensky, B.P.: The V2

transition ratio: a new electrocardiographic criterion for distinguishing left from right ventricular outflow tract tachycardia origin. J. Am. Coll. Cardiol. **57**(22), 2255–2262 (2011)
41. Feldman, T., Henrikson, C.A., Tereshchenko, L.G., Oehler, A.: QRS-T angle: a review. Ann. Noninvasive Electrocardiol. **19**(6), 534–542 (2014)
42. Kamath, U., Bharadwaj, A.: EE times connecting the global electronics community. [Online]. http://www.eetimes.com/document.asp?doc_id=1278571 (2011)
43. Pucik, J., Cocherová, E., Ondracek, O.: Filters for ECG digital signal processing. Int. Conf. Trends Biomed. Eng. **7**(9) (2005)
44. Memane, K., Londhe, T., Thanki, H.J., More, P.: Advance IoT-based BSN healthcare system for emergency response of patient with continuous monitoring and motion detection. Int. J. Modern Trends Sci. Technol. **2**(12) (2016)
45. Bhattacharya, P.P., Sangwa, A.: Wireless body sensor networks: a review. Int. J. Hybrid Inf. Technol. **8**(9), 105–120 (2015)

Social Networking in Higher Education in India

Anil Kumar Malleshappa and Tomayess Issa

Abstract Innovations in technology are remarkable as they enable people to be reached almost immediately. This is the beauty and uniqueness of technology as it makes visual communications so easy. The emergence of the Internet is the main reason for further innovations in technology since the world is being interconnected with the web. The web technologies have made the world connect more closely and enabled people to interact by sharing experiences and information. Social networks are integrated with the cognitive aspects of human beings. It makes the technology even better by combining the hardware of the computer with the feelings of human beings. Social networks have grown like never before but are they acceptable in all the sectors? Well, each stream has its own perspective concerning social networks. Higher education is one of those streams, which uses social networking in huge numbers. Educational institutions have been in the process of providing a career path to its students through guidance and recognizing their skills. Social networks would help the education providers by making this much easier by providing platform enabling them to know more about their students. The primary objective of this research is to examine the advantages and disadvantages to students that arise when using social networks and to validate the importance of social networking in education. This research intends to list the advantages that Indian students and educational institutions can take advantage of in the academic environment via using social networking, and to list the disadvantages to help the policy makers and students to be well aware of the negative situations that they might encounter by using social networking in the Indian higher education. The research uses a quantitative method, in this case an online survey, as its main approach. 103 participants completed the online survey from 175 participants. Finally, this study was limited to India higher education, further research will be carried out in the future to examine other higher education from developed and developing countries.

A.K. Malleshappa · T. Issa (✉)
Curtin University, Perth, WA, Australia
e-mail: Tomayess.Issa@cbs.curtin.edu.au

A.K. Malleshappa
e-mail: anilkmrr020@gmail.com

G. Skourletopoulos et al. (eds.), *Mobile Big Data*, Lecture Notes on Data Engineering and Communications Technologies 10,
https://doi.org/10.1007/978-3-319-67925-9_13

Keywords Social networking · India · Advantages and disadvantages

1 Introduction

In recent years, we have seen a revolution in technology, which has made the life simpler in terms of communicating, information sharing and interacting. The web 2.0 has helped in bonding the Internet and web that has taken the development of online tools to the next level. Organizations consider it as a powerful tool which they make use of it to communicate, collaborate and share information not only internally, but also externally. Web 2.0 has been used in various streams; however, the term is unclear because of its adaptable nature of different disciplines. Web 2.0 tools have made communication quicker and more efficient compared to traditional technologies such as telephone. There are various online social tools used for different purposes that help people to make contact and stay in contact with colleagues and friends. They also help organizations to share information and conduct business. Organizations have started using social networking to build their relationship with the employee to improve the social interaction within the professional environment.

The use of social networking provides advantages to the education sector. With the growing numbers of learners and scholars moving to online information sharing, the Internet has become a knowledge management tool. There are significant debates and research going on to study the impact of online tools on education and learning. The universities have already adopted these tools and are encouraging students to participate in wikis and forums. Blogging has become a discussion tool. The classroom study has been making use of wikis to collaborate and as a tool to present the work done by a group. Online social networking tools have become a part of academics integrating collaboration, communication and shared information. The universities have set up individual applications to make courses easier.

The innovations on the Internet have brought major changes to academia; the tools now used for teaching, collaborating and sharing information have presented new challenges and advantages. This research study is an attempt to forecast the impact of online tools on teaching and learning which also forecasts the impacts due to social networks. It also exhibits the perceptions and attitudes of students and teachers. The study aims to examine the barriers and advantages arising from the adoption of social networks in India.

Technology plays a major role in providing a platform to share information in different ways, sometimes the means of communication is just as important as the content of the information. The digital technology provides a hassle free environment in order to communicate which is made much better by wireless technologies. The sharing of information is much easier than ever before, but we need to be aware of the disadvantages involved. The Internet is been accessed by various means, but the technology to access the Internet is very much the same. Modern technologies have followed cognitive principles in order to make them more ‘human’ in nature.

The education that has been provided in the classroom for many years has had positive effects on a student's career. It is time for us to think about building a student's career not only within the school, but also as a part of his life. In order to achieve this, social networking might help in a positive way. During recent years, we have seen the positive response from different communities accepting social networking for different reasons. The universities have not been left behind as they are using social networking and are in a position to understand the thoughts of the student through collaboration. Social networks provide resources from different sources and are a unique means of providing the details of the information stored. There are many social networking websites serving different needs; LinkedIn is a website, which focuses on employment advantages. The growth of social networking has been quicker than the economic growth of the country in terms of serving people. Many advantages are provided to the student who is using social networking, which we believe, is highly appreciable. However, despite the many opportunities provided by social networks, there are a few challenges that a student might face.

A student has to undergo various challenges in his/her student life and has to be flexible in academic as well as other activities. When a student uses the Internet, there is a lot of material that might distract a student; however, a strong commitment to study can dispel this distraction. In our belief, the student will meet various challenges when using social networking. The student has to create a personal profile and has to be aware that others can misuse it for their own benefit. Privacy concerns may also be alarming in the event of information being misused. Trust is a factor when sharing information; the direct recipient of the communication may not reveal information, but the people with bad intentions may attempt to exploit it. There are various countries, including India, that promote social networking in education despite the associated disadvantages. India is well known for its historical background.

The cultural aspects of India are unique because there are various cultural traditions within this one country. India's innovations in technology and contributions to the world are remarkable. In terms of education, the Indian education system follows rigid examination procedures whereby students sit for exams conducted by the educational institutions, many of which are not internationally recognized.

