Fausto Pedro García Márquez
Benjamin Lev *Editors*

Big Data Management

Springer

Big Data Management

Fausto Pedro García Márquez
Benjamin Lev
Editors

Big Data Management

Springer

Editors
Fausto Pedro García Márquez
ETSI Industriales de Ciudad Real
University of Castilla-La Mancha
Ciudad Real
Spain

Benjamin Lev
Drexel University
Philadelphia, PA
USA

ISBN 978-3-319-45497-9 ISBN 978-3-319-45498-6 (eBook)
DOI 10.1007/978-3-319-45498-6

Library of Congress Control Number: 2016949558

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

This book is dedicated to
my beloved wife of 51 years Debbie Lev
2/12/1945–4/16/2016

Benjamin Lev

Preface

Big Data and Management Science has been designed to synthesize the analytic principles with business practice and Big Data. Specifically, the book provides an interface between the main disciplines of engineering/technology and the organizational, administrative, and planning abilities of management. It is complementary to other sub-disciplines such as economics, finance, marketing, decision and risk analysis.

This book is intended for engineers, economists, and researchers who wish to develop new skills in management or for those who employ the management discipline as part of their work. The authors of this volume describe their original work in the area or provide material for case studies that successfully apply the management discipline in real-life situations where Big Data is also employed.

The recent advances in handling large data have led to increasingly more data being available, leading to the advent of Big Data. The volume of Big Data runs into petabytes of information, offering the promise of valuable insight. Visualization is the key to unlocking these insights; however, repeating analytical behaviors reserved for smaller data sets runs the risk of ignoring latent relationships in the data, which is at odds with the motivation for collecting Big Data. Chapter "Visualizing Big Data: Everything Old Is New Again" focuses on commonly used tools (SAS, R, and Python) in aid of Big Data visualization to drive the formulation of meaningful research questions. It presents a case study of the public scanner database Dominick's Finer Foods, containing approximately 98 million observations. Using graph semiotics, it focuses on visualization for decision making and explorative analyses. It then demonstrates how to use these visualizations to formulate elementary-, intermediate-, and overall-level analytical questions from the database.

The development of Big Data applications is closely linked to the availability of scalable and cost-effective computing capacities for storing and processing data in a distributed and parallel fashion, respectively. Cloud providers already offer a portfolio of various cloud services for supporting Big Data applications. Large companies such as Netflix and Spotify already use those cloud services to operate

their Big Data applications. Chapter "Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective" proposes a generic reference to architecture that implements Big Data applications based on state-of-the-art cloud services. The applicability and implementation of our reference to architecture is demonstrated for three leading cloud providers. Given these implementations, we analyze how main pricing schemes and cost factors can be used to compare respective cloud services. This information is based on a Big Data streaming use case. Derived findings are essential for cloud-based Big Data management from a cost perspective.

Most of the information about Big Data has focused on the technical side of the phenomenon. Chapter "The Strategic Business Value of Big Data" makes the case that business implications of utilizing Big Data are crucial to obtain a competitive advantage. To achieve such objective, the organizational impacts of Big Data for today's business competition and innovation are analyzed in order to identify different strategies a company may implement, as well as the potential value that Big Data can provide for organizations in different sectors of the economy and different areas inside such organizations. In the same vein, different Big Data strategies a company may implement toward its development are stated and suggestions regarding how enterprises such as businesses, nonprofits, and governments can use data to gain insights and make more informed decisions. Current and potential applications of Big Data are presented for different private and public sectors, as well as the ability to use data effectively to drive rapid, precise and profitable decisions.

Chapter "A Review on Big Data Security and Privacy in Healthcare Applications" considers the term Big Data and its usage in healthcare applications. With the increasing use of technologically advanced equipment in medical, biomedical, and healthcare fields, the collection of patients' data from various hospitals is also becoming necessary. The availability of data at the central location is suitable so that it can be used in need of any pharmaceutical feedback, equipment's reporting, analysis and results of any disease, and many other uses. Collected data can also be used for manipulating or predicting any upcoming health crises due to any disaster, virus, or climate change. Collection of data from various health-related entities or from any patient raises serious questions upon leakage, integrity, security, and privacy of data. The questions and issues are highlighted and discussed in the last section of this chapter to emphasize the broad pre-deployment issues. Available platforms and solutions are also discussed to overcome the arising situation and question the prudence of usage and deployment of Big Data in healthcare-related fields and applications. The available data privacy, data security, users' accessing mechanisms, authentication procedures, and privileges are also described.

Chapter "What Is Big Data" consists of three parts. The first section describes what Big Data is, the concepts of Big Data, and how Big Data arose. Big Data affects scientific schemes. It considers the limitations of predictions by using Big Data and a relation between Big Data and hypotheses. A case study considers an

electric power of Big Data systems. The next section describes the necessity of Big Data. This is a view that applies aspects of macroeconomics. In service science capitalism, measurements of values of products need Big Data. Service products are classified into stock, flow, and rate of flow change. Immediacy of Big Data implements and makes sense of each classification. Big Data provides a macroeconomic model with behavioral principles of economic agents. The principles have mathematical representation with high affinity of correlation deduced from Big Data. In the last section, we present an explanation of macroeconomic phenomena in Japan since 1980 as an example of use of the macroeconomic model.

Chapter "Big Data for Conversational Interfaces: Current Opportunities and Prospects" is on conversational technologies. As conversational technologies develop, more demands are placed upon computer-automated telephone responses. For instance, we want our conversational assistants to be able to solve our queries in multiple domains, to manage information from different usually unstructured sources, to be able to perform a variety of tasks, and understand open conversational language. However, developing the resources necessary to develop systems with such capabilities demands much time and effort. For each domain, task, or language, data must be collected and annotated following a schema that is usually not portable. The models must be trained over the annotated data, and their accuracy must be evaluated. In recent years, there has been a growing interest in investigating alternatives to manual effort that allow exploiting automatically the huge amount of resources available in the Web. This chapter describes the main initiatives to extract, process, and contextualize information from these Big Data rich and heterogeneous sources for the various tasks involved in dialog systems, including speech processing, natural language understanding, and dialog management.

In Chapter "Big Data Analytics in Telemedicine : A Role of Medical Image Compression," Big Data analytics which is one of most rapidly expanding fields has started to play a vital role in the field of health care. A major goal of telemedicine is to eliminate unnecessary traveling of patients and their escorts. Data acquisition, data storage, data display and processing, and data transfer represent the basis of telemedicine. Telemedicine hinges on transfer of text, reports, voice, images, and video between geographically separated locations. Out of these, the simplest and easiest is through text, as it is quick and simple to use, since sending text requires very little bandwidth. The problem with images and videos is that they require a large amount of bandwidth for transmission and reception. Therefore, there is a need to reduce the size of the image that is to be sent or received, i.e., data compression is necessary. This chapter deals with employing prediction as a method for compression of biomedical images. The approach presented in this chapter offers great potential in compression of the medical image under consideration, without degrading the diagnostic ability of the image.

A Big Data network design with risk-averse signal control optimization (RISCO) is considered to regulate the risk associated with hazmat transportation and

minimize total travel delay. A bi-level network design model is presented for RISCO subject to equilibrium flow. A weighted sum risk equilibrium model is proposed in Chapter "A Bundle-Like Algorithm for Big Data Network Design with Risk-Averse Signal Control Optimization" to determine generalized travel cost at lower level problem. Since the bi-objective signal control optimization is generally non-convex and non-smooth, a bundle-like efficient algorithm is presented to solve the equilibrium-based model effectively. A Big Data bounding strategy is developed in Chapter "A Bundle-Like Algorithm for Big Data Network Design with Risk-Averse Signal Control Optimization" to stabilize solutions of RISCO with modest computational efforts. In order to investigate the computational advantage of the proposed algorithm for Big Data network design with signal optimization, numerical comparisons using real data example and general networks are made with current best well-known algorithms. The results strongly indicate that the proposed algorithm becomes increasingly computationally comparative to best known alternatives as the size of network grows.

Chapter "Evaluation of Evacuation Corridors and Traffic Management Strategies for Short-Notice Evacuation" presents a simulation study of the large-scale traffic data under a short-notice emergency evacuation condition due to an assumed chlorine gas spill incident in a derailment accident in the Canadian National (CN) Railway's railroad yard in downtown Jackson, Mississippi by employing the dynamic traffic assignment simulation program DynusT. In the study, the effective evacuation corridor and traffic management strategies were identified in order to increase the number of cumulative vehicles evacuated out of the incident-affected protective action zone (PAZ) during the simulation duration. An iterative three-step study approach based on traffic control and traffic management considerations was undertaken to identify the best strategies in evacuation corridor selection, traffic management method, and evacuation demand staging to relieve heavy traffic congestions for such an evacuation.

Chapter "Analyzing Network Log Files Using Big Data Techniques" considers the service to 26 buildings with more than 1000 network devices (wireless and wired) and access to more than 10,000 devices (computers, tablets, smartphones, etc.) which generate approximately 200 MB/day of data that is stored mainly in the DHCP log, the Apache HTTP log, and the Wi-fi log files. Within this context, Chapter "Analyzing Network Log Files Using Big Data Techniques" addresses the design and development of an application that uses Big Data techniques to analyze those log files in order to track information on the device (date, time, MAC address, and georeferenced position), as well as the number and type of network accesses for each building. In the near future, this application will help the IT department to analyze all these logs in real time.

Finally, Chapter "Big Data and Earned Value Management in Airspace Industry" analyzes earned value management (EVM) for project management. Actual cost and earned value are the parameters used for monitoring projects. These parameters are compared with planned value to analyze the project status. EVM covers scope, cost,

and time and unifies them in a common framework that allows evaluation of project health. Chapter "Big Data and Earned Value Management in Airspace Industry" aims to integrate the project management and the Big Data. It proposes an EVM approach, developed from a real case study in aerospace industry, to simultaneously manage large numbers of projects.

Ciudad Real, Spain — Fausto Pedro García Márquez
Philadelphia, PA, USA — Benjamin Lev

Contents

About the Editors

Prof. Fausto Pedro García Márquez obtained his European Doctorate in 2004 at the University of Castilla-La Mancha (UCLM), Spain with the highest distinction. He has been honored with the Runner-up Prize (2015) Advancement Prize (2013), and Silver Prize (2012) by the International Society of Management Science and Engineering Management. He is a senior lecturer at UCLM (with tenure, accredited as full professor), honorary senior research fellow at Birmingham University, UK, lecturer at the Postgraduate European Institute, and he was senior manager in Accenture (2013–14). Fausto has managed a great number of projects as either principal investigator (PI) or researcher: five Europeans and four FP7 framework program (one Euroliga, three FP7); he is PI in two national projects, and he has participated in two others; four regional projects; three university projects; and more than 100 joint projects with research institutes and industrial companies (98 % as director). He has been a reviewer in national and international programs. He has published more than 150 papers (65 % in ISI journals, 30 % in JCR journals, and 92 % internationals), being the main author of 68 publications. Some of these papers have been especially recognized, e.g., by "Renewable Energy" (as "Best Paper Award 2014"); "International Society of Management Science and Engineering Management" (as "excellent"), and by the "International Journal of Automation and Computing" and "IMechE Part F: Journal of Rail and Rapid Transit" (most downloaded). He is the author/editor of 18 books (published by Elsevier, Springer, Pearson, McGraw-Hill, Intech, IGI, Marcombo, and AlfaOmega), and he is the inventor of five patents. He is an associate editor of three international journals: Engineering Management Research; Open Journal of Safety Science and Technology; and International Journal of Engineering and Technologies, and he has been a committee member of more than 25 international conferences. He is a director of Ingenium Research Group (www.uclm.es/profesorado/fausto).

Prof. Benjamin Lev is a trustee professor of DS&MIS at LeBow College of Business, Drexel University in Philadelphia, PA. He holds a Ph.D. in operations research from Case Western Reserve University in Cleveland, OH. Prior to joining Drexel University, Dr. Lev held academic and administrative positions at Temple University, University of Michigan-Dearborn, and Worcester Polytechnic Institute. He is the editor in chief of OMEGA, The International Journal of Management Science; co-editor in chief of International Journal of Management Science and Engineering Management; and has served and currently serves on several other journal editorial boards such as JOR, ORP, IIE-Transactions, ERRJ, Interfaces, and IAOR. Dr. Lev has published/edited thirteen books, numerous articles, and organized national and international INFORMS and IFORS conferences.

Visualizing Big Data: Everything Old Is New Again

Belinda A. Chiera and Małgorzata W. Korolkiewicz

Abstract Recent advances have led to increasingly more data being available, leading to the advent of Big Data. The volume of Big Data runs into petabytes of information, offering the promise of valuable insight. Visualization is key to unlocking these insights, however repeating analytical behaviors reserved for smaller data sets runs the risk of ignoring latent relationships in the data, which is at odds with the motivation for collecting Big Data. In this chapter, we focus on commonly used tools (SAS, R, Python) in aid of Big Data visualization, to drive the formulation of meaningful research questions. We present a case study of the public scanner database Dominick's Finer Foods, containing approximately 98 million observations. Using graph semiotics, we focus on visualization for decision-making and explorative analyses. We then demonstrate how to use these visualizations to formulate elementary-, intermediate- and overall-level analytical questions from the database.

Keywords Visualisation · Big Data · Graph semiotics · Dominick's Finer Foods (DFF)

1 Introduction

Recent advances in technology have led to more data being available than ever before, from sources such as climate sensors, transaction records, cell phone GPS signals, social media posts, digital images and videos, just to name a few. This phenomenon is referred to as 'Big Data'. The volume of data collected runs into petabytes of information, thus allowing governments, organizations and researchers to know much more about their operations, thus leading to decisions that are increasingly based on data and analysis, rather than experience and intuition [1].

B.A. Chiera (✉) · M.W. Korolkiewicz
University of South Australia, Adelaide, Australia
e-mail: belinda.chiera@unisa.edu.au

M.W. Korolkiewicz
e-mail: malgorzata.korolkiewicz@unisa.edu.au

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_1

Big Data is typically defined in terms of its *Variety, Velocity* and *Volume*. Variety refers to expanding the concept of data to include unstructured sources such as text, audio, video or click streams. Velocity is the speed at which data arrives and how frequently it changes. Volume is the size of the data, which for Big Data typically means 'large', given how easily terabytes and now petabytes of information are amassed in today's market place.

One of the most valuable means through which to make sense of Big Data is visualization. If done well, a visual representation can uncover features, trends or patterns with the potential to produce actionable analysis and provide deeper insight [2]. However Big Data brings new challenges to visualization due to its speed, size and diversity, forcing organizations and researchers alike to move beyond well-trodden visualization paths in order to derive meaningful insights from data. The techniques employed need not be new—graphs and charts can effectively be those decision makers are accustomed to seeing—but a new way to look at the data will typically be required.

Additional issues with data volume arise when current software architecture becomes unable to process huge amounts of data in a timely manner. Variety of Big Data brings further challenges due to unstructured data requiring new visualization techniques. In this chapter however, we limit our attention to visualization of 'large' structured data sets.

There are many visualization tools available; some come from established analytics software companies (e.g. Tableau, SAS or IBM), while many others have emerged as open source applications.[1] For the purposes of visualization in this chapter, we focus on SAS, R and Python, which together with Hadoop, are considered to be key tools for Data Science [3].

The use of visualization as a tool for data exploration and/or decision-making is not a new phenomenon. Data visualization has long been an important component of data analysis, whether the intent is that of data exploration or as part of a model building exercise. However the challenges underlying the visualization of Big Data are still relatively new; often the choice to visualize is between simple graphics using a palette of colors to distinguish information or to present overly-complicated but aesthetically pleasing graphics, which may obfuscate and distort key relationships between variables.

Three fundamental tenets underlie data visualization: (1) visualization for data exploration, to highlight underlying patterns and relationships; (2) visualization for decision making; and (3) visualization for communication. Here we focus predominantly on the first two tenets. In the case of the former, previous work in the literature suggests a tendency to approach Big Data by repeating analytical behaviors typically reserved for smaller, purpose-built data sets (e.g. [4–6]). There appears, however, to be less emphasis on the exploration of Big Data itself to formulate questions that drive analysis.

[1] http://www.tableau.com, http://thenextweb.com/dd/2015/04/21/the-14-best-data-visualization-tools/, http://opensource.com/life/15/6/eight-open-source-data-visualization-tools.

In this chapter we propose to redress this imbalance. While we will lend weight to the use of visualization of Big Data in support of good decision-making processes, our main theme will be on visualization as a key component to harnessing the scope and potential of Big Data to drive the formulation of meaningful research questions. In particular, we will draw upon the seminal work of Bertin [7] in the use of graph semiotics to depict multiple characteristics of data. We also explore the extension of this work [8] to apply these semiotics according to data type and perceived accuracy of data representation and thus perception. Using the publicly available scanner database Dominick's Finer Foods, containing approximately 98 million observations, we demonstrate the application of these graph semiotics [7, 8] for data visualization. We then demonstrate how to use these visualizations to formulate elementary-, intermediate- and overall-level analytical questions from the database, before presenting our conclusions.