This chapter focuses mainly on the advantages and disadvantages a student can expect when using social networking. The main target of the study is Indian higher education system and the research might help to reveal the factors that could provide solutions to the problems faced.

2 Social Networking

The technology has been consistently changing to meet the requirements of information sharing. The educational system in India that has been followed for many years needs a makeover and has to be updated to meet the requirements of the

students. Social networks have been very useful as a means to improve the knowledge and encourage social interaction. The web technologies are used by many mainstream enterprise systems serving the business as well as the consumers. The user interfaces created during the development of web technologies are amazing and are developed in line with cognitive characteristics, approximating the interactions that take place in society. It has advantages and as well as limitations, so students need to be carefully guided so that they can make the best use of social networks. Online social networking has been improved by web 2.0. It has become a very powerful tool in changing the behavior of the student. Social networking not only can establish a relationship with peers, but can also create a community of people with similar interests. There are situations a student may want to make use of part of it is to his/her academics.

This chapter aims to gather the necessary resources and discusses the evolution of the Internet. The Internet as a web has been used to create many sophisticated online tools one of which is social networking, so it is important to know how the Internet has evolved which led to the development of online social networks. Social networking and various other factors are discussed, followed by a discussion of higher education, and then it shifts to the Indian context. The technology used to communicate; process and exchange information is called 'communication technology'. The early communication methods take us to the invention of spoken language and pictographs on walls in caves, then ultimately lead to electronic communication that can convey information quickly. Newspapers and television were the early electronic communication technologies which reached a large number of people and was popularly known as mass media, not forgetting radio technology which is widely used even today [1].

The invention of telegraph, television, radio, and newspaper helped people to obtain information quickly; similar to technologies, social networking has also evolved effectively. Usernets were the first of the social media technologies used for interaction whereby the users could post to newsgroups and publish articles. Even today, sites like Facebook and Google refer to usernets for the evolution of social networks [2]. Electronic bulletin boards (bboard) are also one of those media that encouraged social interaction. The 'bboards' enabled users to post messages and could be viewed by anyone who had access. It was not used to convey private messages but generally served a wide proportion of members of a community or group so that information is shared only among them [3].

The term 'web' is derived from World Wide Web, and the web plays a most important role in acquiring the required information with the help of the technology. It is obvious that today's online tools, including social networking applications, are based on the web technology. It is important for us to know about these web technologies, so this section provides a brief description of web 1.0, web 2.0 and web 3.0 and their role in developing web applications [4]. Web 1.0 is considered the very first generation of web that served a wide range of communities. It served as a database where the users could search for the information. It also utilised cognitive principles but did not enable interactions. Web 1.0 was used for business purposes as well where the organizations could broadcast information and users

would search for it, but the main feature of this tool was the read-only web. The contribution to the business from this tool was appreciable where the brochures and catalogues were created in the web by an organization, and users located and contacted the organization for the products. The main goal of this tool was to disseminate information to anyone anywhere with the implementation of technology. Even though it served a purpose very similar to that of newspapers or magazines, the organizations took advantage of the graphic user interface to attract the users. The HTML web pages created using this tool were not frequently updated, and were not collaborative. HTTP, HTML and URI are the core applications used in this web [4].

The web 1.0 is a read-only web, whereas the web 2.0 is a read-write web. It is bi-directional where users are able to read and write, enabling them to participate more fully in the interaction. The web 2.0 is considered to have revolutionised business, making the Internet a platform for business purposes. The web 2.0 became increasingly important and became very popular in no time. Users liked the flexibility and collaborative content creation, which was not possible with web 1.0. Web 2.0 signalled the era of applications and collaboration. Many applications were created to serve the community. The services of web 2.0 included RSS feeds, blogs, Wikis, Mashups, social networking, tags and tag clouds, to name just a few of the important technologies served by web 2.0. The development tools also started gaining in importance during this time, as tools such as Google Web Kit, XML, and Flex were used to create the applications.

In 2006, John Markoff suggested the third generation of web. Web 1.0 and 2.0 had served the purposes of business, social interaction, and learning. The technical gurus believe that it is time for web 3.0 and have called web 3.0 the 'intelligent web' saying that it would support the features of web 2.0. The main aim of web 3.0 is to enhance the software and hardware tools used in web 2.0. Web 3.0 concentrates more on security, connectivity and on incorporating more and more of the cloud technologies [4]. The components of the semantic web are the main constituents of Web 3.0. RDF language, which is used in database management, is also increasing in importance because it allows us to store additional data and helps in connecting data with web patterns, which is not available in the present databases used. It helps us to limit the online applications according to the relevant requirements of the user [5].

Web 1.0 and web 2.0 have changed the way of searching for information and added the cognitive functionalities to the business perspective. It is important for us to know the evolution of web 2.0 from web 1.0 to show how web 2.0 was evolved. Web 3.0 is a more advanced way of using the web with the involvement of smart web. These web applications have served various sectors of business, in addition to changing the means of interaction in educational institutions. The following Table 1 shows the differences between web 1.0, 2.0 and 3.0.

The effective mechanisms made available by web 2.0 technologies have enhanced social networking. There are various reasons for the use of social networking by students. Pempek et al. [6] Conducted a survey in the US to discover the reasons for usage of social networking by college students and found that

Table 1 Difference between web 1.0, 2.0 and 3.0—prepared by authors

Features	Web 1.0	Web 2.0	Web 3.0
Communication	Broadcast	Collaboration	Engaged/Invested
Focus	Organization	Groups	Individual
Technologies	Client-server HTML, portals	Peer to Peer XML, RSS, Java	RDF, RDFS, OWL
Content	Owning	Sharing	Curation
Interaction	Web forms	Web applications	Smart applications

students place greater trust in information obtained through social interaction. Profiles are very effective in determining whether one accepts the content on the web; social networking reveals the profiles, which make the students trust the content, and if the students are unsure about the content, tagging their classmates could help them to decide. An empirical study conducted by Valkenburg et al. [7] showed that the responses received on the user's social networking posts had an effect on the self-esteem of the user. Using social networking in the education sector will enhance the following factors namely: peer feedback, critical thinking, tag friends, join groups and sharing of knowledge.

3 Indian Higher Education and Social Networking

India is a democratic nation with many religions each with its own religious customs. Tourists around the world are attracted to its rich history and fantastic landscape. The global culture has been increasing in importance with the new technology enabling access to and sharing of information. There are more than 1000 communities with their own individual social characteristics. Technology has become a tool to connect the communities and work towards a better society. The various communities and volunteer organizations can use the social networking as a platform to raise awareness of social issues. Volunteer organizations play an important role in supporting those people who need them the most. Social networking provides them with the platform needed to publicise their plans for society. Social networking has been widely used by government organizations to brand and market their policies [8]. The following report generated by Tobias and Carlson [9], suggests that social media is popular in India. The active social media users are found to be 8%, which is a good sign. The mobile phone users have been increasing widely. The report shows 87% of users are accessing social networking through mobile technology. The report also shows that there are plenty of social networking users; however, in India as a whole, the percentage of social media users is not up to the mark although it is improving day by day.

The development of the nation depends on the growth of the economy, and this can be achieved by improving the higher education sector. This research on Indian higher education does not shift the thinking of an entire educational system;

however, it provides food for thought that may be helpful in reducing technical constraints. Indian institutions are performing well, but they have to perform better to obtain recognition internationally [10]. India, with a population of 600 million below 25 years of age, is likely to have the second largest graduate talent by the end of 2020 next to China and followed by the USA [11]. Indian educational institutions have failed to gain recognition globally. This issue was raised by the Indian president in a recent convocation ceremony held by a prestigious university in India.

There are several reasons for not India achieving global recognition. We cannot entirely blame the educational system as it stands; there are several other reasons. The Indian population is vast and accessibility to knowledge has been difficult for the ordinary person for many years, although it has improved considerably. India has more than 343 universities but even the reputed institutions are finding it difficult to keep up with changing global requirements. Researchers believe that it is time for a complete overhaul of the higher education system in India. The number of dropouts in India was astounding in 2006; the report revealed 56% of students had dropped out of university [12].

Social networking helps in various ways to integrate the process of rethinking before accepting the content. It also helps educational institutions to promote their courses and their university in general. One of the reasons for the high dropout rates discussed in the previous section may be that students are not well informed about the courses in which they enrol. Social networking provides an opportunity of describing the course in detail, together with the job opportunities and career advice. Students are often locked into a set of subjects, not all of which may be useful for a career. Social networking would enable a conversation to be initiated between the students and the schools. It helps to create a common ground for interaction between students, education institutions, and communities.

Social networking could help India to examine closely its economic growth. The Indian GDP has been very inconsistent for many years, whereas China, which has a higher population than Indian, has a steady GDP growth and is far better than India in terms of GDP. India is an efficient country that has the capability to have consistent growth every year. However, policy making procedures and investment are the main reasons for this paradox [13].

The people of India are often unaware of the procedures, which take place for the development. Indian educational institutions and communities have skill sets, which can improve the system implementations. If the government proclaims that its plans are for the benefit of its people, it is better to consult the people. In India, it is more often that the faith is built on the individuals rather than the policies. Social networking could address this by providing better communication between the content of the policy, policy makers, and the people.

In terms of practicality, this is highly achievable. Indian educational institutions can give students an opportunity to develop and create ideas that could change the perspective of the government. For example, in India the average class size for a course of higher education would be 50 students. Let us say there is one unit focusing on policy, which relates to the policy. We can divide the class into two, let

us say class A and B. Class A would be responsible for creating ideas and submitting them to class B with supporting evidence. Wiki would be the best tool for scrutinizing this process. Class B would formulate the adverse effects of the plans provided by class A. The discussion would start, and they can debate about the advantages and the disadvantages. The lecturers can play their part in motivating the students to use social networking and making it more fun rather than the usual teaching approach. The educational institutions can make use of the wiki, blogspots and newsletter to highlight their students' works. In this way, the information could reach the government in no time and may help to resolve some part of the issue.

Social networking provides a platform for teachers and educational institutions to develop new ways of engaging with students. Communication between the school and students plays a major role in determining the path that the system will take in providing for the needs of the student. Social networking is derived from web 2.0 and provides users with contents generated by other users [14]. At the beginning of the development of online social networking, the universities made good use of them by integrating instant messaging and profiles with the bulletin boards. Different social networking sites were created year after year, which benefited a large number of student communities. In order to develop education, government organizations encourage collaboration to improve communication and interaction between the education providers and students.

Kahne et al. [15], followed a theory-driven approach to examine the development of schools through collaborative networks. Kahne et al. [15] state that initiatives by the government organizations to improve collaboration have had a considerable positive effect on schools rather than communities, and Kahne et al. [15] suggests that trust factors play a significant role in collaboration, as students are not likely to express themselves via untrusted networks. Collaboration consists of groups with shared values working for a sole purpose but also improving skills. The ability to obtain information from the other people in the group and to allow the flow of new information would enrich the discussion.

Researchers believe that innovative development of cities can be achieved with cooperation; currently, cities compete to display their advantage over another city [16]. Collaboration encourages students to think globally. The universities in India can make communities of their respective field of study and contribute to the study. Educational institutions have the opportunity of providing a platform for their students by means of which they can collaborate and monitor the students' work frequently. The motivation and involvement of teachers plays an important role in encouraging the students to participate. Students can make great use of collaboration as it helps them to experience new ways of learning. Social networking helps students to communicate quickly with students who have similar interest(s). Osborne [14], argues in his article that online social networking would strengthen the partnership between the students and communities because it involves better interaction and engagement. In another research conducted by Prasad and Ramakrishna [17], it was shown that online social networking tools assist peer communication globally.

4 Advantages and Disadvantages of Social Networking

Social networks help students to interact with the community of their own educational institution, and help them to interact with the students of other institutions. Social networks can provide students with information that is authentic, relevant, and up-to-date. Social networks help the individuals to validate the information before accepting it [18]. The students obtain several other advantages when accessing the content, such as: Updated information; New knowledge; Global awareness and Aspects of past. Social networking would help the students to maintain the inter-crossing relationships in every different aspect of their lives. Furthermore, collaboration, communication skills, be environment-friendly and acquire new acquaintances.

The research conducted by UNESCO [19] shows that, world-wide there is a decline in formal education. The reason would be a failure of socialization in an educational context. Social networking enables users with a common interest to communicate and develop relationships. Social networking applications can help students to perform better in their studies. Google Scholar is the best online social networking website which has a collection of resources from different databases which are linked to Google's database [20]. There are various other online databases such as ProQuest, Springer Link, Science Direct, IEEE explorer and many more, which might help to improve students' analytical skills.

The Indian education system is rigid and depends on a fixed set of resources, forcing the student to learn within established boundaries. Social networking gives students an opportunity to overcome intense classroom study. This is feasible in India given that the educational institutions support and accept resources available from databases. Accessibility of databases is not formally practised in India; databases, which are made accessible, are only those that correspond with the ideas of the authors of textbooks.