2 Case Study Data: Dominick's Finer Foods

To illustrate Big Data visualization, we will present a case study using a publicly available scanner database from Dominick's Finer Foods[2] (DFF), a supermarket chain in Chicago. The database has been widely used in the literature, ranging from consumer demand studies through to price point and rigidity analysis, as well as consumer-preferences studies. The emphasis in the literature has been on using the data to build analytical models and drive decision-making processes based on empirically driven insights [4–6, 9–12].[3]

The DFF database contains approximately nine years of store-level data with over 3,500 items, all with Unique Product Codes (UPC). Data is sampled weekly from September 1989 through to May 1997, totaling 400 weeks of scanner data and yielding approximately 98 million observations [13]. The sample is inconsistent in that there is missing data and data that is non-homogeneous in time, for a selection of supermarket products and characteristics. The database is split across 60 files, each of which can be categorized broadly as either:

1. **General files**: files containing information on store traffic such as coupon usage and store-level population demographics (cf. Table 1); and
2. **Category-specific files**: grocery items are broadly categorized into one of 29 categories (e.g. *Analgesics, Bath Soap, Beer*, and so forth) and each item category is associated with a pair of files. The first file of the pair contains product description information such as the name and size of the product and UPC, for all brands of that specific category. The second file contains movement information for each UPC, pertaining to weekly sales data including *store, item price, units sold, profit margin, total dollar sales* and *coupons redeemed* (cf. Table 2).

[2] http://edit.chicagobooth.edu/research/kilts/marketing-databases/dominicks/dataset.

[3] An expanded list of literature analyzing the Dominick's Finer Foods Database can be found at https://research.chicagobooth.edu/kilts/marketing-databases/dominicks/papers.

Table 1 Sample of customer information recorded in the DFF database. Coupon information was recorded only for those products offering coupon specials

Variable	Description	Type
Date	Date of observation	Date (yymmdd)
Week	Week number	Quantitative
Store	Unique store ID	Quantitative
Bakcoup	Bakery coupons redeemed	Quantitative
Bakery	Bakery sales in dollars	Quantitative
Beer	Beer sales in dollars	Quantitative
…	…	…
Wine	Wine sales in dollars	Quantitative

Table 2 Sample of demographic information recorded in the DFF database. A total of 510 unique variables comprise the demographic data. This brief excerpt gives a generalized overview

Variable	Description	Type
Name	Store name	Qualitative
City	City in which store is located	Qualitative
Zone	Geographic zone of store	Quantitative
Store	Unique store ID	Quantitative
Age60	Percentage of population over 60	Quantitative
Hsizeavg	Average household size	Quantitative
…	…	…

In total, across the general and category-specific files there are 510 store-specific demographic variables and 29 item categories recorded as 60 separate data sets. A 524-page data manual and codebook accompanies the database. Given the amount of information recorded in the database, we are able to present a *breadth-and-depth* data overview. Specifically, we will demonstrate data visualization across a range of characteristics of a single supermarket product, to provide a summary of the breadth of the data set, as well as an in-depth analysis of beer products to demonstrate the ability of visualization to provide further insight into big databases.

3 Big Data: Pre-processing and Management

Prior to visualization, the database needs to be checked for inconsistencies and, given the disparate nature of the recorded data, merged in a meaningful way for informative visualization. Unlike smaller databases however, any attempt to view the data in its raw state will be overwhelmed by the volume of information available, due to the

prohibitive size of Big Data. What is ordinarily a rudimental step of any statistical analysis — checking data validity and cleaning — is now a difficult exercise, fraught with multiple challenges. Thus alternative approaches need to be adopted to prepare the data for visualization.

The two main areas which need to be addressed at this stage are:

1. Data pre-processing; and
2. Data management.

Of the three software platforms considered here, the Python programming language provides tools which are both flexible and fast, to aid in both data pre-processing and manipulation, a process referred to as either *data munging* or *wrangling*.

Data munging encompasses the process of data manipulation from cleaning through to data aggregation and/or visualization. Key to the success of any data munging exercise is the flexibility provided by the software to manipulate data. To this end, Python contains the specialized *Pandas* library (PANel DAta Structures), which provides the *data frame* structure. A data frame allows for the creation of a data table, mirroring e.g. Tables 1 and 2, in that variables of mixed type are stored within a single structure.

The advantage of using a Python data frame is that this structure allows for data cleaning, merging and/or concatenation, plus data summarization or aggregation, with a necessarily fast and simple implementation. It should be noted that R, and to an extent SAS, also offer a data frame structure which is equally easy to use and manipulate, however Python is preferred for Big Data as the Pandas data frame is implemented using a programming construct called *vectorization,* which allows for faster data processing over non-vectorized data frames [14]. R also provides vectorization functionality, however it is somewhat more complicated to use than Pandas. For this reason we will use Python for data pre-processing and management.

3.1 Data Pre-processing

We first addressed the general store files individually (Tables 1 and 2) to perform an initial investigation of the data. The two files were read into separate Python data frames, named *ccount* and *demo*, and were of dimension 324,133 × 62 and 108 × 510 respectively, with columns indicating unique variables and rows giving observations over those variables. A sample of each data frame was viewed to compare the database contents with the data manual, at which time it was determined that the store data was not perfectly mirrored in the manual. Any variable present in the database that did not appear in the manual was further investigated to resolve ambiguity around the information recorded, and if a resolution could not be achieved, the variable was removed from the analysis.

Rather than pre-process the two complete data frames, we elected to remove columns not suitable, or not of immediate interest, for visualization. Given an end goal was to merge disparate data frames to form a cohesive database for visualization,

we identified common variables appearing in *ccount* and *demo* and used these variables for the merging procedures, as will be discussed in what follows. We removed missing values using Python's *drop.na()* function, which causes the listwise removal for any record containing at least one missing value. We opted for listwise deletion since the validation of imputed data would be difficult due to inaccessibility of the original data records, and sample size was not of concern. Other operations performed included the removal of duplicate records and trailing whitespace characters in variable names, since statistical software could potentially treat these whitespaces as unique identifiers, and introduce erroneous qualitative categories.

In what follows, rather than attempt to read the database as a whole, category-specific files for a selection of products were processed and held in computer memory only as needed. All three software applications considered here—SAS, Python and R—are flexible and allow easy insertion of data, thus supporting the need to keep the data frames as small as possible, with efficient adaptation on-the-fly.

We focused on a single product item for visualization, in this case *Beer*, as given the scope of the data there was a plethora of information available for a single product and the data was more than sufficient for our purposes here. Each product was represented by two files, as was the case for Beer. The first captured information such as the Unique Product Code, name, size and item coding. The second file contained movement information indicating price, profit margins, sales amounts and codes, as well as identifying information such as the store ID and the week in which the data was recorded. We elected to use the movement data only, since: (1) the information contained therein was more suited to meaningful visualization; and (2) the information in the movement data overlapped with the *ccount* and *demo* data frames, namely the variables *store* (a number signifying the store ID) and *week*, allowing for potential merging of the data frames. Finally, a map was made available on the DFF website, containing geographic information of each store (City, Zone, Zip code) as well as the ID and the *Price Tier* of each store, indicating the perceived socio-economic status of the area in which each store was located. In what follows, *Store, Zone* and *Price Tier* were retained for the analysis.

3.2 Data Management

As previously noted, the DFF database is comprised of 60 separate files. While the data in each file can be analyzed individually, there is also much appeal in meaningfully analyzing the data as a whole to obtain a big-picture overview across the stores. However given the full database contains over 98 million observations, the practicalities of how to analyze the data as a whole becomes a matter of data manipulation and aggregation. The initial pre-processing described above is the first step towards achieving this goal, as it provides flexible data structures for manipulation, while the management process for creating and/or extracting data for visualization forms the second step.

An attraction of using Python and R is that both languages allow the manipulation of data frames in the same manner as a database. We continued to work in Python during the data management phase for reasons cited above, namely the fast implementation of Python structures for data manipulation. It should be noted however, the functionality discussed below applies to both Python and R. While not all database functionality is implemented, key operations made available include:

- **concat**: appends columns or rows from one data frame to another. There is no requirement for a common variable in the data frames.
- **merge**: combines data frames by using columns in each dataset that contain common variables.
- **groupby**: provide a means to easily generate data summaries over a specified characteristic.

The database-style operation *concat* concatenates two data frames by adding rows and/or columns, the latter occurring when data frames are merged and each structure contains no variables in common. Python will automatically insert missing values into the new data frame when a particular row/column combination has not been recorded, to pad out the concatenated structure. Thus care needs to be taken when treating missing values in a concatenated data frame—a simple call to *drop.na()* can at times lead to an empty data frame. It is suggested that only those variables immediately of interest for visualization should be treated for missing data.

The *merge* operation joins two or more data frames on the basis of at least one common variable—called a *join key*—in the data frames [14]. For example, the data frames *ccount* and *demo* both contain the numerical variable *store*, which captures each Store ID. Merging the *demo* and *ccount* data frames would yield an expanded data frame in which the observations for each store form a row in the data frame while the variables for each store form the columns.

There is some flexibility as to how to join data frames, namely *inner* and *outer* joins. An *inner* join will merge only those records which correspond to the same value of the join key in the data frame. For example, while *Store* appears in *ccount* and *demo*, not every unique store ID necessarily appears in both data frames. An inner join on these data frames would merge only those records for which the store ID appears in both *ccount* and *demo*. Other inner join operations include merging data by retaining all information in one data frame and extending it by adding data from the second data frame, based on common values of the join key. For example, the *demo* data frame can be extended by adding columns of variables from *ccount* that do not already appear in *demo*, for all store IDs common to both data frames. Python also offers a full join, in which a Cartesian combination of data frames is produced. Such structures can grow quite rapidly in size and given the prohibitive nature of Big Data, we opted to use an inner join to reduce computational overhead.

Data summarization can take place via the *groupby* functionality, on either the original or merged/concatenated data frames. For example, if it is of interest to compute the total product profits per store, a *groupby* operation will efficiently perform this aggregation and calculation, thereby producing a much smaller data structure

which can then be visualized. The *groupby* operation also has the flexibility to group at multiple levels simultaneously. For example, it might be of interest to group by the socioeconomic status of the store location, and then for each store in each of the socioeconomic groups, compute store profits. Providing the data used to define the levels of aggregation can be treated as categorical, *groupby* can perform any of these types of aggregation procedures in a single calculation. As *groupby* is defined over the Python data frame structure, this operation is performed quickly over large amounts of data.

It should be noted that SAS also provides database-style support for data manipulation through Structured Query Language (SQL), which is a widely-used language for retrieving and updating data in tables and/or views of those tables. PROC SQL is the SQL implementation within the SAS system. Prior to the availability of PROC SQL in Version 6.0 of the SAS System, DATA step logic and several utility procedures were the only tools available for creating, joining, sub-setting, transforming and sorting data. Both non-SQL base SAS techniques or PROC SQL can be utilized for the purposes of creating new data sets, accessing relational databases, sorting, joining, concatenating and match-merging data, as well as creating new and summarizing existing variables. The choice of approach—PROC SQL or DATA step—depends on the nature of the task at hand and could be also accomplished via the so-called Query Builder, one of the most powerful 'one stop shop' components of the SAS® Enterprise Guide user interface.

4 Big Data Visualization

The challenge of effective data visualization is not new. From as early as the 10th century data visualization techniques have been recorded, many of which are still in use in the current day, including time series plots, bar charts and filled-area plots [15]. However in comparatively more recent years, the perception of effective data visualization as being not only a science, but also an art form, was reflected in the seminal work on graph semiotics [7] through to later adaptations in data visualization [8, 16, 17]. Further influential work on statistical data displays was explored in [15] with an emphasis on avoiding data distortion through visualization, to more recent approaches [18] in which a tabular guide of 100 effective data visualization displays, based on data type, has been presented.

4.1 Visualization Semiotics

Data visualization embodies at least two distinct purposes: (1) to communicate information meaningfully; and (2) to "*solve a problem*" [7]. It is defensible to suggest that 'solving a problem' in the current context is to answer and/or postulate questions about (big) data from visualization, as was the approach adopted in the originating

Table 3 Retinal variables for the effective communication of data visualization [7, 19]

Variable	Description	Best for data type
Position	Position of graphing symbol relative to axes	Quantitative, qualitative
Size	Space occupied by graphing symbol	Quantitative, qualitative
Color value	Varied to depict weight/size of observation	Quantitative differences
Texture	Fill pattern within the data symbol	Qualitative, quantitative differences
Color hue	Graduated RGB color to highlight differences	Qualitative differences
Orientation	Used to imply direction	Quantitative
Shape	Graphic symbol representing data	Quantitative

work [7]. It is thus in the same spirit we adopt graphic semiotics and reference the fundamental data display principles, in the visualization that follows.

At the crux of the works on visualization and graphic semiotics are the retinal variables identified in [7]. These variables are manipulated to encode information from data for effective communication via visualization (Table 3) with application to data type as indicated [19].

The usefulness of the retinal variables was experimentally verified in subsequent research [20]. The authors focused solely on the accurate perception of visualization of quantitative data and developed a ranking system indicating the accuracy with which these variables were perceived. The variables *Position* and *Size* were the most accurately understood in data visualizations, whilst *Shape* and *Color* were the least accurate, with area-based shapes somewhat more accurate than volume-based shapes [20]. This work was later extended to include qualitative data in the heavily cited research of [8], in which further distinction was made between the visualization of ordinal and nominal categorical variables and is an approach which we adopt here. The revised ordering, including an extended list of retinal variables and their associated accuracy, is depicted in Fig. 1. The extended list centers around the original retinal variables introduced in [7]—for example *Shape* was extended to consider area- and volume-based representations while *Color* was considered in terms of saturation and hue.

The retinal variables are typically related to the components of the data that are to be visualized. Even from the smaller set of the event retinal variables in Table 3, there is a large choice of possible graphical constructions, with the judicious selection of several retinal variables to highlight data characteristics being perceived as more effective than use of the full set [7]. In the visualizations presented here, we opt to avoid the use of different colors and instead focus on color hue and saturation. Often in the literature there is a restriction to grayscale printing; we wish to demonstrate the ability to effectively visualize aspects of Big Data in these circumstances.

A motivation for forming data visualizations is the construction and/or answering of questions about the data itself. It was suggested in [7] that any question about data can be defined firstly by its *type* and secondly by its *level*. In terms of question

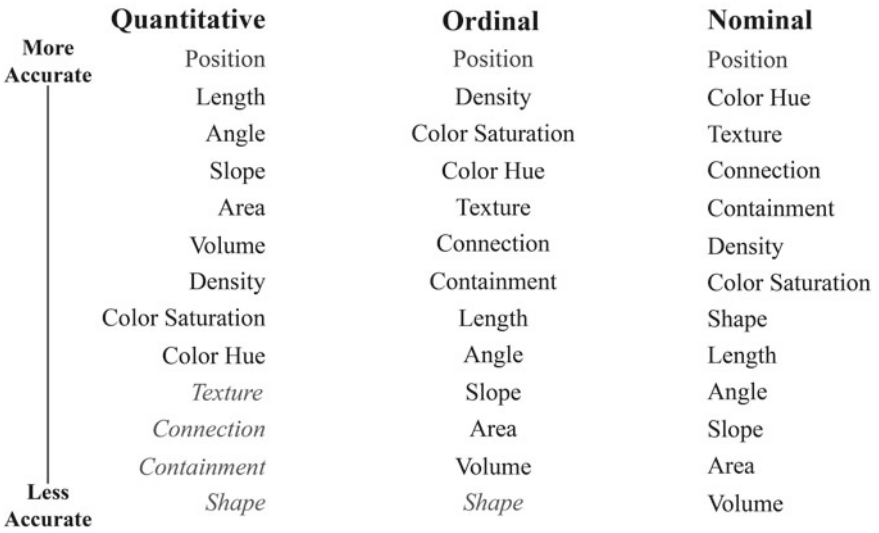


Fig. 1 Accuracy of the perception of the retinal variables by data type [8]. *Position* is the most accurate for all data types, whilst items in gray are not relevant to the specified data type

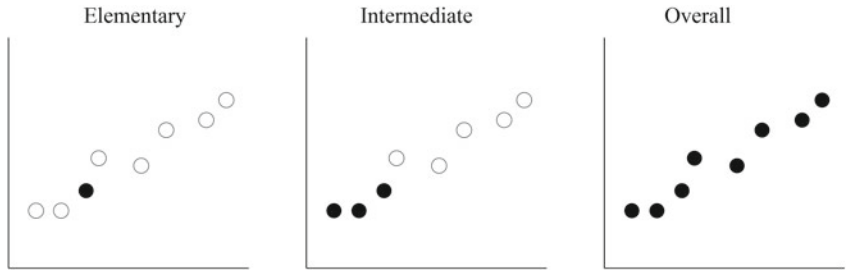


Fig. 2 Elementary-, intermediate- and overall-level questions, based on data [7]. The *filled circles* indicate the number of data points involved in the answer to each question type

type, the suggestion is that there are at least as many *types* of questions as physical dimensions used to construct the graphic in the first place. However, [7] derived these conclusions on the basis of temporal data only, while [21] demonstrated that for spatio-temporal data, the distinction between question types can be independently applied to both the temporal and spatial dimensions of the data.

Questions about the data can be defined at an *elementary-*, *intermediate-* or *overall-level* [7]. From Fig. 2 it can be seen that the answer to elementary-level questions results in a single item of the data (e.g. product sales on a given day), answers to intermediate-level questions typically involve at least several items (e.g. product sales over the past 3 days) while overall-level questions are answered in terms of the entire data set (e.g. what was the trend of the product sales over the entire period?).

Questions combining spatio-temporal scales could be phrased as, e.g. *What is the trend of sales for Store 2?*, which is elementary-level with regards to the spatial component, but an overall-level question with regards to the temporal component. We will further elucidate on these question types in what follows, in which we present visualization of the DFF Database by question level as defined in [7] and indicate combinations between spatio- and temporal-scales as appropriate.

4.2 Visualization of the DFF Database

To achieve data visualization in practical terms, all three software systems considered here (SAS, Python, R) allow for graphical representation of data, however while Python is the preferred choice for data wrangling, we have selected R and SAS for data visualization, as they offer a comprehensive suite of graphical display options that are straightforward to implement. In contrast, data visualization in Python is typically not as straightforward and in practice it has been the case that, e.g. several lines of code in R are reproduced by over tens of lines of code in Python. Although Python generally boasts faster processing speeds due to the data frame structure, once data has been pre-processed as described in Sect. 3, the resulting data set is typically much smaller than the original, thus R and SAS are able to produce graphics with efficiency.

The introduction of Statistical Graphics (SG) Procedures and Graph Template Language (GTL) as part of the ODS Graphics system in SAS® 9.2 has been a great leap forward for data presentation using SAS. The SG procedures provide an easy to use, yet flexible syntax to create most commonly-used graphs for data analysis and presentation in many domains, with visualization options including the SGPLOT, SGSCATTER and SGPANEL procedures. In subsequent SAS releases more features were added that make it easy to customize graphs, including setting of group attributes, splitting axis values, jittering, etc. for SAS users of all skill levels.

The Graph Template Language allows the creation of complex multi-cell graphs using a structured syntax, and thus provides highly flexible ways to define graphs that are beyond the abilities of the SG procedures. Alternatively, SAS now offers SAS® Visual Analytics, which is an autocharting solution with in-memory processing for accelerated computations, aimed at business analysts and non-technical users. In this chapter, SG procedures and GTL were used in SAS® Enterprise Guide to generate selected visualizations.

To transfer data between Python and R for visualization, the options provided by Python include: (1) saving the Python data frame structures to file, which is then read into R in an identical data frame structure; and (2) direct communication with R from within Python via the *rpy2* interface. While the latter approach is more elegant and reduces computational overhead, the *rpy2* library is poorly supported across computing platforms and for this reason we have opted for the former approach. However it should be noted that as the merged data frames are the results of data wrangling and management rather than the raw data, these files are typically smaller

and thus manageable for file input/output operations. It is also worth noting that R has the facility to read SAS files, and the latest releases of SAS include the facility to process R code, for added flexibility.

There are four primary graphical systems in R: *base, grid, lattice* and *ggplot2*, each of which offers different options for data visualization:

1. **base**: produces a basic range of graphics that are customizable;
2. **grid**: offers a lower-level alternative to the base graphics system to create arbitrary rectangular regions. Does not provide functions for producing statistical graphics or complete plots;
3. **lattice**: implements trellis graphics to provide high-level data visualization to highlight meaningful parts of the data. Useful for visualizing data that can be naturally grouped; and
4. **ggplot2**: creates graphics using a *layering* approach, in which elements of the graph such as points, shape, color etc. are layered on top of one another. Highly customizable and flexible.

Of the four graphic systems, only the library *ggplot2* needs to be explicitly installed in R, however this is readily achieved through the in-built installation manager in R with installation a once-for-all-time operation, excepting package updates. In this work we have predominantly used the *ggplot2* library to produce data visualizations as well as the *lattice* library. The lattice package allows for easy representation of time series data, while the layering approach of ggplot2 naturally corresponds with the retinal variables for visualization [7, 8].

Next we present a small selection of elementary-level questions to highlight the use of the retinal variables [7, 8], as these types of questions are the most straightforward to ask and resolve, even with regards to Big Data. The bulk of the visualization following elementary-level questions will be on the most challenging aspects with regards to visualization; intermediate- and overall-level questions.

4.2.1 Elementary-Level Question Visualizations

To produce an elementary-level question from a visualization, the focus on the data itself needs to be specific and quite narrow. In this regard, it is straightforward to produce one summary value of interest, per category.

For example, Fig. 3 shows a dot plot summarizing the total dollar sales of Beer in a selection of 30 stores in week 103 of the database. For this seemingly straightforward graphic, there are a number of retinal variables at play, including Position, Color Hue and Color Saturation. We recall there are at least as many types of questions to be asked as physical dimensions used to construct the plot [7] and as the temporal quantity is fixed (week 103), we can adopt this approach over the spatial domain [21]. Thus sample questions could be: *What is the total dollar sales of beer in Store 131?* or *Which store has the maximum total dollar sales of beer?* The former is a very specific question requiring some level of knowledge of the database whereas the latter is purely exploratory in spirit and would be a typical question of interest.

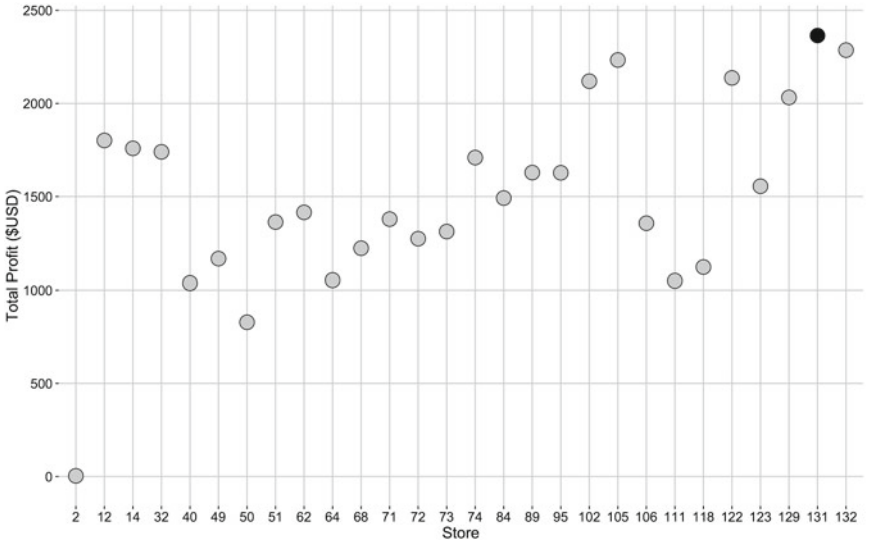


Fig. 3 Dot plot of beer sales by store

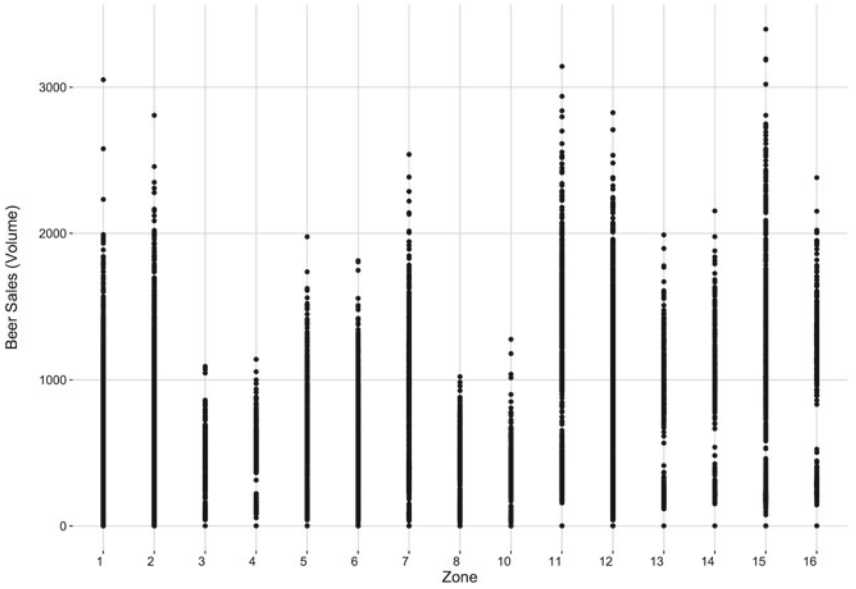


Fig. 4 Rainfall plot for elementary questions regarding beer sales by zone

In Fig. 4, a rainfall plot of the beer sales data is shown as a natural extension to the dot plot of Fig. 3, using the retinal variables Position and Length [22]. Color saturation is now used for aesthetic purposes and not to highlight information, as

was the case in Fig. 3. All stores have been included in Fig. 4 for all weeks of the data set, however the introduction of the qualitative variable Zone also increases the opportunity to pose informative, elementary questions, e.g. *Which zone has the largest variability in beer sales?* or *Which zone sells the largest volume of beer?*

On the other hand, Fig. 5 depicts a somewhat more intricate 100 % stacked bar chart of the average beer price in each store over time, with the chart showing two types of categorical information: the year and the Price Tier of all stores. It is now also possible to observe missing data in the series, represented as vertical white bars, with four weeks starting from the second week of September in 1994 and the last two weeks of February in 1995. In our exploration of the database we noted that the same gaps appear in records for other products, suggesting a systematic reason for reduced or no trading at Dominick's stores during those weeks.

The retinal variables used in Fig. 5 include Position, Color Saturation, Color Hue and Length. The graph varies over both the temporal and spatial dimensions and besides conveying information about a quantitative variable (average beer price), two types of qualitative variables (ordinal and nominal) are included as well. Thus questions can be formulated over either or both of these dimensions, for instance: *Which Price Tier sets the highest average beer price? In 1992, in which Price Tier do stores set the highest average price for beer?* and *In which year did stores in the "High" Price Tier set the lowest average price for beer?*

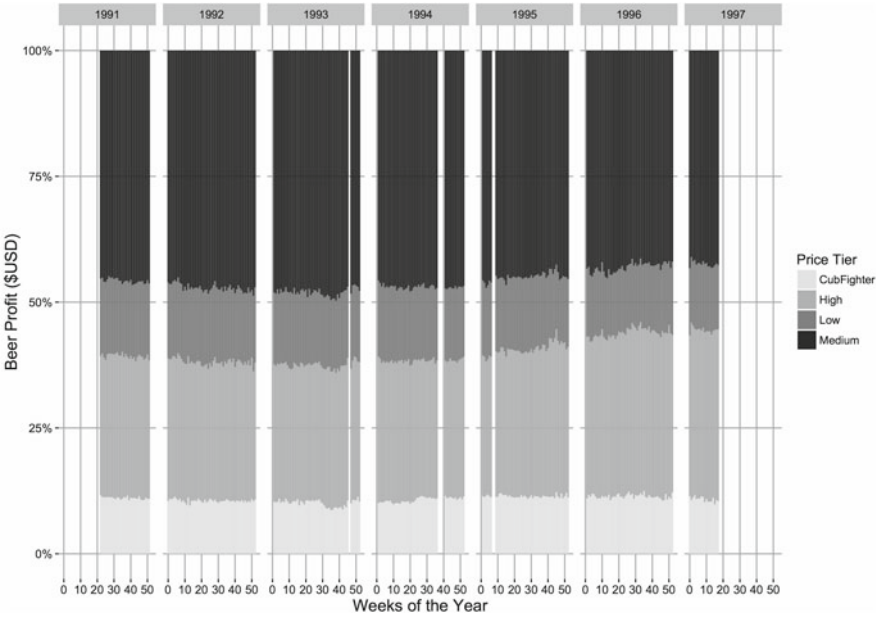


Fig. 5 A 100 % stacked bar chart of beer profit by price tier between 1991 and 1997. Missing values appear as *vertical white bars* (e.g. 1994, near Week 40)

4.2.2 Intermediate-Level Question Visualizations

While elementary-level questions are useful to provide quick, focused insights, intermediate-level questions can be used to reveal a larger amount of detail about the data even when it is unclear at the outset what insights are being sought. Given data visualizations take advantage of graph semiotics to capture a considerable amount of information in a single graphic, it is reasonable to expect that it would be possible to extract similarly large amounts of information to help deepen an understanding of the database. This is particularly valuable as Big Data is prohibitive in size and viewing the database as a whole is not an option. While visualizations supporting intermediate-level questions do not capture the entire database, they do capture a considerable amount of pertinent information recorded in the database.

Figure 6 depicts total beer sales for all stores across geographic zones. This graphic resembles a box plot however is somewhat more nuanced, with the inclusion of 'notches', clearly indicating the location of the median value. Thus not only does Fig. 6 capture the raw data, it provides a second level of detail by including descriptive summary information, namely the median, the interquartile range and outliers. Much information is captured using the retinal variables Position, Shape, Length and Size, meaning that more detailed questions taking advantage of the descriptive statistics can now be posed. For example, *Which zones experience the lowest and highest average beer sales, respectively? Which zones show the most and least variability in beer sales?* and *Which zones show unusually high or low beer sales?* Due

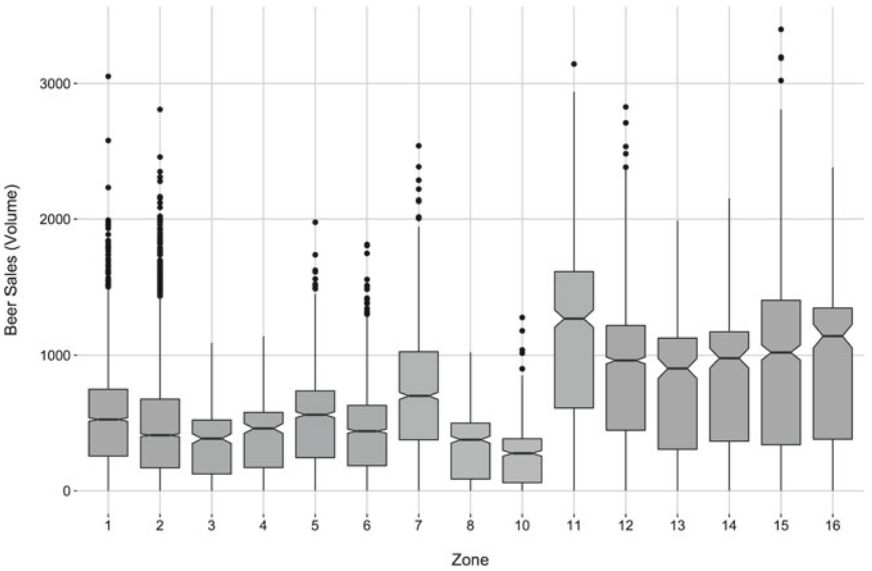


Fig. 6 Notched boxplot for elementary questions regarding beer sales by zone

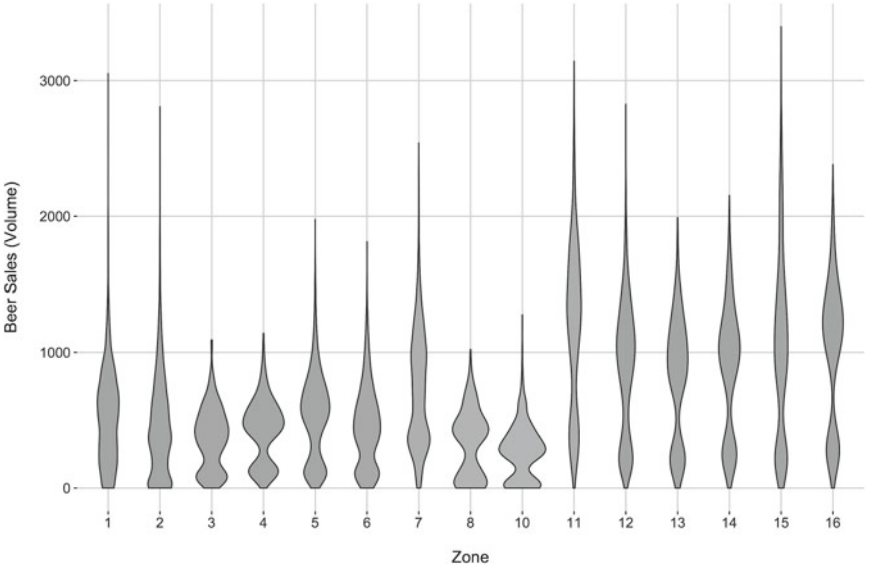


Fig. 7 Violin plot for elementary questions regarding beer sales by zone

to the shape of the notched boxes, comparisons between zones are enabled as well, e.g. *How do stores in Zones 12 and 15 compare in terms of typical beer sales?*

Adjusting this visualization slightly leads to the violin plot of Fig. 7, which is created for the same data as in Fig. 6. However while a similar set of retinal variables are used in this case, Fig. 7 captures complementary information to that captured by Fig. 6. In particular, the retinal variable *Shape* has been adjusted to reflect the distribution of the data, while notches still indicate the location of the average (median) quantity of beer sales. What is gained, however, in terms of data distribution, is lost in terms of detailed information about the stores themselves, namely the exact outliers captured in Fig. 6 versus the more generic outliers, depicted in Fig. 7.