Enabling students to access information online can have another advantage. Students are often stressed about the research or study because they cannot engage in and enjoy the study process, but rather, they consider it as a chore. This stress can be decreased by social interaction occurring within the student groups [21]. By using social networking students will obtain several advantages i.e. study unconventionally; scrutinize research; complete study quickly; overcome study stress and problems; and complete study quickly.

Social networking would help the students to maintain the inter-crossing relationships in every different aspect of their lives. The research conducted by UNESCO [19] shows that, world-wide there is a decline in formal education. The reason would be a failure of socialization in an educational context. Social networking enables users with a common interest to communicate and develop relationships. Social networking applications can help students to perform better in their studies. Google Scholar is the best online social networking website which has a collection of resources from different databases which are linked to Google's database [20]. There are various other online databases such as ProQuest, Springer

Link, Science Direct, IEEE explorer and many more, which might help to improve students' analytical skills.

The Indian education system is rigid and depends on a fixed set of resources, forcing the student to learn within established boundaries. Social networking gives students an opportunity to overcome intense classroom study. This is feasible in India given that the educational institutions support and accept resources available from databases. Accessibility of databases is not formally practised in India; databases, which are made accessible, are only those that correspond with the ideas of the authors of textbooks. Enabling students to access information online can have another advantage. Students are often stressed about the research or study because they cannot engage in and enjoy the study process, but rather, they consider it as a chore. This stress can be decreased by social interaction occurring within the student groups [21]. By using social networking students will obtain several advantages i.e. Study Unconventionally; Scrutinize research; complete study quickly; overcome study stress and problems; and complete study quickly.

Communication skills are a basic requirement for any individual who wants to be successful in a competitive environment. The above advantages of collaboration and inter-cross relationships aim to improve the communication skills of an individual. Burke et al. [22], state that the skills developed through interactions are unique, and term these 'social communication skills'. Strong Communication skills can be acquired with the usage of social networking, and it does not require any particular effort in a social environment. Communication skills are part of the curriculum of the social network. When an individual expresses his views in a social environment, it involves more of the cognitive principles and can and the ideas of others can be grasped more readily.

The skills acquired within the social environment are highly encouraged in a workplace environment as well. The recent study by Brown and Vaughn [23] shows that the organizations hire employees based on the screening procedure of their profiles in online sites. We have found examples of bloggers who found career as a journalist due to the skills presented in their blogs. Social networking helps to demonstrate the skills to the employers. Sustainability has been a part of urban policy and development from past many years. Dempsey et al. [24] claim that social equity and justice aims to minimize the gap between sustainability community and social sustainability. Social sustainability questions the development that is taking place from a sustainability perspective. The sustainability community refers to the people who live and work in the present and future, who are capable of understanding the environment for the welfare of the community [24]. Social networking provides an opportunity to explore the factors associated with sustainability.

The organizations with various departments can work together to create the awareness and follow the sustainability principles. Sustainability starts from a piece of paper and extends to the waste generated through technology, factories, and various other factors. In the research conducted by Kotler [25], he claims that social marketing is a term that has been used for the past 40 years, and also, social marketing can generate awareness and has the ability to change behaviour. There are various government and non-government agencies working for the welfare of

the environment just as there are many political parties opposing the carbon emissions in factories. Social networking helps us to understand the consequences of neglecting the environment and ignoring the factors that harm it. There are various tools available to students that can help them to know how they can help society and the environment. Sustainable measures in developing nations such as India are not a high priority. Social networking can help with the marketing of products that are less harmful to society.

The users of social networking have to be well aware of disadvantages that they might face when using the application. Users tend to utilize it to its maximum extent without being aware of the consequences that they may encounter. One drawback of social networking comes from the information required to access the network. It is seen that people with different attitudes and perspectives access the information. In the research conducted by [26], it was found that accepting multiple contents from the web at the same time may scatter attention and can have adverse effects on memory and cognitive development. While searching for the information online, there are situations where the users end up in viewing irrelevant information. There are plenty of applications online that might confuse some users, but this situation can be relatively improved as the user gets used to searching for information [27].

In a social networking environment, there are situations that could embarrass an individual or educational institution. Social networking promotes sharing of information. In an educational system, if the staffs are asked to share their profiles with the students, the institution is at risk. It has been seen that individuals, when conversing on social networks, may lose control; but when an individual is representing an organization, it can be an embarrassment to the institution. Institutions need to have well-defined policies before engaging in a social network [14].

The malicious use of the Internet has increased and spread to social networking. Malicious hackers have been known to access users' profiles to steal private information. Trust plays an important factor in privacy issues [26]. Social interaction in previous years was mostly face-to-face. Social networking has undoubtedly reduced the occurrence of such social meetings. The physical presence of an individual is being taken away. The expression, happiness, and joy to be found in face-to-face meetings are decreasing due to the impact of technology. The younger generation finds the wall posts, updates; activity feeds etc., on the social networking websites more attractive. Access to social networking has been banned in many schools. The research conducted by [28] shows that youth are more likely to become addicted to social networking and are finding joy in updates made by their colleagues. Mutual understanding might be better with social networking but sometimes has a deleterious effect in the real world. Individuals have been known to be offensive to strangers in the real world because their social interactions have been limited to social networking.

Furthermore, the users of social networking have to be well aware of disadvantages that they might face when using the application. Users tend to utilize it to its maximum extent without being aware of the consequences that they may encounter. One drawback of social networking comes from the information

required to access the network. It is seen that people with different attitudes and perspectives access the information. In the research conducted by [26], it was found that accepting multiple contents from the web at the same time may scatter attention and can have adverse effects on memory and cognitive development. While searching for the information online, there are situations where the users end up in viewing irrelevant information. There are plenty of applications online that might confuse some users, but this situation can be relatively improved as the user gets used to searching for information [27].

In a social networking environment, there are situations that could embarrass an individual or educational institution. Social networking promotes sharing of information. In an educational system, if the staffs are asked to share their profiles with the students, the institution is at risk. It has been seen that individuals, when conversing on social networks, may lose control; but when an individual is representing an organization, it can be an embarrassment to the institution. Institutions need to have well-defined policies before engaging in a social network [14]. The malicious use of the Internet has increased and spread to social networking. Malicious hackers have been known to access users' profiles to steal private information. Trust plays an important factor in privacy issues [26]. Finally, Social interaction in previous years was mostly face-to-face. Social networking has undoubtedly reduced the occurrence of such social meetings. The physical presence of an individual is being taken away. The expression, happiness, and joy to be found in face-to-face meetings are decreasing due to the impact of technology. The younger generation finds the wall posts, updates; activity feeds etc., on the social networking websites more attractive. Access to social networking has been banned in many schools. The research conducted by Ahn [28] shows that youth are more likely to become addicted to social networking and are finding joy in updates made by their colleagues. Mutual understanding might be better with social networking but sometimes has a deleterious effect in the real world. Individuals have been known to be offensive to strangers in the real world because their social interactions have been limited to social networking.