Figure 8 merges the best of both of the notched and violin plots to produce an RDI (Raw data/Description/Inference) plot, with Fig. 8a representing the same data as in Figs. 6 and 7. However, due to the breadth and depth of data representation, it is possible to easily capture other pertinent information about the database, including maximum beer sales, beer price and beer profit (Fig. 8b–d, respectively), allowing easy comparison between multiple quantitative variables in the database, on the basis of identical qualitative information (Price Tier, Zone). Now more detailed intermediate-level questions can be posed, e.g. *How do beer price and profit compare across stores in the Medium price tier?* Generic questions can be now asked of the data as well, such as *How do beer prices in Low price tier stores compare with stores in other tiers?* or *Does any price tier consistently show the most variability across the beer variables Sales, Maximum Sales, Price and Profit?*

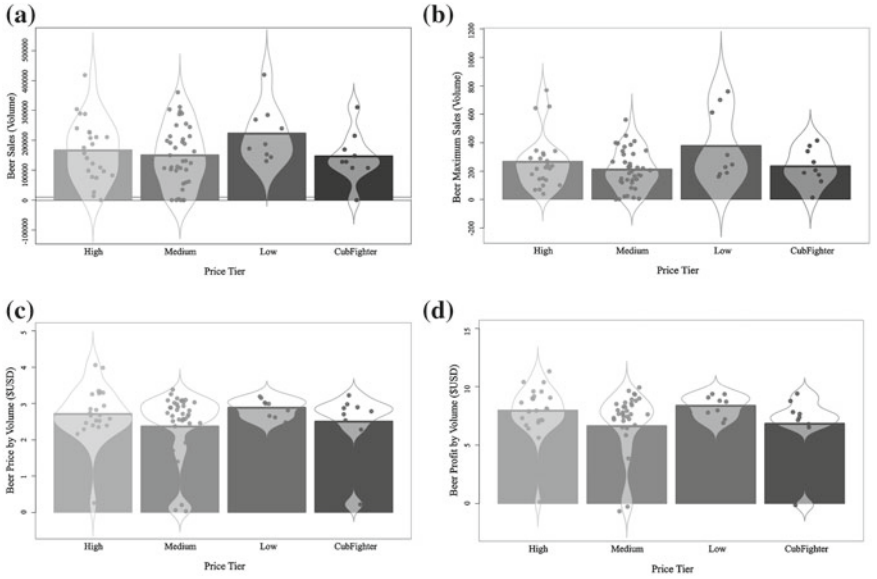


Fig. 8 RDI (raw data/description/inference) plots of beer price, profit and movement

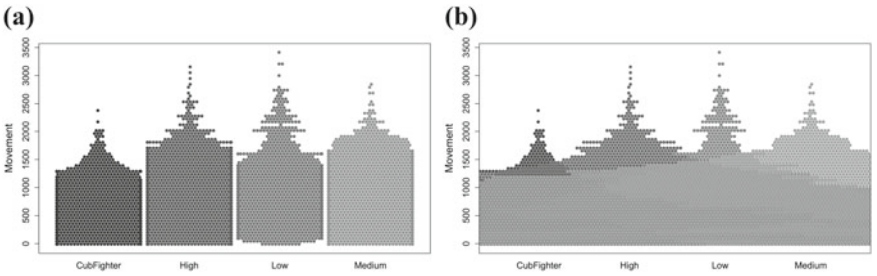


Fig. 9 A swarm plot of beer sales by price tier. The plot in (**a**) has been altered for visually aesthetic purposes, distorting the true nature of the data while the plot in (**b**) shows the true data, however is not easily readable

On a cautionary note, alternative RDI plots are shown in Fig. 9a and b, in which the retinal variable Density has been added to represent the spread of the data. However it is worth observing the construction of the vertical borders of the columns representing each price tier. This is an example in which visual aesthetics have been promoted over data validity and observations that did not fall within the vertical bounds were instead plotted on the vertical boundaries, resulting in the thicker walls of each bar (Fig. 9a). The true spread of the data is shown in Fig. 9b which is highly unreadable. Thus while the raw data can be viewed and a descriptive statistic can be obtained from each price tier, interpretation of this type of plot is necessarily more restrained than for R DI plots such as those in Fig. 8.

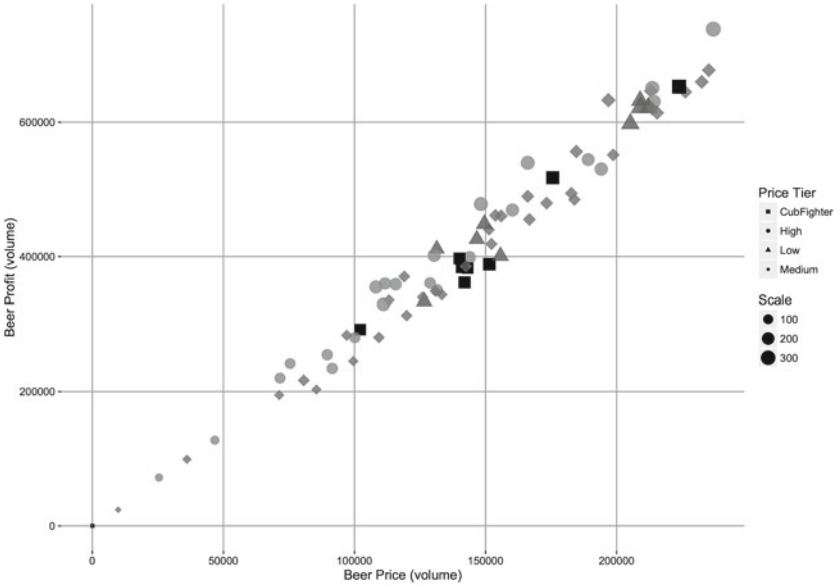


Fig. 10 Bubble plot of beer price versus profit, relative to product sales, over each price tier

Thus far, questions have been formulated about characteristics of a single variable, however it is also often of interest to determine the association between two variables. Figure 10 depicts a bubble plot, in which the retinal variables Position, Color Hue, Color Saturation, Shape, Density and Size are used to convey information about two quantitative variables (beer profit and price), summarized over a qualitative variable (Price Tier). Questions about the relationship between the quantitative variables can be posed, e.g. *Are beer prices and profits related? Is the relationship between beer price and profit consistent for lower and higher prices?* Or including the qualitative variable: e.g. *Are beer prices and profits related across price tiers? Which price tier(s) dominate the relationship between high prices and profits?*

A complementary focus for intermediate-level questions is the interplay between specific and general information. Figure 11 depicts this juxtaposition via a bubble plot, however now focusing on two beer brands and their weekly sales across the four price tiers, using the retinal variables Color Hue, Shape, Size, Position and Length. Questions that can be posed include e.g. *Amongst which price tiers is Miller the most popular beer? Is Budweiser the most popular beer in any price tier?* Alternatively, the focus can instead be on the second qualitative variable (price tier) e.g. *Which is the most popular beer in the Low price tier?*

As a second cautionary tale, we note that there is no one graph type that is infallible, and often the selection of variables of interest will determine the usefulness of a visualization. For example, Figs. 12 and 13 both display the same data, namely beer sales over the four price tiers, for two popular brands. However while the box plot of Fig. 12 is an RDI plot in that it captures a great deal of detail about each variable,

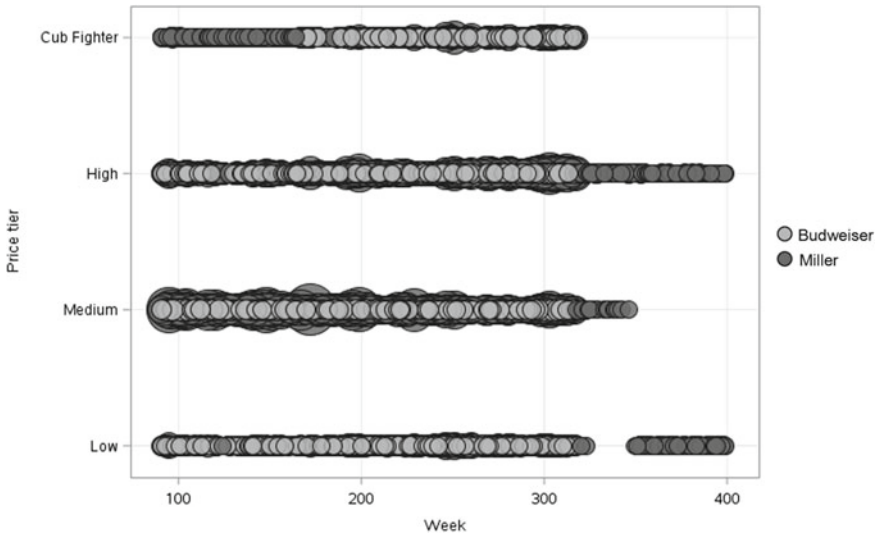


Fig. 11 Bubble plot of beer sales of two popular brands over four store price tiers

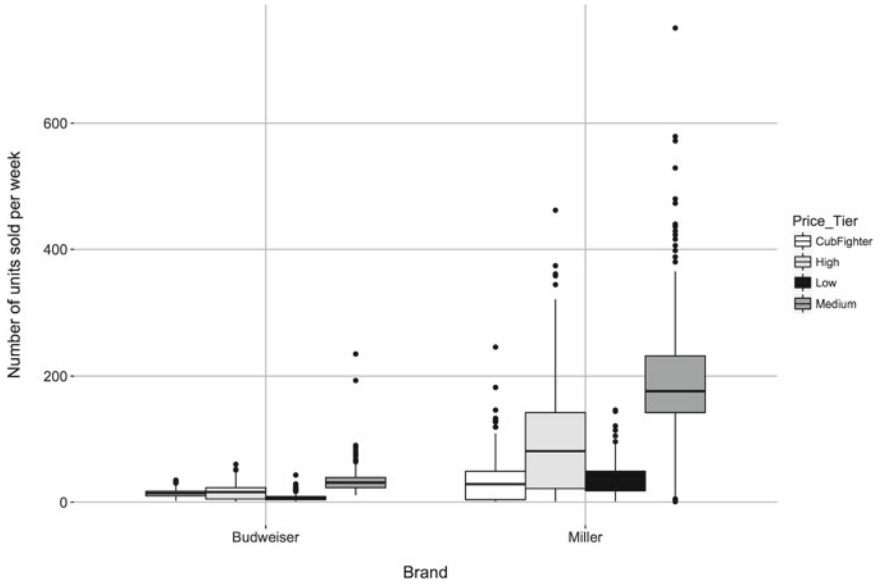


Fig. 12 A box plot of beer sales of two popular brands, over the four price tiers

much of the data features are obscured by the differences in scale of sales for the two selected beer brands. On the other hand, the butterfly plot (Fig. 13) offers a variation on a simple bar chart by utilizing an extra retinal variable—Orientation—and in doing so provides a more informative comparison of the average sales levels.

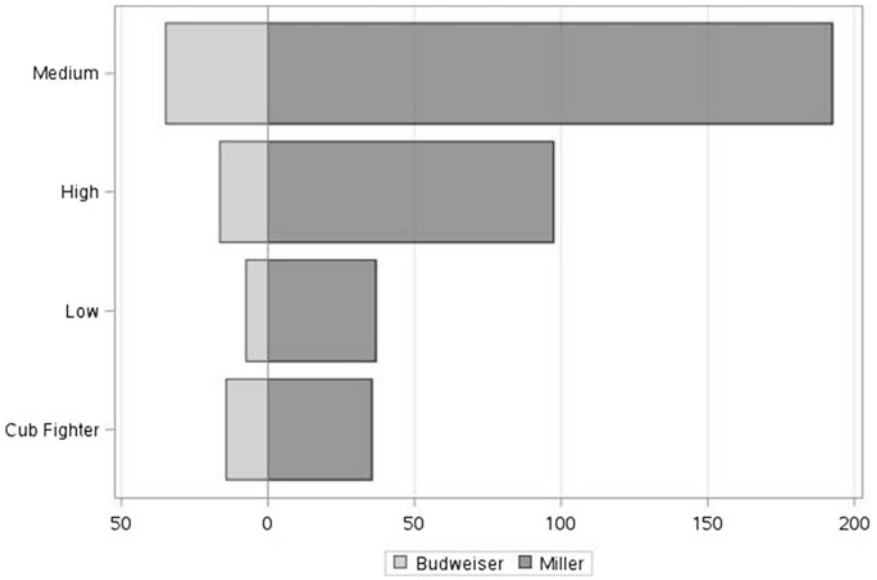


Fig. 13 A butterfly plot of beer sales of two popular brands, over the four price tiers

A useful extension to intermediate-level questions comes from using the retinal variables Color Hue and Orientation, to produce a ternary plot [22]. In Fig. 14 the palette used to represent Hue is shown at the base of a ternary heat map, with the interplay between the three quantitative variables Beer movement (sales), price and profit being captured using shades from this palette.

Intermediate-level questions can now focus on the combination of quantitative variables and can either be specific e.g. *For beer sold in the top 20 % of prices, what percentage profit and movement (sale) are they experiencing?* or generic, so as to uncover information, e.g. *do stores that sell beer at high prices make a high profit? Is it worth selling beer at low prices?* The power of question formulation based solely on the judicious selection of retinal variables makes extracting insights from Big Data less formidable than original appearances may suggest, even when combined with standard graphical representations.

4.2.3 Overall-Level Question Visualizations

Questions at the overall level focus on producing responses that cover the data as a whole, with an emphasis on general trends [7]. Time series plots are useful in this regard, however traditional time series plots which show all data points over the entire data collection usually renders very little information and are often difficult

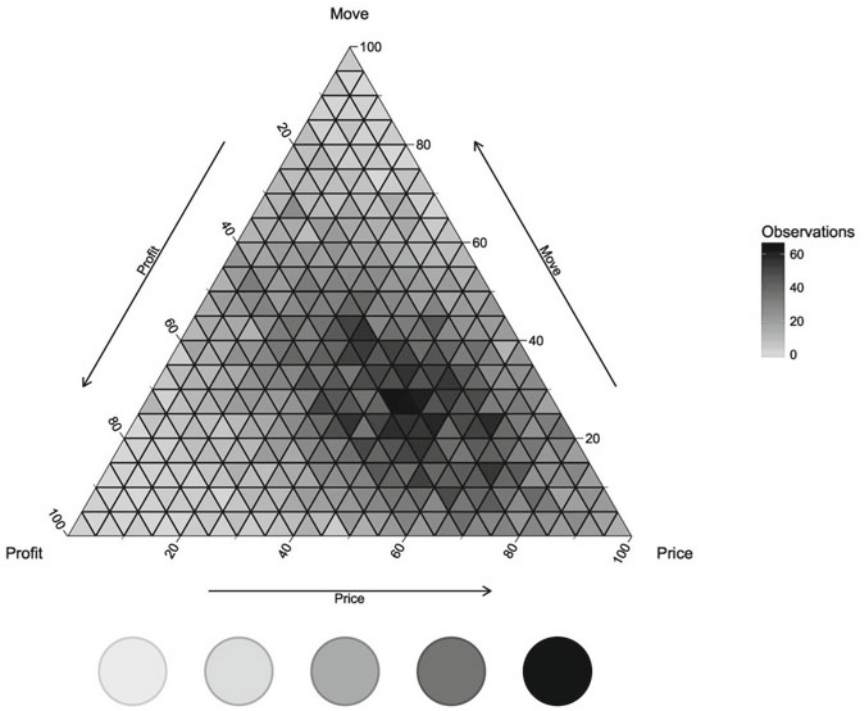


Fig. 14 Ternary plot for intermediate questions about beer over three quantitative variables: price, profit and move (sales). The *color hue palette* is shown at the base of the plot and is reflected in the plot and legend

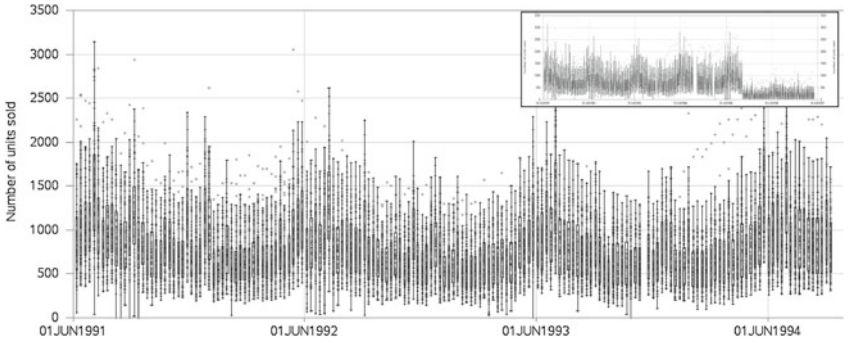


Fig. 15 Time series plot of the distribution of weekly beer sales in each store

to read, unless specialized visualizations of the time series are shown, such as an inset (Fig. 15). Although it is possible to glean very general trends in this case, there is still an opportunity loss in that retinal variables are not being properly utilized to convey more subtle information.

Horizon plots are a variation on the standard time series plot by deconstructing a data set into smaller subsets of time series that are visualized by stacking each series upon one another. Figure 16 demonstrates this principle using beer profit data for all stores between the years 1992–1996. In this case the shorter time interval was chosen to present easily comparable series that each span a single year. Even so, similar questions to those posed for Fig. 15 apply in this case. The attraction of the horizon plot lies however in the easy comparison between years and months which is not facilitated by the layout of Fig. 15.

Interpretation of the beer data is improved again when the overall time series is treated as an RDI plot (Fig. 17) and information about each store can be clearly ascertained over the entire time period, with the focus now on spatial patterns in the data, rather than temporal.

In contrast, the same information is depicted in Fig. 18 however through the addition of the retinal variable Colour Hue, the data presentation allows for easier interpretation and insight. In this case questions such as *When were beer profits at a low and when were they at a high? What is the general trend of beer profits between 1991 and 1997?* can be asked. Similar questions can also be asked of the data in the time series RDI plot (Fig. 17) however the answer will necessarily involve the stores, providing alternative insight to the same question.