5 Research Gap

The research in improving the education sector with social networking examined the characteristics of social networking. New research has to be conducted focusing on developing social networking in a way that takes into account the country's culture. The cultural characteristics of different countries might not be compatible with all aspects of social networking and this notion needs to be investigated by researchers. A study related to the effects of social networking on student behaviour would benefit the student community. This research focuses on the advantages and disadvantages of social networking in terms of students; however, a study that has the potential to encourage teachers to motivate students with the help of social,

networking, would be beneficial. There is also scant research on differentiating classroom practices by the use of online social networks and the digital education required to implement social networking.

6 Research Method and Question

This research aims to address and answer the following question "What are the advantages and disadvantages of using social networking in India's higher education sector?" In this research, an online survey was generated based on the current literature review by using Qualtrics, a tool used to create an online survey and providing various other facilities. The Internet is the main advantage of the online survey as it is been used by different sectors of society and helps us to reach people quickly. The online survey helps us to communicate easily with potential respondents through emails, social networking websites, mobile applications etc. [29]. The online survey must meet certain requirements, including approval from the ethics committee and the research supervisor.

The survey tool generates a link via Qualtrics through which a respondent obtains information about the purpose of the research and the university personnel involved in the research well before starting the survey. These factors encourage the respondents to take the survey seriously, because it looks professional. People are fed up with online surveys, which make them spend the least amount of time in completing it, mainly due to a lack of awareness of its impact, but this tool helped us to explore the importance of the survey, which in turn obtained the responses considerably from the respondents.

The online survey was generated and developed based on the current literature review and it consists of five distinct parts. The first part of the survey presents information regarding the aims of the research and the associated legalities. It also includes a questionnaire about gender and age. The second part asks respondents to state their profession and/or field of study. The third part asks respondents to give information about the number of hours spent daily in social networking activities and Internet for email. It also provides the opportunity for the respondent to elaborate on the hours spent over the Internet and social networking. The fourth part of the online survey consists of a 5-point Likert scale designed to cover the positive and negative effects of using social networking for education and for personal interactions. The last part gives participants the opportunity to elaborate on what they perceive to be the negatives associated with social networking.

The online survey design has been used by many researchers to collect and examinee data, such as [30–34]. Moreover, the Internet is the quickest way to obtain responses from people located in different parts of the world, which is accomplished through an online survey. The online survey has several features, which allow us to obtain a range of opinions about a particular topic. A number of other researchers have described the positive and negative factors involved in the online survey. Jøsang et al. [35], believes that the online surveys are one of the most

important factors in decision-making. In the study, the respondents need to be acquired from different parts of the world, so the online survey was used for data collection. In addition, data can be collected quickly with the help of various online applications such as email and social networking websites.

Tiene [36], states that the research conducted through an online survey would trigger online discussion with students. Online survey methodology has been used in this research for various reasons such as [37]: Obtains data through various resources. The online survey uses the technology to reach the person quickly, allows the user to complete it at any time, whereas the face-to-face interview is time-consuming, and requires a certain protocol to be followed. The online survey also helps the researcher to target a specific population. Hence, it makes grouping and analysing of data more efficient compared to face-to-face interviews. The online survey helps the respondents to express themselves better as they can do the survey anytime and they have the opportunity to think more about the questions. It is also cost-effective, as the researcher does not have to travel to different locations in order to conduct a face-to-face interview.

The online survey also has the following disadvantages [32, 37]: It requires that the participant to have some knowledge and experience with online applications. Apart from the above issues, several other issues such as sample and accessibility issues are faced. The sample issues involve grouping of the participants according to their location, but the recent online survey tools would solve such issues. The disadvantages were considered in this research, were found to be minimal, and can be avoided through recently available online survey tools such as Qualtrics. The researcher was able to distribute an online survey through email and online networking sites involving the online tools experience.

A sample of 131 (valid respond rate) is obtained from the southern part of India; the online survey was created to suit the current situation of education and social networks. Using Qualtrics, the survey was distributed to people in the southern part of India. Data collected undergo further quantitative analysis in order to evaluate the responses generated by the survey, and are analyzed using the statistical software analysis tool, SPSS. The data collected from Qualtrics are entered into SPSS, and various data analysis techniques take place. The analysis in intended to obtain information about the factors that is pertinent to the southern part of India. The research explores both the advantages and limitations of using social networks in education.

The sampling method used in this study is adopted from the empirical study conducted by Lin et al. [38] Heckathorn [39], whereby we contact three friends requesting that they complete the survey, and ask them to suggest three more people, who in turn will suggest three more and so on; this is how more respondents were recruited for this study.

The data which is been collected will be analyzed using SPSS version 24. The data obtained through the survey are entered into the SPSS tool and various statistical tests are performed. The result obtained by the SPSS tool depends on the type of tests that we perform along with the type of tests required for this study. The factors are selected after various tests, which then will be recognized as the critical

components in the outcome of the research. In order to check the variation in the sample variables with that of the actual population, factor analysis is conducted. Reliability tests help us to ascertain the consistency of our findings.

As per Mark et al. [40] the reliability and validity of the data can be derived through the application of three principles: By checking whether the results are similar on other occasions; By testing whether the same observations are made by other researchers; By ensuring the transparency of the data; In this study, to test the reliability and validity, the statistical technique Cronbach's Alpha test is conducted, and the result is analyzed.

7 Results

Data was compiled mainly with the help of social networking websites, instant messaging applications, and email. The online survey link generated by Qualtrics was distributed using these tools and data was collected. The lecturers in India were contacted by mail and asked to distribute the questionnaires to the students. Few of the lecturers were available on Facebook or LinkedIn; therefore, they were contacted via instant messaging. The groups in instant messaging applications such as WhatsApp and Viber were also used to contact the students. Table 2 shows gender, and age of the online survey respondents. The participants for this study were 175 (131 valid respond rate) from Garden City College, which is located in the Southern part of India. 52% are male, while 48% are female. For this study, the highest participants' age is 48% from ranged 22–32.

While Table 3 presents the participants' fields of study; the accounting sector is considered the highest respond rate (34.34%) for this study followed by science and engineering (11.45%).