An alternative representation of overall trends in the data come from a treemap visualization, which aims to reflect any inherent hierarchy in the data. Treemaps are flexible as they can be used to not only capture time series data, but also separate data by a qualitative variable of interest, relying on the retinal variables Size and Colour

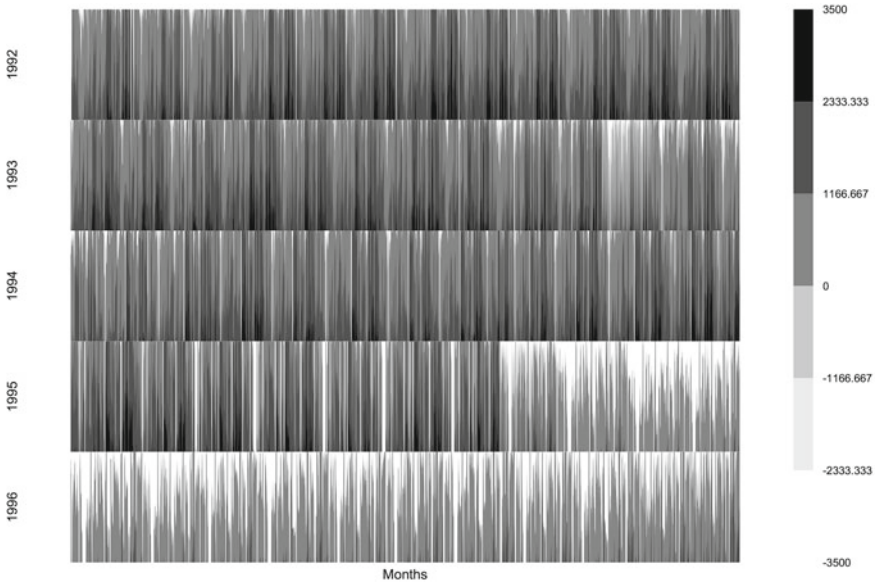


Fig. 16 Horizon plot of beer profit over all stores each week between 1992 and 1996

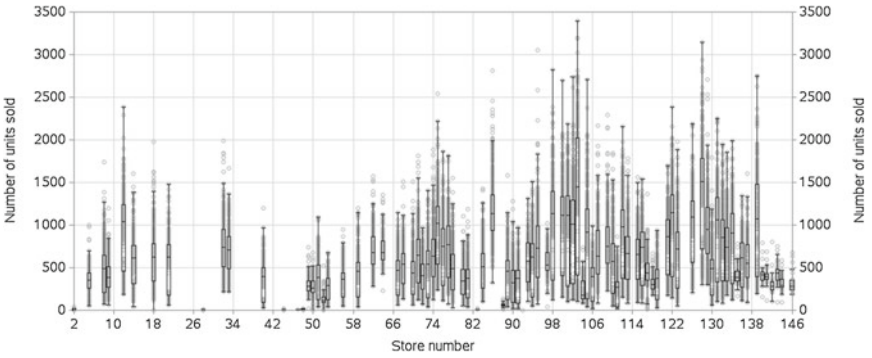


Fig. 17 Time series RDI plot of weekly beer sales in each store

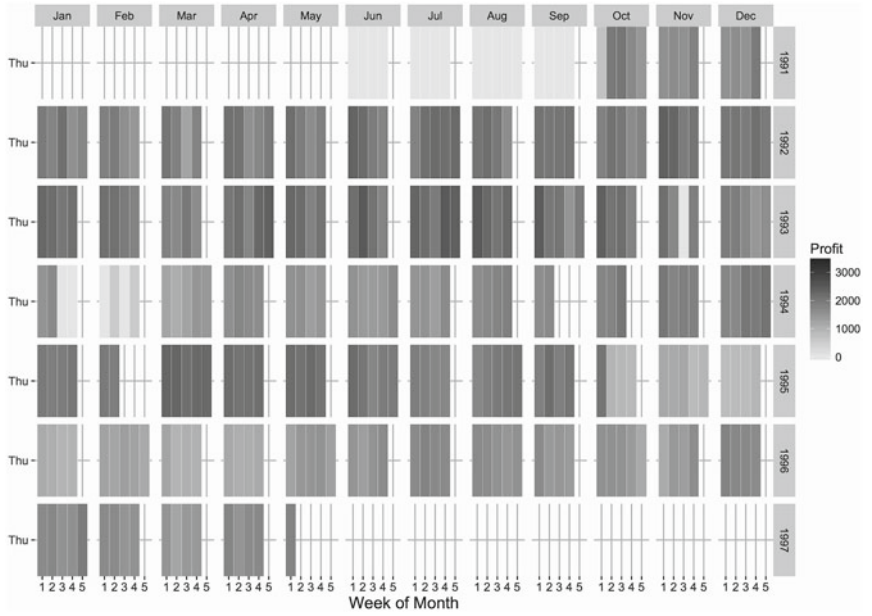


Fig. 18 Calendar heat map of beer profit over all stores in each week

Hue to indicate differences between and within each grouping. The advantage of such a display is the easy intake of general patterns that would otherwise be obfuscated by data volume. For this reason, treemaps are often used to visualize stock market behavior [23].

To create a treemap two qualitative and two quantitative variables are required. The item of interest (qualitative) is used to form the individual rectangles or 'tiles', while the group to which the item belongs (qualitative) is used to create separate areas in the map. A quantitative variable to scale the size of each rectangle is required

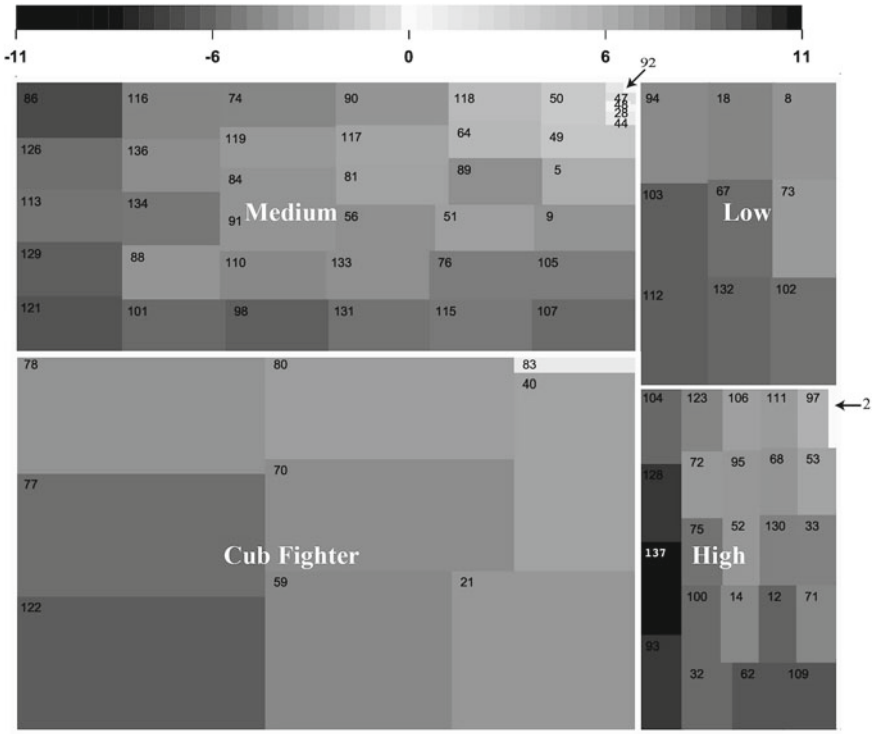


Fig. 19 A treemap showing the relationship between beer price and profit across price tiers and at each individual store

and a second quantitative variable assigns the colour hue to each tile. Figure 19 displays a treemap of the beer data using the qualitative variables Store and Price Tier and the quantitative variables Beer Price and Beer Profit. Each store corresponds to a single tile in the map while Price Tier is used to divide the map into four separate areas. The size of each tile corresponds to the price of beer at a given store, while the color hue represents the profit made by each store, with the minimum and maximum values indicated by the heat map legend. Questions postulated from a treemap include *Which stores are generating high profits? What is the relationship between beer price and profit? Which price tiers make the largest profit by selling beer? Do stores within a price tier set the price of beer consistently against one another?*

The last graph presented here is another variation of a time series plot, however with added functionality to depict the behavior of multiple groups simultaneously. Figure 20 shows a streamgraph of beer profit made in every single store over the entire time period represented in the data set. In this case the retinal variables Length, Orientation and Color Hue are being used to combine quantitative information (beer profit) with qualitative groups (stores) to give an overall view of the general trend. The streamgraph in R however has an added feature that is a modernization of the

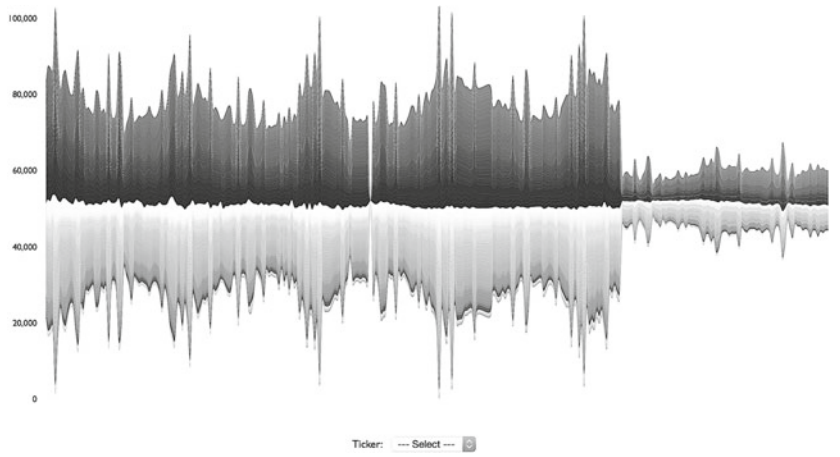


Fig. 20 Streamgraph of beer profit over all stores in each week

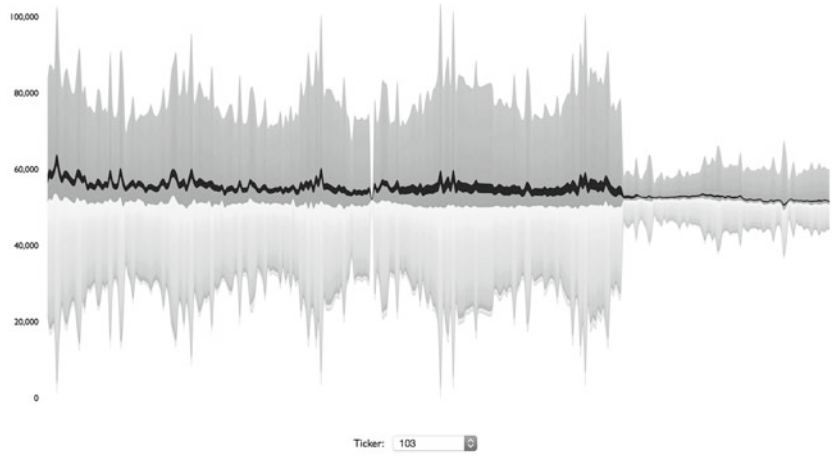


Fig. 21 Streamgraph of beer profit for store 103 in each week

retinal variable Color Saturation. The R streamgraph is interactive, with a drop-down menu to select a particular store of interest (labeled *Ticker* in Fig. 20). The streamgraph is also sensitive to cursor movements running over the graph, and will indicate in real time over which store the cursor is hovering. Figure 21 demonstrates the selection of Store 103 and how the modernization of an 'old' technique further enhances the types of insights that can be drawn from this graph.

Thus questions that can be asked about the data based on streamgraphs include *What is the overall trend of beer prices? Does beer price behaviour change over time? Are there repetitive patterns to beer price behaviour?* Then, coupled with

Color Saturation to select a store of interest, *What is the overall trend of beer prices at a particular store? Does the trend of this store behave similarly to the overall pattern?* and so forth, allowing for overall-level questions that compare specific item behavior (e.g. a store) with the overall trend in the data.

5 Conclusions

Visualization of data is not a new topic—for centuries there has been a need to summarize information graphically for succinct and informative presentation. However recent advances have challenged the concept of data visualization, through the collection of 'Big Data', that is data characterized by its variety, velocity and volume and is typically stored within databases that run to petabytes in size.

In this chapter, we postulated that while there have been advances in data collection, it is not necessarily the case that entirely new methods of visualization are required to cope. Rather, we suggested that tried-and-tested visualization techniques can be adopted for the representation of Big Data, with a focus on visualization as a key component to drive the formulation of meaningful research questions.

We discussed the use of three popular software platforms for data processing and visualization, namely SAS, R and Python and how they can be used to manage and manipulate data. We then presented the seminal work of [7] in the use of graph semiotics to depict multiple characteristics of data. In particular, we focused on a set of retinal variables that can be used to represent and perceive information captured by visualization, which we complemented with a discussion of the three types of questions that can be formulated from such graphics, namely elementary-, intermediate- and overall-level questions.

We demonstrated application of these techniques using a case study based on Dominick's Finer Foods, a scanner database containing approximately 98 million observations across 60 relational files. From this database, we demonstrated the derivation of insights from Big Data, using commonly known visualizations and also presented cautionary tales as a means to navigate graphic representation of large data structures. Finally, we also showcased modern graphics designed for Big Data, however with foundations still traceable to the retinal variables of [7], in support of the view that in terms of data visualization, everything old is new again.

References

1. McAfee A, Brynjolfsson E (2012) Big data: the management revolution. Harvard Bus Rev 90(10):60–68
2. SAS (2014) Data visualization techniques: from basics to big data with SAS visual analytics. SAS: White Paper
3. Oancea B, Dragoecsu RM (2014) Integrating R and hadoop for big data analysis. Revista Romana de Statistica 2(62):83–94

4. Gelper S, Wilms I, Croux C (2015) Identifying demand effects in a large network of product categories. J Retail 92(1):25–39
5. Toro-González D, McCluskey JJ, Mittelhammer RC (2014) Beer snobs do exist: estimation of beer demand by type. J Agric Resour Econ 39(2):1–14
6. Huang T, Fildes R, Soopramanien D (2014) The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. Eur J Oper Res 237(2):738–748
7. Bertin J (1967) Semiology of graphics: diagrams, networks, maps. The University of Wisconsin Press, Madison, Wisconsin, p 712
8. Mackinlay J (1986) Automating the design of graphical presentations of relational information. ACM Trans Graph 5(2):110–141
9. Eichenbaum M, Jaimovich N, Rebelo S (2011) Reference prices, costs, and nominal rigidities. Am Econ Rev 101(1):234–262
10. Chen Y, Yang S (2007) Estimating disaggregate models using aggregate data through augmentation of individual choice. J Mark Res 44(4):613–621
11. Chintagunta PK, Vishal J-PDS (2003) Balancing profitability and customer welfare in a supermarket chain. Quant Mark Econ 1:111–147
12. Nevo A, Wolfram C (2002) Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. RAND J Econ 33(2):319–339
13. Levy D, Lee D, Chen HA, Kauffman RJ, Bergen M (2011) Price points and price rigidity. Rev Econ Stat 93(4):1417–1431
14. McKinney W (2012) Python for data analysis. O'Reilly Media, p 466
15. Tufte ER (1983) The visual display of quantitative information. Graphics Press, Cheshire, Connecticut, p 197
16. Card S (2009) Information visualisation. In: Sears A, Jacko JA (eds) Human-computer interaction handbook. CRC Press, Boca Raton, pp 181–215
17. Card SK, Mackinlay JD, Shneiderman B (1999) Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc., San Francisco
18. Lengler R, Eppler MJ (2007) Towards a periodic table of visualization methods of management. In: Proceedings of graphics and visualization in engineering (GVE 2007), Florida, USA, ACTA Press, Clearwater, pp 1–6
19. Krygier J, Wood D (2011) Making maps a visual guide to map design for GIS, 2nd edn. Guilford Publications, New York, p 256
20. Cleveland WS, McGill R (1984) Graphical perception: theory, experimentation and application to the development of graphical methods. J Am Stat Assoc 79(387):531–554
21. Koussoulakou A, Kraak MJ (1995) Spatio-temporal maps and cartographic communication. Cartographic J 29:101–108
22. Breckon CJ (1975) Presenting statistical diagrams. Pitman Australia, Carlton, Victoria, p 232
23. Jungmeister W-A (1992) Adapting treemaps to stock portfolio visualization. Technical report UMCP-CSD CS-TR-2996, College Park, Maryland 20742, U.S.A

Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective

Leonard Heilig and Stefan Voß

Abstract The development of big data applications is closely linked to the availability of scalable and cost-effective computing capacities for storing and processing data in a distributed and parallel fashion, respectively. Cloud providers already offer a portfolio of various cloud services for supporting big data applications. Large companies like Netflix and Spotify use those cloud services to operate their big data applications. In this chapter, we propose a generic reference architecture for implementing big data applications based on state-of-the-art cloud services. The applicability and implementation of our reference architecture is demonstrated for three leading cloud providers. Given these implementations, we analyze main pricing schemes and cost factors to compare respective cloud services based on a big data streaming use case. Derived findings are essential for cloud-based big data management from a cost perspective.

Keywords Big data management · Cloud-based big data architecture · Cloud computing · Cost management · Cost factors · Cost comparison · Provider selection · Case study

1 Introduction

The cloud market for big data solutions is growing rapidly. Besides full-service cloud providers that offer a large portfolio of different infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) solutions, there are also some niche providers focusing on specific aspects of big data applications. In general, such big data applications are highly dependent on a scalable computing infrastructure, programming tools, and applications to efficiently process large data

L. Heilig (✉) · S. Voß
Institute of Information Systems (IWI), University of Hamburg, Hamburg, Germany
e-mail: leonard.heilig@uni-hamburg.de

S. Voß
e-mail: stefan.voss@uni-hamburg.de

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_2

sets and extract useful knowledge [17]. In this regard, cloud computing represents an attractive technology-delivery model as it promises the reduction of capital expenses (CapEx) and operational expenses (OpEx) [11] and further moves CapEx to OpEx, closely correlating expenses with the actual use of tools and computing resources [5]. A recent scientometric analysis of cloud computing literature further indicates that there is a huge research interest in scalable analytics and big data topics [10]. As cloud-based big data applications are usually composed of several managed cloud services, it becomes increasingly important to identify important cost factors in order to evaluate potential use cases and to make strategic decisions, for instance, concerning the choice of a cloud provider (for an extensive overview on decision-oriented cloud computing, the reader is referred to Heilig and Voß [9]). The variety of possible configurations and pricing schemes makes it difficult for consumers to estimate overall costs of cloud-based big data applications. Often, consumers appear in the form of cloud application providers, as companies like Netflix and Spotify, outsourcing operations of their services to third-party cloud infrastructures. To benefit from big data technologies and applications in companies, it is meanwhile essential to address the economic perspective and provide means to evaluate the promises cloud computing gives with regard to the use of highly scalable computing infrastructures in order to unlock competitive advantages and to maximize value from the application of big data [18]. To the best of our knowledge, a cost perspective for implementing big data applications in cloud environments has not yet been addressed in the current literature.