Table 2 Gender and age—prepared by the authors

Gender		
	Response	%
Male	91	52
Female	84	48
Total	175	100
Age		
18–22	54	30.86
22–32	84	48.00
32–42	29	16.57
42–52	7	4.00
Over 52	1	0.57
Total	175	100

Table 3 Fields of study—prepared by the authors

Field of study		
	Response	%
Accounting	57	34.34
Business law	16	9.64
Economics and finance	8	4.82
Information systems	4	2.41
Information technology	10	6.02
Computer science	14	8.43
Management	11	6.63
Marketing	4	2.41
Health sciences	1	0.60
Humanities	3	1.81
Science and engineering	19	11.45
Art and design	6	3.61
Others—please specify	13	7.83
Total	166	100

Table 4 Highest education level—prepared by the authors

Highest education level		
	Response	%
Primary education	18	10.91
Higher secondary/Pre-university	6	3.64
Professional certificate	13	7.88
Diploma	20	12.12
Advanced/Higher/Graduate diploma	11	6.67
Bachelor's degree	55	33.33
Post graduate diploma	9	5.45
Master's degree	33	20.00
Total	165	100

Table 4 shows that Bachelor's Degree holders are the highest participants who completed the online survey with 33.33% followed by Master's degree (20%) and Diploma (12.12%).

Table 5 shows that Indian participants (46.99%) are spending less than 1 h on social networking daily (not including email), followed by 35.54% up to 5 h daily.

From the survey results (see Table 6), it was noted that the majority (48.41%) of the participants spend on social networking (not including email) less than an hour daily; on the other hand, 53 of the participants spend 33.76% while 34% spend up to 5 h on the social networking.

The Cronbach's Alpha is conducted to test the internal consistency of the variables. Values ranging from 0.7 to 0.9 are recommended, even though values as low as 0.5 are slightly significant [41]. In this study, to test the reliability and validity, the statistical technique Cronbach's Alpha test is conducted, and the result

Table 5 Number of hours spent on social networking daily, not including email? (Per day)—prepared by the authors

How many hours do you spend on the social networking daily, not including email? (Per day)		
	Response	%
Less than an hour	78	46.99
Up to 5 h	59	35.54
5–10 h	21	12.65
10–20 h	4	2.41
Over 20 h	4	2.41
Total	166	100

Table 6 Number of hours spent on the Internet for email (per day?)—prepared by the authors

How many hours do you spend on the internet for email? (Per day)		
	Response	%
Less than an hour	76	48.41
Up to 5 h	53	33.76
5–10 h	19	12.10
10–20 h	7	4.46
Over 20 h	2	1.27
Total	157	100

is analyzed. For the advantages and disadvantages the Cronbach's Alpha is 0.99 and based on the rules of thumb >0.9 is Excellent [37, 42] (Table 7).

Furthermore, the KMO for the Advantages and disadvantages are Marvellous according to Beavers et al. [43]. The Bartlett's test of sphericity is highly significant for both advantages and disadvantages, $\chi^2 = 5632.949$, 6657.093 df = 300 and 435 respectively and $p < 0.000$, indicating that the items of the scale are sufficiently correlated to factors to be found [9] (see Table 8).

Table 9, the component matrix—advantages, generated three new factors from the Indian perspective, and these new advantages are namely: rapidly research study and personal skills; global awareness, new knowledge and communication; and romance relationship. These advantages will assist the Indian students to obtain the necessary skills and knowledge via assimilating social networking in the education sector especially in India.

Table 10 present the component matrix for the disadvantages via the social networking in India. From the online survey results two new disadvantages are generated namely Elude social activities; deep thinking and stress and averts from traditional activities and privacy. A warning message generated from this result, as

Table 7 Reliability statistics—Cronbach's alpha for advantages and disadvantages—prepared by the authors

	Reliability statistics	
	Cronbach's alpha	N of items
Advantages	0.991	25
Disadvantages	0.994	30

Table 8 KMO and Bartlett's test—advantages and disadvantages—prepared by the authors

KMO and Bartlett's test		Advantages	Disadvantages
Kaiser-Meyer-Olkin measure of sampling adequacy		0.926	0.926
Bartlett's test of sphericity	Approx. chi-square	5632.949	6657.093
	df.	300	435
	Sig.	0.000	0.000

using the social networking tools in the education sector especially in India, can prevent and isolate users from the social and personal activities and develop type of anxiety; traditional activities and Intellectual property and security. The respondents are worried about the privacy concerns related to social networking. Privacy issues depend on how the technology is used; hence, it is incumbent upon the organization providing a platform for social networking to address this issue and safeguard the users as much as possible. The Indian educators should address these disadvantages by preparing and presenting workshops and seminars among the Indian students to raise and promote social networking usage and awareness.

8 Discussion, New Findings, Limitations and Recommendations

There is evidence that students have generally had positive experiences associated with the use of social networking, apart from a few negative concerns. The data collected provided new findings on social networking in terms of the Indian context. The analysis helped us to identify the advantages that can be concentrated and implemented as part of the curriculum, bearing the disadvantages in mind. The research question for this study concerns finding the advantages and disadvantages of using social networking in the higher education sector of India, and the study primary goal is to conduct a study on social networking to ascertain its importance in higher education (see Table 11).

The researchers then study the requirements of social networking from the Indian perspective. India is a vast country with the second largest population in the world; it is difficult to implement procedures, which can meet everyone's needs and wishes.

The researchers explore the advantages that an Indian student can obtain through social networking. The researcher also provides suggestions that can be used by institutions when considering the adoption of social networking in line with government policies on education. The researchers answered the objectives of the study; as they identified the evidences, which lead to the generation of advantages and disadvantages of using social networking. Social networking creates advantages, but the social connectedness is not similar to normal interactions. It completely depends on the users and the way they see social networking; some users are comfortable with face-to-face interaction and some preferred social networking.

Table 9 Component matrix—advantages—prepared by the authors

	Rotated component matrix[a]			
		Component		
	New advantages for India	1	2	3
Scrutinize my research study more easily	Rapidly research study and personal skills	0.809	0.424	0.347
Understand and solve study problems easily		0.783	0.423	0.359
Concentrate more on my reading and writing skills		0.724	0.381	0.496
Complete my study more quickly		0.718	0.481	0.393
Develop my personal and communication skills		0.683	0.452	0.478
Acquire new acquaintances—work related		0.678	0.414	0.519
To prepare my professional attitude toward study and work		0.675	0.431	0.512
Reduce carbon footprint in my activities		0.642	0.439	0.555
Overcome study stress		0.630	0.472	0.461
Study independently		0.617	0.550	0.416
Be more sustainable person		0.579	0.453	0.568
Provide reliable and scalable services		0.566	0.518	0.538
Be more aware of global issues/local issues	Global awareness, new knowledge and communication	0.392	0.831	0.277
Communicate with my peers frequently		0.300	0.790	0.390
Learn new information and knowledge		0.462	0.784	0.206
Gain up-to-date information		0.550	0.759	0.227
To remember facts/aspects of the past		0.430	0.755	0.354
Communicate with my peers from different universities		0.344	0.733	0.499
Collaborate with my peers frequently		0.297	0.726	0.509
Acquire new acquaintances—romance relationship	Romance relationship	0.417	0.313	0.785
Do whatever I want, say whatever I want, and be whoever I want		0.544	0.293	0.709
Acquire new acquaintances—friendship relationship		0.508	0.409	0.705
Communicate with my different communities		0.335	0.609	0.626
Develop intercrossing relationships with my peers (i.e. Artistic talents, sport and common interests)		0.438	0.580	0.594
Become more "Greener" in my activities		0.558	0.483	0.563
	Extraction method: principal component analysis Rotation method: Varimax with Kaiser normalization[a]			