In this chapter, we propose a generic reference architecture for implementing big data applications in cloud environments and analyze pricing schemes and important cost factors of related cloud services. The cloud reference architecture considers state-of-the-art technologies and facilitates the main phases of big data processing including data generation, data ingestion, data storage, and data analytics. Both batch and stream processing of big data is supported. We demonstrate the applicability and implementation of the proposed architecture by specifying it for the cloud services of the, according to Gartner's magic quadrant [6], three leading cloud providers, namely *Amazon Web Service*, *Google Cloud*, and *Microsoft Azure*. Practical implementations of large companies like Netflix and Spotify verify the relevancy of the defined architectures. The individual architectures provide a basis for evaluating important cost factors. For each of the main phases of big data processing, we identify and analyze the scope and cost factors of relevant cloud services based on a case study. In cases a comparison is useful, we compare cloud services of the different cloud providers and derive important implications for decision making. Thus, the contribution of this chapter is twofold. First, the chapter provides a blueprint for implementing state-of-the-art cloud-based big data applications and gives an overview about available cloud services and solutions. Second, the main part is concerned with providing a cost perspective on cloud-based big data applications, which is essential for big data management for cloud consumers.

The chapter is structured as follows. Section 2 defines the main phases of big data processing and presents the generic reference architecture. Moreover, we describe the implementation of the reference architecture using cloud services of the three

leading cloud providers. For each big data processing phase, cloud pricing schemes and relevant cost factors of those cloud services are analyzed based on a case study focusing on streaming analytics in Sect. 3. In Sect. 4, we discuss main findings and implications. Finally, we draw conclusions and identify activities for further research.

2 Big Data Processing in Cloud Environments

The calculation of costs is highly dependent on the utilized cloud services. Major cloud service providers offer a plethora of different tools and services to address big data challenges. In this section, we define a common reference architecture for big data applications. The reference architecture corresponds to the state-of-the-art and supports main phases of big data processing from data generation to the presentation of extracted information, as depicted in Fig. 1. After briefly explaining these phases and the corresponding reference architecture, we give an overview on its implementations with cloud services of the three leading cloud service providers.

2.1 Generic Reference Architecture

The processing of big data can be divided into five dependent phases. In the first phase, data is generated in various applications and systems. This might include internal and external data in various forms and formats. Depending on the rate of occurrence and purpose of collected data, velocity requirements may differ among data sources. The second phase involves all steps to retrieve, clean, and transform the data from different sources for further processing. This may include, for instance, data verification, the extraction of relevant data records, and the removal of duplicates in order to ensure efficient data storage and exploitation [3]. Typically, the data is permanently stored in a file system or database. In some cases of streaming applications, however, value can only be achieved in the first seconds after the data is produced, making a persistent storage obsolete. Nevertheless, information and results being extracted during processing and analysis usually need to be stored and managed permanently. In the fourth phase, different methods, techniques, and systems are used to analyze and utilize the data in order to extract information relevant for

Fig. 1 Phases of big data processing

supporting business activities and decision making. The information and results of the data analytics phase need to be visualized, allocated, distributed, and presented to its users in the final phase.

To define a generic reference architecture, we reviewed several technical documentations, recommendations, and use cases provided by cloud providers (see, e.g., [2]) and interviewed experts of one of the largest cloud providers. Moreover, we analyzed practical implementations of cloud-based big data platforms of large cloud consumers, such as Netflix [12] and Spotify [15], using cloud services of different large cloud providers. Considering both batch and real-time data stream processing, we identify a common structure of systems interacting with each other to support the different phases of big data processing. To express this structure, we present a state-of-the-art generic reference architecture in Fig. 2. The arrows represent the data flows. Note that variations of this generic architecture and associated data flows are possible. In the following, we briefly explain the basic components of the reference architecture.

Data Generation: The number of data producers and the amount of data being produced is continuously increasing. This involves business data from internal systems (e.g., production data, inventory data, sales data, e-commerce platform data, etc.) and data from external third-party systems (e.g., social network data, government data, weather data, finance data, search trends, etc.) being offered through the Internet. The emergence of the internet of things (IoT), enabling physical objects to sense and act on their environment by interacting with each other, represents another big data source.

Data Ingestion: For the data ingestion, it is essential to consider velocity requirements, which mainly determine how fast the data is fed into the overall system and the processing latency between data generation and presentation. Thus, a system architecture must consist of components that support both real-time and batch processing of data. The former must be supported by systems that enable a fault-tolerant, scalable, and consistent real-time processing of *data streams* from a large number of

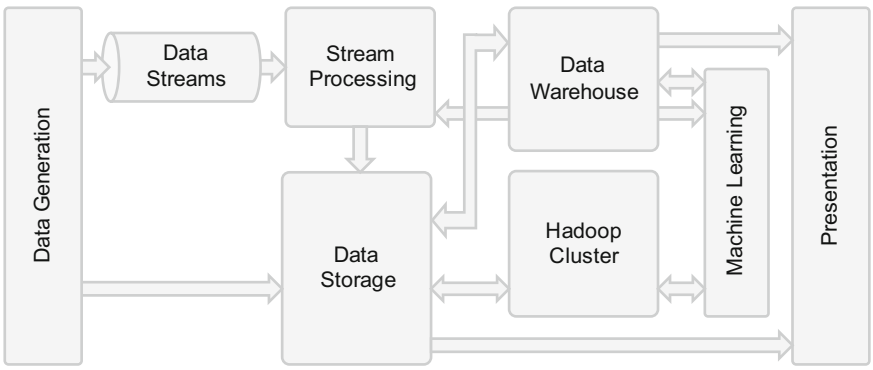


Fig. 2 Generic reference architecture

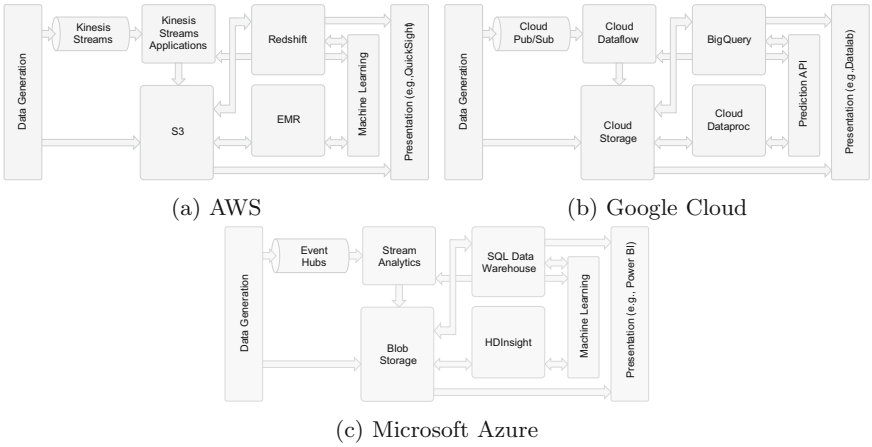
data sources. To harvest its potential, it is necessary to implement applications and methods to further process or analyze streaming data immediately, depending on the processing latency requirements. Therefore, the *stream processing* component contains all logic to immediately utilize streaming data, for instance, by generating alerts or making recommendations based on machine learning tools. If useful for future processing, streaming data is transferred to the central persistent *data storage* component to be stored permanently. This may involve extract, transform, and load (ETL) jobs that reliably export the data into a central storage or directly to a central processing cluster (see, e.g., a recent streaming solution of Spotify using *Google Cloud* [15]). Typically static data with a low velocity is initially stored in a database or file system. To further process and analyze this data, for instance in a data warehouse, those data sources may need to be consolidated and integrated with additional ETL jobs.

Data Storage: For permanent use in form of batch processing, data should be persisted as files or data records in file systems or databases designed to provide scalability, high availability, and low latency. From there on, the data can be used by several distributed applications in parallel. For data analytics in a data warehouse, it is necessary to load the data into database tables of the data warehouse using ETL. A permanent storage of data in an Apache Hadoop cluster is often economically unreasonable as it involves huge costs for using necessary cluster nodes in form of virtual machines (VM). Large cloud consumers, like Netflix [12], show that the integration with low cost cloud storage services instead of using local storage based on the Hadoop Distributed File System (HDFS) [16] can be beneficial. For instance, multiple clusters can access and process the same data for different workloads depending on the data analysis task [12]. However, reading and writing to a central file system is slower than using local storage and thus requires a low-latency and high-bandwidth access. As a compromise, local storage of HDFS is often used for all intermediate stages of MapReduce processes. Consequently, a mixture of a *shared nothing architecture*[1] and *shared storage architecture*[2] approach is often used in practice for operating Hadoop clusters efficiently.

Data Analytics: The generic architecture further supports ad-hoc data queries and advanced data analytics. For providing related cloud services, cloud providers leverage their capabilities in providing massive and scalable computing infrastructure. While a data warehouse aids the processing and analysis of structured data, a Hadoop cluster can be used to process and transform unstructured and semi-structured data into structured data for further processing in databases and *data warehouses*. For the latter, the Hadoop ecosystem provides several extensions for data processing, querying, storage in NoSQL databases (e.g., HBase) and data warehouses (e.g., Hive) as well as advanced statistical and machine learning algorithms

[1] A shared nothing architecture denotes a distributed computing architecture consisting of nodes that only possess and utilize their own computing resources including memory and disk storage. This facilitates, inter alia, a large scale horizontal scaling using commodity machines based on a distributed file system.

[2] In a shared storage environment, a central file storage system is shared among the nodes.



(a) AWS (b) Google Cloud

(c) Microsoft Azure

Fig. 3 Reference architectures of leading cloud service providers

(e.g., Mahout). Thus, a data warehouse is often build upon a Hadoop cluster. Instead of using the MapReduce programming model for processing data in a cluster, it has been shown that Apache Spark, building on in-memory storage to avoid slow disk reads/writes as well as directed acyclic graph (DAG) scheduling to better parallelize processing stages, offers up to two orders of magnitude performance increase [14]. For structured data queries, it is essential that the data warehouse service supports SQL[3] commands. Moreover, a component that provides *machine learning* algorithms and technology for supporting artificial intelligence and predictive analytics shall be provided as a service.

Presentation: Once the data has been analyzed and stored, normally in a data warehouse, resulting insights and results need to be transformed into rich visualizations, dashboards, and reports for individual stakeholders in order to avoid information overload and decrease transaction costs. Therefore, tools to prepare and manage rich visualizations and reports need to be provided.

2.2 Implementations of the Generic Reference Architecture

In this section, we show the technical implementation of the proposed generic big data processing reference architecture. For this purpose, we have reviewed the technical details and architectures of managed cloud services offered by the three leading cloud providers, namely *Amazon Web Services (AWS)*, *Google Cloud*, and *Microsoft Azure*. For each cloud provider, we present a reference architecture in Fig. 3 describing the use of managed cloud services for supporting each phase of

[3] Abbr. for Structured Query Language.

big data processing as described in Fig. 1. A prerequisite for considering a cloud service was that it is fully compatible and integrated with other cloud services of the respective provider. In this regard, all three cloud providers are able to address each phase of a big data processing lifecycle with at least one managed cloud service and further support their integration as well as the use of third-party services. The cloud services conform to the above mentioned requirements (see Sect. 2.1). The proposed composition and an associated integration of those cloud services is fully supported and in line with best practices. As the primary focus is on cost management, details on the implementation of the proposed architectures are out of scope in this chapter. For an overview and technical details, the reader is referred to the documentation of each cloud service given by the respective cloud provider.[4] Generally, those examples demonstrate the applicability of the proposed generic reference architecture.

3 Cloud Pricing and Cost Perspective

After defining the reference architectures for each cloud provider, we investigate possible configurations and the pricing schemes for each category of cloud service in this section. In doing so, we identify main cost factors. Based on a real-time data streaming example, we present a cost comparison for the cases where similar pricing schemes and configurations are possible.

3.1 Data Streams and Stream Processing

By analyzing the cost models of data streaming cloud services, we identify two different approaches. *Amazon Kinesis Streams* charges for each message event occurring during data ingestion and delivery, whereas no additional costs are charged for the required throughput. Although volume-tiered pricing is supported, the prices per million message events are comparatively high. *Kinesis Streams* and *Azure Event Hubs* use a common cost model that calculates the costs based on the volume of incoming message events and required throughput. That is, the stream is grouped into smaller streams (i.e., substreams) with a maximum throughput. Thus, the number of substreams needs to be scaled according to the current message load for achieving real-time processing. Consequently, both volume and speed scaling is covered in the cost models of those cloud providers.

In Table 1, we see that *Kinesis Streams* has competitive prices and is therefore able to provide the most inexpensive solution, also in terms of scaling. A difference between *Kinesis Streams* and *Event Hubs* is the maximum message event size.

[4]The technical documentation and pricing details can be found on the website of the respective cloud providers: AWS (https://aws.amazon.com), Google Cloud (https://cloud.google.com), and Microsoft Azure (https://azure.microsoft.com/).

Table 1 Cost comparison of cloud streaming services (Scenario 1: 100 data records/s, 35 KB record size, 3.42 MB/s input stream, 1 subscriber)

Configuration/pricing	Kinesis streams	Pub/sub*	Event hubs
Max. message event size in KB	25	64	64
Max. thoughput ingress in MB/s per substream	1	N/A	1
Max. thoughput egress in MB/s per substream	2	N/A	2
Message retention in days	1	7**	1
Pricing scheme	Fixed	Tiered	Fixed
Price per million message events per hr ($)	0.014	0.40/0.20/0.10/0.05***	0.028
Price streaming per throughput unit per hr ($)	0.015	N/A	0.03
Required number of substreams (ingress)	4	N/A	4
Additional number of substreams (egress)	0	N/A	0
Number of message events per day in million	17.28	17.28	8.64
Costs ($)			
Overall costs per day	**1.68**	6.91	3.12
Overall costs per month	**52.14**	214.27	96.78

*Assuming that the push subscription mode is used
**If the subscriber is not present; otherwise, messages are dropped after delivery/failure
***First to 250M/next 500M/next 1000M/next 1750M message events

Assuming that the prices of the two providers would be the same, *Event Hubs* would provide a cost advantage for data records sizes greater than 25 KB due to the smaller amount of message events (see Table 1). Thus, also the size of data records of an application may need to be taken into account for cost considerations. When assuming that *Google Cloud Pub/Sub* would provide more attractive prices, the service might be economically advantageous in terms of throughput scaling. In general, we identify two main cost factors: throughput capacity (ingress and egress) and number of message events.

Next, we analyze the portfolio of cloud services for real-time stream processing and analytics. All cloud services allow an individual creation of jobs to read, transform, and analyze streaming data. While Google's *Dataflow* has established a programming model to simplify the implementation of data processing jobs, *Azure Stream Analytics* focuses on structured data processing and allows running SQL-like queries. Although *AWS* has announced that the managed service *Kinesis Analytics* will be available soon, consumers currently have to implement applications using the *Kinesis Client Library* (*KCL*) and different connectors for establishing a link to other cloud services (e.g., *S3*), similar to the *Dataflow* approach. All three cloud services allow an integration with both their own and third-party machine learning cloud services. In general, the costs of running all three cloud solutions are highly dependent on the computational demands of processing and analytics tasks in terms of compute and storage requirements.

Table 2 Cost factors of streaming analytics cloud services

Kinesis analytics	• Number and types of VM cluster nodes
	• Storage capacity
Dataflow	• Number and types of VM cluster nodes
	• Number of GCEU (batch/streaming)
	• Storage capacity
Stream analytics	• Volume of streaming data to be processed
	• Compute capacity in streaming units

Besides costs for using available VM instance types and local storage capacity, *Dataflow* additionally charges per GCEU[5] and differentiates between streaming and batch mode. As shown in a recent implementation of Spotify's event delivery system, the streaming mode considerably decreases end-to-end latency for exporting message events from *Pub/Sub* to *Cloud Storage* [15]. Applying *KCL* further implies costs for using a *DynamoDB* that tracks state information in a cluster of workers that process the data from the stream and transfer the result to respective applications. In general, we see that both cloud providers offer a great flexibility regarding the configuration of the infrastructure and streaming application, but also require a high expertise. The approach of *Stream Analytics* aims to abstract from the infrastructure and defines a price per streaming unit[6] and data volume. Due to the different measures and approaches, it is difficult to compare those cloud services. Estimating the costs requires the collection of empirical data concerning the use of infrastructure for different processing and analytics jobs. However, the flexibility implied by *Dataflow* and *KCL* allows consumers to adapt infrastructure to their individual requirements, e.g., for achieving cost reductions and/or performance boosts. Due to complexity reasons, it is important to implement brokerage mechanisms for supporting related decisions (see, e.g., [8]). To analyze costs, simulation studies that consider different workload scenarios and configurations may provide more insights. In Table 2, the different cost factors of all solutions are shown. In general, we can identify two main cost drivers: compute and storage capacity.

3.2 Data Storage

The costs for storing data in a cloud are a critical aspect to be considered when planning cloud-based big data applications. All three cloud providers offer a variety of

[5]The Google Compute Engine Unit (GCEU) is used as a measure to calculate the total capacity of a virtual central processing unit (vCPU). Google's *Compute Engine* defines the GCEU for each VM instance type depending on the number of vCPUs.

[6]A streaming unit is a measure for expressing the computing capacity in terms of CPU and memory with a maximum throughput of 1 MB/s.

storage options and database systems. In our cost analysis, we focus on inexpensive standard object storage services that can be used to persist data in different formats and massive quantities. Implementations of Netflix [12] and Spotify [15] further emphasize the critical role of a central data object storage component for big data processing. Both volume-tiered and unit pricing schemes are used, as depicted in Table 3. We see that *Azure Storage* undercuts the prices of its competitors. Moreover, it is not differentiated between different types of object requests. The price per request is considerably low. While cost savings of 34.8 % on average can be achieved, the percentage of cost savings slightly increases by the amount of storage capacity. Fig. 4 shows the increase of costs dependent on the volume of data to be persisted. Thus, in particular in terms of scaling, *Azure Storage* offers attractive pricing.

Comparing *Amazon S3* and Google's *Cloud Storage*, we see that the high costs for "expensive" requests (e.g., PUT, LIST, etc.) greatly influence the overall storage costs. Thus, the costs are highly dependent on the use intensity and access patterns of consumers. This emphasizes the importance of data preprocessing activities (e.g., based on ETL). Moreover, all three cloud providers charge for data transfers. Incoming data flows and data transfers within a region is generally free; outgoing data flows and inter-regional data transfers are charged based on different unit prices. The latter needs to be considered, for example, if third-party services, such as machine learning services, are used. In general, we identify three main cost factors of cloud storage services: amount of virtual storage capacity, number and type of data requests, and data transfer.

Table 3 Cost comparison of cloud storage services (assuming that 10 % of the streaming data collected in scenario 1 is persisted in the same region and requires a PUT request and a GET request for each data record)

Pricing ($)	S3	Cloud storage	Azure storage
Pricing model	Tiered	Fixed	Tiered
Storage per GB per month	0.03/0.0295/0.029*	0.026	0.024/0.0236/0.0232*
PUT, COPY, POST, LIST requests	0.005**	0.10***	0.0036****
GET and other requests	0.004***	0.01***	0.0036****
Data transfer (same region)	Free	Free	Free
Costs			
Storage	784.03	680.06	617.69
GET requests	10.71	26.78	0.96
PUT/POST/other requests	133.92	267.84	0.96
Overall costs per day	29.96	31.44	**19.99**
Overall costs per month	928.66	974.69	**619.62**

*First 1 TB/next 49 TB/next 450 TB per month
**Per 1000 requests
***Per 10000 requests
****Per 1000000 requests

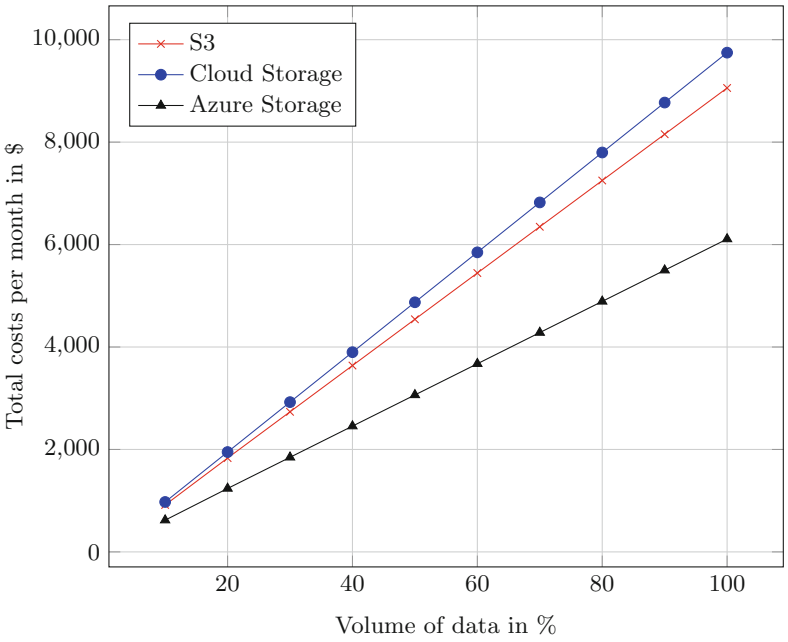


Fig. 4 Cost trends for increasing data volumes (26156–261563 GB)

3.3 Hadoop Cluster

For running a Hadoop cluster, all three cloud providers allow to use a wide range of different VM types and follow comparable pricing schemes. While for *Amazon EMR* and *Google Cloud Dataproc*, it is possible to choose among all VM types provided by *Amazon EC2* and *Google Cloud Compute Engine*, respectively, *Azure HDInsight* limits the range to certain VM types. As shown in Table 4, the price per VM node includes already the costs for using *HDInsight* and is about two times more expensive than using regular nodes of the same VM type. In the other two cases, extra costs for using the Hadoop service occur. *EMR* charges an individual fee per VM type and *Dataproc* calculates an extra fee per vCPU per hour. Thus, three main cost factors need to be considered: number and VM type of cluster nodes, additional usage fee (if applicable), and running time.

To approximately compare costs, we choose a comparable cluster configuration for each cloud provider (see Table 4). In general, we see that the increased price per cluster node leads to considerably higher costs in the case of *HDInsight*. While comparably low storage costs attract consumers to store their data in *Azure Storage*, the inherent comparative cost advantages disappear when it comes to large-scale processing of this data. This further emphasizes the need for an overall cost estimate. We further see that *Dataproc* offers competitive prices for using Google's *Compute*

Table 4 Cost comparison of Hadoop cluster services (clusters with 50 nodes, operated 5 h per day for processing 10 % of the streaming data)

Configuration/pricing	EMR	Dataproc	HDInsight
VM type	m3.xlarge	n1-standard-4	D3 v2
Number of vCPU	4	4	4
Memory in GB	15	15	14
Local SSD storage in GB	80	N/A	200
Additional SSD storage	120	200	N/A
Price VM per hour	0.27	0.20	Incl.
Price SSD per GB per month	0.10	0.22	Incl.
Price Hadoop per hour	0.07	0.04*	Incl.
Price per EMR node per hour	0.35	0.30	0.62
Costs			
Costs for VMs per day	88.03	**74.65**	155.50
Overall costs per month	2729.00	**2314.17**	4820.50

*Number of vCPU × Price per vCPU per hour

Engine services. However, the costs for additional SSD[7] storage are twice as high as the costs in *EMR*. Although the cluster configurations are comparable, the running time needs to be measured in order to estimate costs more accurately.

3.4 Data Warehouse

All three cloud providers offer managed solutions with data warehouse functionality for large-scale data analytics. Querying of massive amounts of data with SQL queries is one of their main features. Besides, the solutions support ETL processes based on different file formats. While the three cloud services offer similar functionality, different pricing approaches are used, as depicted in Table 5. In *Amazon Redshift* consumers define a computing cluster and are charged for the number and hours of utilized VM types. The available VM types are specifically designed for the purposes of *Redshift*. Each node comes with a fixed amount of local storage and it is not possible to separate storage capabilities. Consequently, a trade-off concerning the utilization of processing power and storage capacity will likely occur. That is, increasing the number of nodes for improving the querying performance might lead to unused storage capacities. Instead of charging for infrastructure components, Google's *BigQuery* prices storage capacity, streaming inserts,[8] and querying. Data operations like loading, copying, and exporting are free of charge. While *BigQuery* calculates costs based on the query capacity, which is dependent on the processed

[7] Abbr. for Solid-State Drive.

[8] Allows a direct transfer of data from Pub/Sub to BigQuery.

Table 5 Cost factors of data warehouse cloud services

Redshift	BigQuery	SQL data warehouse
• Number and type of VM cluster nodes	• Queries per capacity*	• Querying in DWU**
	• Amount of storage capacity	• Amount of storage capacity
	• Amount of streaming data***	

*According to the total amount of data processed in the selected columns
**A Database Warehouse Unit (DWU) measures the query performance
***Applies if streaming data is directly transferred to BigQuery

data volume per column and the column's data type, *Azure SQL Data Warehouse* (*DW*) charges in terms of query performance. Thus, the consumer is able to scale the speed of queries in *SQL DW*. As storage is charged separately per volume, consumers are only charged for the exact amount of storage needed to store table data. Due to the differences in the used pricing approaches, it is not possible to compare costs without determining the requirements of individual data analytics tasks. In general, we identify two main cost factors: compute and storage capacity.

3.5 Machine Learning

For predictive analytics, all three cloud providers provide managed machine learning engines that are integrated with both the storage and data warehouse solution of the respective cloud provider. *Amazon Machine Learning* (*ML*) and Google's *Prediction API*[9] focus on basic features to support common activities like the selection of data sources, explorative data analysis, model training, model evaluation, and model deployment. Less rigid is the approach of *Azure ML*, which provides an integrated SaaS application, referred to as *Azure ML Studio*, to individually define all steps of the data mining process as known from other data mining tools (e.g., *RapidMiner*). Besides, artificial intelligence algorithm APIs for vision, language, speech, and recommendations are available.

In general, we can identify three main cost factors: amount of computing capacity, number of transactions, and subscription fees. In Table 6, we give an overview of the individual cost factors per cloud service and provide two cost examples. Although the pricing approach is quite similar between *Amazon ML* and *Prediction API*, the main difference is that the latter calculates computing costs for training based on the volume of datasets. As the running time is closely linked to both the volume of training data and the complexity of the chosen learning algorithm, Google's pricing approach might be beneficial for the consumer as it does not consider complexity.

[9]Abbr. for Application Programming Interface.

Table 6 Cost factors and examples of machine learning cloud services (15 GB dataset size, 15000 streaming updates per day, 150 MB model size, 36 h of model generation, 30000 predictions per month)

Amazon ML	• Hours of data analysis and model training	
	• Number of batch predictions	
	• Number of real-time predictions	
	• Amount of reserved memory capacity	
Example ($)	Price per hour data analysis model training	0.42
	Price per real-time prediction	0.0001
	Price for reserved memory per 10 MB per hour	0.001
	Costs for computing capacities per month	15.12
	Costs for real-time predictions per month	41.16
	Overall costs per month	56.28
Prediction API	• Data volume for model training	
	• Number of streaming updates	
	• Number of predictions	
	• Number of projects (subscription)	
Example ($)	Monthly usage fee per project	10.00
	Price per MB bulk trained	0.002
	Price per streaming update	0.00/0.50*
	Price per prediction	0.00/0.05*
	Costs for computing capacities per month	30.00
	Costs for streaming updates per month	7.75
	Costs for predictions per month	145.00
	Overall costs per month	182.75
Azure ML	• Number of users (subscription)	
	• Number of compute hours	
	• Number of transactions	

*First 10000 predictions/10001 + predictions

Amazon ML does not imply any subscription fees and further differentiates between batch and real-time predictions. For allowing real-time predictions, reserved memory capacity needs to be rented for storing the model. However, streaming updates to further train the model, as supported by the *Prediction API*, are not possible. *Azure ML* charges a monthly subscription fee as well as an hourly fee for using its data mining tools. For its ML APIs, an additional fee per transaction and per computing hour occurs. Due to the different pricing approaches and functionality, it is not possible to precisely compare the cloud services with each other. However, our examples indicate that the number of predictions is an essential cost factor while the cost for generating the model are comparatively low. This emphasizes the need for estimating the business value, i.e., impact of predictions (e.g., for justifying, guiding, and prescribing business actions [13]), which becomes increasingly important, for instance,

to calculate the return of investment (ROI) of cloud-based big data applications. A comprehensive cost and performance benchmark study (as shown in, e.g., [4, 7]) is necessary to further evaluate the different machine learning cloud services.

4 Discussion

In the previous sections, we have shown that the proposed generic cloud-based big data architecture can be implemented in the cloud environments of the three market leading cloud providers. In general, we see that the cloud pricing approaches are quite similar, but also strongly differ in some aspects. To strategically select a cloud provider, consumers have to specify and estimate the characteristics and resource demands of use cases. Given the identified cost factors, a total cost of ownership (TCO) approach will help to better estimate the overall costs of using big data cloud services, for instance, to compare it with an inhouse solution. Besides costs for operating the cloud environment, a TCO approach may need to consider additional costs including costs for installing and configuration, support, and backsourcing. In general, we have seen that each cloud provider has its strengths and weaknesses, for instance, in terms of costs and flexibility. Cost benefits achieved in one cloud service can be exhausted by another cloud service, as shown in the previous sections. All cloud providers support linear scalability in such a way that cost functions are linear. Common pricing schemes for on-demand cloud services are unit-based and tiered pricing, mainly based on the compute or storage capacity. Other pricing schemes, such as for reserved cloud services, have not been considered in this study. Regarding our cost comparisons, we see that the cloud market leader, *AWS*, generally performs well and may be the best choice for implementing a big data streaming application. To better estimate and evaluate total costs, however, the dynamics of big data applications in terms of resource requirements need to be taken into account, for instance, using simulations. For this purpose, application profiling might be necessary [1]. Moreover, benchmark studies are necessary to evaluate the running time of computing clusters for performing certain data processing and data analytics tasks in order to be able to better compare associated costs. The study furthermore emphasizes the need of measures for estimating the value of big data analytics, for instance, in terms of ROI or service quality.

5 Conclusions and Outlook

While the demand for big data applications is growing with the continuous increase of digital data, cloud computing has become essential for meeting infrastructure requirements for big data in terms of computational power and storage capacity. The rapid development of managed big data cloud services makes it possible to support certain activities of big data processing in an on-demand fashion. These cloud ser-

vices need to be integrated to implement sophisticated big data application environments. Likewise, it is important to estimate the costs and value of those environments in order to make strategic decisions.

In this paper, we have presented a generic reference architecture for implementing big data application in cloud environments based on best practices. Moreover, we have defined three specific implementations of this reference architecture, applying cloud services of three leading cloud providers, and analyze important pricing schemes and cost factors. In particular for big data streaming analytics, we identify differences and similarities as well as strengths and weaknesses between cloud providers. Taking a cost perspective, important implications can be derived from the applied case studies. The case studies indicate the importance of an integrated view on big data application environments for estimating and evaluating the overall costs. Therefore, the contribution of this chapter is twofold. First, we proposed a state-of-the-art reference architecture and explained important aspects for implementing big data applications in cloud environments. Second, we analyzed relevant cloud services from a cost perspective and derived important implications for big data management.

Given the insights and implications of this study, the development of a holistic TCO model for big data applications is the object for future research. Moreover, we aim to integrate this model into a simulation framework in order to provide a tool for decision support that is able to consider the dynamics of big data applications in terms of usage patterns and resource demands.

References

1. Assunção MD, Calheiros RN, Bianchi S, Netto MA, Buyya R (2015) Big data computing and clouds: trends and future directions. J Parallel Distrib Comput 79:3–15
2. AWS (2016) Big data analytics options on aws. https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf
3. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171–209
4. Chen Y, Alspaugh S, Katz R (2012) Interactive analytical processing in big data systems: a cross-industry study of MapReduce workloads. Proc VLDB Endowment 5(12):1802–1813
5. Creeger M (2009) Cloud computing: an overview. ACM Queue 7(5):2
6. Gartner (2015) Magic quadrant for public cloud storage services, worldwide. http://www.gartner.com/technology/reprints.do?id=1-2IH2LGIŹct=150626Źst=sb
7. Ghazal A, Rabl T, Hu M, Raab F, Poess M, Crolotte A, Jacobsen HA (2013) BigBench: towards an industry standard benchmark for big data analytics. In: Proceedings of the ACM SIGMOD international conference on management of data. ACM, New York, NY, USA, pp 1197–1208
8. Heilig L, Lalla-Ruiz E, Voß S (2016) A cloud brokerage approach for solving the resource management problem in multi-cloud environments. Comput Ind Eng 95:16–26
9. Heilig L, Voß S (2014) Decision analytics for cloud computing: a classification and literature review. In: Newman A, Leung J (eds) Tutorials in operations research—bridging data and decisions. INFORMS, San Francisco, pp 1–26
10. Heilig L, Voß S (2014) A scientometric analysis of cloud computing literature. IEEE Trans Cloud Comput 2(3):266–278

11. Jensen M, Schwenk J, Gruschka N, Iacono LL (2009) On technical security issues in cloud computing. In: Proceedings of the IEEE international conference on cloud computing (CLOUD). IEEE, Bangalore, India, pp 109–116
12. Krishnan S, Tse E (2013) Hadoop platform as a service in the cloud. Technical report, Netflix. http://techblog.netflix.com/2013/01/hadoop-platform-as-service-in-cloud.html
13. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big data, analytics and the path from insights to value. MIT Sloan Manage Rev 52(2):21
14. Li M, Tan J, Wang Y, Zhang L, Salapura V (2015) SparkBench: a comprehensive benchmarking suite for in memory data analytic platform Spark. In: Proceedings of the 12th ACM international conference on computing frontiers (CF). ACM, Ischia, Italy, pp 53:1–53:8
15. Maravić I (2016) Spotify's event delivery—the road to the cloud (part III). https://labs.spotify.com/2016/03/10/spotifys-event-delivery-the-road-to-the-cloud-part-iii/
16. Shvachko K, Kuang H, Radia S, Chansler R (2010) The Hadoop distributed file system. In: Proceedings of the 26th IEEE symposium on mass storage systems and technologies (MSST). Incline Village, NV, USA, pp 1–10
17. Talia D (2013) Clouds for scalable big data analytics. IEEE Comput 46(5):98–101
18. Tallon PP (2013) Corporate governance of big data: perspectives on value, risk, and cost. Computer 46(6):32–38

The Strategic Business Value of Big Data

Marco Serrato and Jorge Ramirez

Abstract Most of the information about Big Data has focused on the technical side of the phenomenon. This chapter makes the case that business implications of utilizing Big Data are crucial to obtain a competitive advantage. To achieve such objective, the organizational impacts of Big Data for today's business competition and innovation are analyzed in order to identify different strategies a company may implement, as well as the potential value that Big Data can provide for organizations in different sectors of the economy and different areas inside such organizations. In the same vein, different Big Data strategies a company may implement towards its development are stated, as well as insights on how enterprises such as businesses, non-profits, and governments can use data to gain insights and make better decisions. Current and potential applications of Big Data are presented for different private and public sectors, as well as the ability to use data effectively to drive rapid, precise and profitable decisions.