[a]Rotation converged in 9 iterations

Table 10 Component matrix—disadvantages—prepared by the authors

	Rotated component matrix[a]			
		Component		
	New disadvantages for India	1	2	3
Prevents me from participating in social activities	Elude social activities, personal skills and anxiety	0.808	0.366	0.342
Stresses me		0.771	0.414	0.348
Decreases my grammar and proofreading skills		0.761	0.372	0.406
Decreases my deep thinking		0.754	0.452	0.371
Prevents me from remembering the fundamental knowledge and skills		0.715	0.404	0.411
Depresses me		0.708	0.461	0.393
Bores me		0.692	0.549	0.311
Makes me sick and unhealthy		0.686	0.460	0.443
Distracts me easily		0.679	0.403	0.505
Prevents me from completing my work/study on time		0.673	0.532	0.340
Prevents me from concentrating more on writing and reading skills		0.643	0.494	0.321
Prevents me from having face to face contact with my family		0.628	0.589	0.363
Scatters my attention		0.609	0.505	0.503
Makes me addict		0.581	0.571	0.403
Prevents me from talking on the phone/mobile	Avert traditional activities	0.454	0.760	0.347
Prevents me from watching television		0.460	0.747	0.415
Prevents me from shopping in stores		0.427	0.731	0.457
Prevents me from having face to face contact with my friends		0.439	0.730	0.454
Prevents me from reading the newspapers		0.430	0.721	0.476
Prevents me from completing my work on time		0.520	0.710	0.373
Prevents me from completing my study on time		0.520	0.704	0.361
Makes me feel lonely		0.569	0.670	0.348
Prevents me from participating in physical activities		0.423	0.661	0.530
Makes me more gambler		0.606	0.639	0.391
Makes me lazy		0.496	0.625	0.501

(continued)

Table 10 (continued)

	Rotated component matrix[a]			
		Component		
	New disadvantages for India	1	2	3
Increase intellectual property concerns	Intellectual property and security	0.414	0.406	0.783
Increase security concerns		0.420	0.410	0.771
Increase privacy concerns		0.407	0.429	0.767
Makes me receive an immoral images and information from unscrupulous people and it is difficult to act against them at present		0.488	0.563	0.581
Makes me insecure to release my personal details from the theft of personal information		0.548	0.486	0.555
	Extraction method: principal component analysis Rotation method: Varimax with Kaiser normalization[a]			

[a]Rotation converged in 8 iterations

The main functionality which social networking takes over from regular interaction is the creation of awareness along with the potential for constant interaction.

The volunteer organizations and social networking applications are working together to resolve the social issues. The level of response is very high. In India, social networking can increase digital education. The policy makers of government organizations should become more flexible and have a common understanding with the social networking applications to work towards a common goal. The current social networking applications work as a business, which needs to be changed to bring about a change in the society.

The educational system of India can be enhanced with better use of social networking. The top international institutions provide excellent examples of social networking within the learning environment. The government of India has taken many steps to implement social networking in the top universities. The population and changing technological requirements are considered as central issues; however, more research has to be conducted to reach the students in remote area who are forced to travel long distances in search of knowledge and careers.

The web technologies are changing constantly which results in the development of online tools, which in turn influences the academics. The effects of social networking in education have been observed in the research. This chapter reveals the impact of social networking on education in India; however, the adoption of social networks in the Indian education sector has not been yet been trailed. The major objective of this research is to make a concerted effort to forecast the advantages and disadvantages associated with social networking in higher education. The study explored a number of positive and negative factors that need to be considered by education providers, learners, and developers of online tools.

Table 11 New advantages and disadvantages by using social networking in Indian higher education—prepared by the authors

The Indian education system has made web content available to a certain extent in urban areas. The people from rural areas usually migrate to urban areas in search of education and careers. Web technologies have to concentrate more on making information available in rural areas. In India, there is a high student drop-out rate from educational institutions due to stress, failing units, or losing interest in the course after a few months, to name a few reasons. In order to address this, future research on the attitudes of Indian drop-outs would be helpful. Moreover, research on ways to achieve better enrolment and information about courses through social networking would be beneficial. Finally, the same research should be replicated in other areas in India to justify the research aims and objectives.

Finally, it is evident that social networks provides a platform by means of which one can express, work and interact. It has the capability to withstand population changes and various internal issues pertaining to India. Information can be made easily accessible to students who live remote villages, enabling them to work with other students. Digital education has to be the mantra of the government, but it has to be well handled with consideration given to the potential disadvantages.

9 Conclusion

Next to China and USA, India plays a major role in providing a workforce that contributes to global development. Social networking helps to people to form relationships with others who have similar interests, and this is a good thing. However, we can see conflict arising between the traditional and modern media. The rural students have to make more contribution in future. Accessibility of the information in rural areas would help to improve the global knowledge. Social networking is the best way to access and share the information. Digital education will play a crucial in coming years and has to be initiated. The technologies have been a boon but if consumed more than expected, they are going to destroy the planet bit by bit. Finally, to the researchers conclude that the list of disadvantages associated with social networking lengthens with the new online technologies, so in order to increase social relationships we need to think as one. In the researchers' opinion, social networking helps people to be part of the community and it is our responsibility to make sure the community does not receive a bad reputation because of our actions. Finally this research generated three new advantages and disadvantages of social networking usage by students from the southern part of India. Further research will be carried out by the researchers to replicate the same study in other areas in India to justify the research aims and objectives.

References

1. Rogers, E.M.: Communication Technology, vol. 1. Simon and Schuster (1986)
2. Christoffersen, M.B., Boukaouit, D., Weeke Hervit, B.H., Winther Brødreskift, D., Makilä, R. M., Pingel Vogel, K.E., Wolter Strate, S.: Social Media Ethics (2012)
3. Nickerson, R.S.: Electronic bulletin boards: a case study of computer-mediated communication. Interact. Comput. **6**(2), 117–134 (1994)
4. Aghaei, S., Nematbakhsh, M.A., Farsani, H.K.: Evolution of the world wide web: from web 1.0 to web 4.0. Int. J. Web Semant. Technol. **3**(1), 1–10 (2012)
5. Lassila, O., Hendler, J.: Embracing web 3.0. Internet Comput. IEEE **11**(3), 90–93 (2007)
6. Pempek, T.A., Yermolayeva, Y.A., Calvert, S.L.: College students' social networking experiences on facebook. J. Appl. Dev. Psychol. **30**(3), 227–238 (2009). https://doi.org/10.1016/j.appdev.2008.12.010
7. Valkenburg, P.M., Schouten, A.P., Peter, J.: Adolescents' identity experiments on the internet. N. Media Soc. **7**(3), 383–402 (2005)
8. Mahajan, P.: Use of social networking in a linguistically and culturally rich India. Int. Inf. Libr. Rev. **41**(3), 129–136 (2009)
9. Tobias, S., Carlson, J.E.: Brief report: Bartlett's test of sphericity and chance findings in factor analysis. Multivar. Behav. Res. **4**(3), 375–377 (1969)
10. Jain, P., Sikka, P.: Corporate Sector Participation: Much Needed Elixir of Life to Indian Higher Education System (2014)
11. Council B: Understanding India: The future of higher education and opportunities for international cooperation (2014)

12. Umashankar, V., Dutta, K.: Balanced scorecards in managing higher education institutions: an Indian perspective. Int. J. Educ. Manag. **21**(1), 54–67 (2007). https://doi.org/10.1108/09513540710716821
13. Atir, M.: Efficient policy of India. J. Public Adm. Policy. Res. **5**(6), 141–144 (2013)
14. Osborne, N.: Using social media in education, part 1: opportunity, risk, and policy. IBM Developer Works (2011)
15. Kahne, J., O'Brien, J., Brown, A., Quinn, T.: Leveraging social capital and school improvement: the case of a school network and a comprehensive community initiative in Chicago. Educ. Adm. Q. **37**(4), 429–461 (2001)
16. Lever, W.F.: Innovation in urban policy: collaboration rather than competition between cities. Urban Compet. Innovation **91** (2014)
17. Prasad, E., Ramakrishna, P.: Social networks and online communities in higher education. Int. J. Sci. Eng. **4**(1) (2013)
18. Donal, O.M., Duma Cornel, L., Elena, R., Lidija, K., Norbertas, A., Susana, B., Svetla, M., Pedro, P.: Challenges and opportunities for schools and teachers in a digital world. Lessons Learned from the 2012 SMILE Action Research Project (2012)
19. Social Media for Learning by Means of ICT (2011)
20. Conover, M.D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., Flammini, A.: The geospatial characteristics of a social movement communication network. PLoS One **8**(3), e55957 (2013)
21. Thoits, P.A.: Mechanisms linking social ties and support to physical and mental health. J. Health Soc. Behav. **52**(2), 145–161 (2011)
22. Burke, M., Kraut, R., Marlow, C.: Social capital on facebook: differentiating uses and users. Paper Presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada (2011)
23. Brown, V., Vaughn, E.D.: The writing on the (facebook) wall: the use of social networking sites in hiring decisions. J. Bus. Psychol. **26**(2), 219–225 (2011). https://doi.org/10.1007/s10869-011-9221-x
24. Dempsey, N., Bramley, G., Power, S., Brown, C.: The social dimension of sustainable development: defining urban social sustainability. Sustain. Dev. **19**(5), 289–300 (2011)
25. Kotler, P.: Reinventing marketing to manage the environmental imperative. J. Mark. **75**(4), 132–135 (2011). https://doi.org/10.1509/jmkg.75.4.132
26. Ophir, E., Nass, C., Wagner, A.D.: Cognitive control in media multitaskers. Proc. Natl. Acad. Sci. **106**(37), 15583–15587 (2009)
27. Andreas, M.K., Michael, H.: Users of the world, unite! The Challenges and Opportunities of Social Media. 59–68 (2010)
28. Ahn, J.: The effect of social network sites on adolescents' social and academic development: Current theories and controversies. J. Am. Soc. Inform. Sci. Technol. **62**(8), 1435–1445 (2011)
29. Fox, S., Rainie, L., Larsen, E., Horrigan, J., Lenhart, A., Spooner, T., Carter, C.: Wired Seniors. The Pew Internet and American Life Project (2001)
30. Cheung, C.M.K., Chiu, P.-Y., Lee, M.K.O.: Online social networks: why do students use facebook? Comput. Hum. Behav. **27**(4), 1337–1343 (2011)
31. Dodds, P.S., Muhamad, R., Watts, D.J.: An experimental study of search in global social networks. Science **301**(5634), 827–829 (2003)
32. Issa, T.: Online Survey: Best Practice. In: Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies. IGI Global, pp. 1–19 (2013). https://doi.org/10.4018/978-1-4666-2491-7.ch001
33. Kwon, O., Wen, Y.: An empirical study of the factors affecting social network service use. Comput. Hum. Behav. **26**(2), 254–263 (2010)
34. Malhiwsky, D.R.: Student Achievement Using Web 2.0 Technologies: A Mixed Methods Study (2010)
35. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decis. Support Syst. **43**(2), 618–644 (2007)

36. Tiene, D.: Online discussions: a survey of advantages and disadvantages compared to face-to-face discussions. J. Educ. Multimedia Hypermedia **9**(4), 369–382 (2000)
37. Tavakol, M., Dennick, R.: Making sense of Cronbach's alpha. Int. J. Med. Educ. **2**, 53 (2011)
38. Lin, H., Fan, W., Chau, P.Y.K.: Determinants of users' continuance of social networking sites: a self-regulation perspective. Inf. Manag. **51**(5), 595–603 (2014)
39. Heckathorn, D.D.: Respondent-driven sampling: a new approach to the study of hidden populations. Soc. Probl. **44**(2), 174–199 (1997)
40. Mark, S., Philip, L., Adrian, T.: Research Methods for Business Studies (2007)
41. Bland, J.M., Altman, D.G.: Statistics notes: Cronbach's alpha, vols 314, 7080 (1997). https://doi.org/10.1136/bmj.314.7080.572
42. Connelly, L.M.: Cronbach's alpha. Medsurg Nurs. **20**(1), 44–45 (2011)
43. Beavers, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J., Esquivel, S.L.: Practical considerations for using exploratory factor analysis in educational research. Pract. Assess. Res. Eval. **18**(6), 1–13 (2013)