Keywords Big data · Analytics · Decision-making · Strategy

1 Introduction

Nowadays, the volume of data surpassed 2.8 zettabytes and is expected to continue its rise to a 50 times higher volume in 2020. The ability to generate and access this huge amount of data has been propelled by the increasing number of people, devices, and sensors that are now connected by digital networks and at such networks more than 30 million sensor nodes are now present in different economy sectors: transportation, automotive, industrial, utilities, retail, health services, among others. The number of such sensors is increasing at a rate of more than 30 %

M. Serrato (✉) · J. Ramirez
Tecnológico de Monterrey, Monterrey, Mexico
e-mail: mserrato@itesm.mx

J. Ramirez
e-mail: jorge.ramirez@itesm.mx

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_3

a year [1] and Big Data is being used for processing and analyzing such large amounts of heterogeneously structured and un-structured data. This data explosion represents a new opportunity and challenge for enterprises and organizations, as far as few of them understand how to extract insights, value and manage these ever-increasing amounts of data [2]. The overarching disruptive power of Big Data demands that organizations engage with it at a strategic level, since it is capable of changing competition by transforming processes, altering corporate ecosystems, and facilitating innovation while creating strategic value for organizations.

Three challenges arise for managers and decision-makers in order to take advantage of Big Data. The first is to think critically about these techniques and the analyses based on such data—whether conducted by the organization or by someone else [3, 4]; the second is to identify opportunities for creating value using Big Data [1, 4, 5]; while the third one is to estimate the value created while using Big Data to address an opportunity [2, 6, 7]. As stated throughout this chapter, a specific foundation is required to face such challenges, as well as to understand and apply these methods to drive value for an organization. Big Data is not a theoretical discipline: these techniques are only interesting and important to the extent that they can be used to provide real insights and improve the speed, reliability, and quality of decisions. The concepts presented in this chapter allow the identification of opportunities in which Big Data can be used to improve performance and support important decisions, as well as to be alert to the ways that Big Data can be used—and misused—within an organization.

First, the strategic and organizational opportunities, benefits and impacts of Big Data for today's business competition and innovation are outlined in this chapter. In the same vein, different Big Data strategies a company may implement towards its development are also presented. Based on these elements, insights on how enterprises such as businesses, non-profits, and governments can use data to gain insights and make better decisions are also outlined, in order to present current and potential applications of Big Data in several functions. Finally, conclusions and recommendations to use data effectively to drive rapid, precise and profitable decisions, are described.

2 Strategic and Organizational Opportunities

The everyday consumer world in the near future will look radically different from today's [8]. Many ordinary products and devices like heating systems, televisions, cars, watches, toys, light bulbs, sporting goods, home appliances, among others, will have gone digital. They will be connected to the Internet and to each other in altogether new ways. Consumers will increasingly access, monitor, and control their connected digital products and services remotely over the Internet, using smartphones, tablets, laptops, desktop PCs, and other devices. Massive streams of complex, fast-moving big data from these digital devices will be stored as personal profiles in the cloud, along with related customer data. Such changes will

significantly modify the way our world works, while introducing new opportunities and challenges for organizations and individuals to compete and create value through Big Data.

Data generation in our world has already reached an exponential growing rate that shows no signs of abating and which generates the Big Data opportunities and challenges that individuals and organizations are already facing today. The technology, media, and telecommunications sectors present significant advances in this vein, while financial services, health care, and consumer goods are still on early stages in this new era.

Energy, industrial goods, construction, and public services are arguably just behind the most developed sectors under Big Data, and about to start their steepening ascent. Moreover, industries are increasingly colliding as digital moves beyond screens and software to enter the world of things and businesses.

Fast-changing landscapes of mobility and transportation provide illustrations on how automotive, technology, and start-up companies are both complementing and competing with one another nowadays. Their strategies suggest new ways to think about mapping the landscape, deciding where to play, and embarking on the journey of Big Data while taking advantage of the arising opportunities and facing new challenges at the same time.

2.1 Mapping a New Position on a Dynamic Industry

Three fundamental questions related to Big Data need to be asked inside an organization, in order to understand and map its strategic position under today's and tomorrow's involvement under this field. The first question is '*What can be forecasted?*, which contributes to prepare and open the mind, analyze the breadth of digital and Big Data trends, their expected time frames, and potential tipping points as what is possible evolves under this new dynamic world. The second question is '*Where can our organization be disrupted through Big Data?*' which provides the foundations on a search for vulnerable profit pools, as well as ways to use their own assets and capabilities to be disruptive in other sectors. Finally, the third question is '*What can our organization shape and where do we need to adapt?*' which creates awareness on the way brands, distribution networks, supplier relationships, strong capabilities, and superior cost positions need to be changed or improved in a Big Data world [9].

The organizational challenge arises on figuring out when to draw on existing or latent strengths to shape a market, and when to acquire new capabilities or adapt to ambient forces.

Charting the current fields for competition is necessary but insufficient. It's also essential to frame, explore, and prioritize strategic choices in the near future. As stated by Gerbert et al. [10], Big Data opportunities can be framed through two dimensions: *reengineering the value chain* and *reimagining the offering*.

The first one can vary in complexity and impact, since it may range from a single value-chain step to processes that cross multiple corporate functions and even the organization's boundaries. Big Data and analytics that improve sales effectiveness across physical, mobile, and online channels can promote significant value creation. Also, the development of entirely automated order-to-delivery processes represents an opportunity in this vein. In the back office, human resources management and finance processes typically offer opportunities for immediate optimization. *Reimagining the offering* is a more open ended dimension which requires creativity and vision. Big Data creates ample opportunities for novel products and services, and these innovations typically exploit new data and powerful analytics that allow new offerings created by several organizations.

2.2 A Dynamic Organization in a Dynamic World

Striving to sustain a competitive advantage demands a perpetual process of transformation as today's game quickly morphs into tomorrow's. The ongoing cycle has three stages that are relevant to outline. First of all, organizations need to get involved in an *Enhancement* process, which is about extrapolating from your current position while creating immediate value. It is the least radical stage of Big Data opportunities; however, it can improve the organization's Big Data skills and provide tremendous and immediate value creation that can fund the broader Big Data journey. Examples include predictive maintenance, streamlined digital links to suppliers and customers that provide Big Data for better and more accurate decision-making, as well as recommendation engines.

In the same vein, organizations need to go further and fall into an *Exploration* stage, which requires investigating offerings adjacent to the current business or pursuing larger adjustments of the value chain. Exploratory digital strategies become C-suite topics. Companies need to invest significant resources in Big Data infrastructure and technology, and closely track their performance. Finally, a *Transformation* stage is relevant for organizations that need to face deeper changes, since it is an all-encompassing strategic move that has the greatest potential to generate competitive advantage, often over several years, but also the greatest risk. From a strategic standpoint, an organization needs to envision a target state five to ten years out and then "retropolate" from that vision back to the present. Such transformation requires major investments and often the development of new partner ecosystems.

To achieve these vision, Senior decision makers have to embrace evidence-based decision making. Companies need to hire scientists or organizations whom can find patterns in data and translate them into useful business information. And whole organizations need to redefine their understanding of "judgment." To achieve so, Decision Support Systems are a valuable asset for analysts since they transform

performance data into useful information, and in turn, such information is transformed into knowledge to support decision-making.

However, traditional Business Intelligence and Decision Support Systems tools require something more than the use of mere historical data and rudimentary analysis tools to be able to predict future actions, identifying trends or discovering new business opportunities. Reducing the time needed to react to noncompliant situations can be a key factor in maintaining competitiveness. Real-time, low latency monitoring and analyzing of business events for decision making is key, but difficult to achieve. The difficulties are intensified by those processes and supply chains which entail dealing with the integration of enterprise execution data across organizational boundaries.

One of the most critical aspects associated to these Big Data opportunities and challenges is its corresponding impact on how decisions are made and who gets to make them. When data are scarce, expensive to obtain, or not available in digital form, it makes sense to let well-placed people make decisions, which they do on the basis of experience they have built up and patterns and relationships they have observed and internalized.

However, it is important to acknowledge that throughout the business world today, several people and organizations still rely too much on experience and intuition and not enough on the data-based decisions they are capable of. Executives interested in leading a Big Data transition can start with two simple techniques. First, they can get in the habit of asking "*What do the data say*?" when faced with an important decision and following up with more-specific questions such as "*Where did the data come from*?," "*What kinds of analyses were conducted*?," and "*How confident are we in the results*?". Second, they can allow themselves to be overruled by the data; few things are more powerful for changing a decision-making culture than seeing a senior executive concede when data have disproved a hunch [9].

When it comes to knowing which problems to tackle, domain expertise remains critical. Traditional domain experts, specially those deeply familiar with an area, are the ones who know where the biggest opportunities and challenges lie.

Ultimately, strategy is about choice since not everything can be done. An organization can not even do all the things it should do simultaneously. It has to make conscious choices to prioritize and stage initiatives. Likewise, data is the opposite of oil in that one is ancient and scarce and the other is new and growing exponentially. In fact, the most valuable data does not rest in dusty databases but has yet to be created. Data is a classic example of one of tomorrow's highly dynamic multilevel games. Sustained success requires actively managing a portfolio of initiatives across time. A useful analogy might be the famous Nash equilibrium in game theory [11]. The optimal strategy is always to focus on each action for a part of your time. Your strengths and weaknesses will determine their relative weight.

3 Big Data Strategies

Nowadays technology has changed rules of business competition and it involves the way as companies seek for its competitive advantage. In this modern revolution, companies compete in open networks producing shared public goods and creating shared value amongst networks [7, 12].

The sharply declining cost per performance level of computing power and data storage has boosted this revolution [4]. Every form of information technology has been doubling performance, bandwidth, capacity and halving price every 12 months [13], leading in 1 trillion more products/devices connecting to the Internet across industries such as manufacturing, health care, and mining [14].

The use of the data coming of those millions of networked sensors integrated in such products, empowers businesses to learn about their customers, suppliers, and operations. Big Data is characterized by this ongoing increase in volume, variety, velocity, and veracity of data [2].

Most of these data arise from user generated content in an unstructured way (graph data, voice, images, video, etc.). According to Ebner et al. [2] in general there are four data categories, namely; *External structured* data such as credit history data, *Internal structured* data such as inventory data, *External unstructured* data such as Facebook or Twitter posts, as well as *Internal unstructured* data such as text documents and sensor data.

The minimum data size, which qualify as a Big Data is moving definition as the available technology evolves. While 1 TB (= 10^{12} bytes = 1000 GB) was *huge* a few year back, right now 1 PB (= 10^{15} bytes = 1000 TB) is the definition of *huge* for most, and is moving to 1 Exabyte (= 10^{18} bytes = 1000 PB).

Modern literature points to Big Data is so large that you can't adequately store and process it with your regular technology and requires advanced techniques and technologies to capture, store, distribute, manage, and analyze these data [2] as shown in Table 1.

Table 1 Examples of Big Data

Size	Manage with	How it fits	Examples
1 TB −1 PB	Hadoop, analytics, machine learning, and cognitive computing	Stored across company distributed clusters	eBay uses two data warehouses at 7.5 petabytes and 40 PB as well as a 40 PB Hadoop cluster for search, consumer recommendations, and merchandising
1 EB	Object store. Cloud services	Technologies on cloud containers (Darrow 2015)	Billions of web clicks. At the moment, every day 1 EB of data is created in the internet, that is the equivalent to 250 millions of DVDs

Source Adapted from Driscoll [15], Factor [16] and van Rijmenam (2016)

As Table 1 shows, the general explosion of the data quantity in all industries has outgrown conventional technologies, unstructured data for example do not fit traditional database schemes [16] so a new generation of data tools is needed.

Besides this technological requirements, is completely recognized that the lack of clear rules, skills and HR are important inhibitors to unlock the business value from Big Data, but in spite of the importance of the business Big Data strategy topic, very little has been written about it. The vast majority of literature (34 %) points to 'technology and techniques' issues, followed by 'access to data' issues (28 %), while a small amount of the research deals with 'Organizational change and talent' issues (16 %) and 'data policies' (9 %) [17].

Companies need a Business Big Data digital strategy to cope with these ever-increasing amounts of data, therefore they can capitalize on their data in order to gain a competitive advantage [5].

Modern strategy theory, state that competitive advantage is not something a company owns, instead it is a position game-changer innovation as key to business success [7]. At its core, business analytics is about leveraging value from data. How extracting value from data requires aligning strategy to business performance management. Both of them (business strategy and bpm) are driven by decision based on statistical models that fit, optimize, and predict the data using analytic/BI/visualization technologies [4, 18] and all this is supported on business technological infrastructure [19] as shown in Fig. 1.

The idea behind Big Data Digital Strategy is to transform data in information and knowledge; this isn't just about infrastructure but also about fundamentally transforming the company's business [3, 5]. As Dreischmeier et al. [5] alleged "to ensure the delivery of value to customers best-in-class companies think end-to-end, not only in analytics applications but also in operations and the back office, and they tackle many efforts in parallel, using standardized processes and agile techniques—business performance—to accelerate execution and inject more flexibility into strategy".

This flexibility is needed due that pace of disruption is rapidly increasing. Leading companies are increasingly under attack from new incumbents that are out to reinvent businesses and industries by addressing consumer needs in completely new ways, as Uber did in the taxi business and Airbnb in travel, even companies such as Facebook, itself once a disrupter are themselves constantly under attack [5].

Due to the rapidly evolving competitive business landscape, "agile" methods are required instead a long-term strategy. Leading digital companies test and refine, or prototype, products and strategies in close cooperation with customers using techniques like the Agile Approach for Digital Transformation from BCG Analysis (2015) shown in Fig. 2.

A company must take into account both; external environment and internal capabilities to support digital transformation. As a result, digital strategy is an iterative process that is continually adapting to and seize new opportunities.

According to literature [5, 10, 20] leading companies typically do three things to understand and map the strategic landscape;

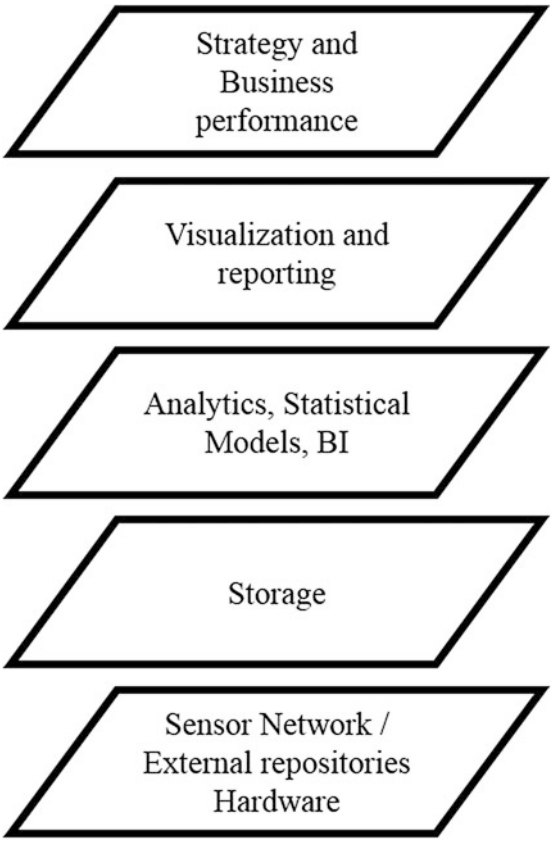


Fig. 1 Big Data Layers. *Source* Adapted from [4, 6, 19]

First, they create a role for a chief digital officer CIO, whose mandate is running a transformation and who is able to make decisions above and beyond the technology architecture and the traditional processes. One of the CIO's main tasks is to analyze the breadth of digital trends, their expected time frames, and potential tipping points as what is possible evolves.

Second, based on former CIO analysis, they do a diagnostic to understand the sources of value that can be captured from digital. At the same time, they'll look at the threat they face from not making those investments. That picture helps the organization prioritize its focus area or examine how their own assets and capabilities could be disruptive in other sectors and decide to what can they shape and where do they need to adapt,

And *Third*, companies typically apply "agile" strategies in place a long-term strategy. They learn by doing, test and refine, or prototype products and services in close cooperation with customers and at a dizzying pace. As an example Amazon has introduced e-readers, tablets, smartphones, cloud services, delivery services, and online marketplaces—all within the past ten years.

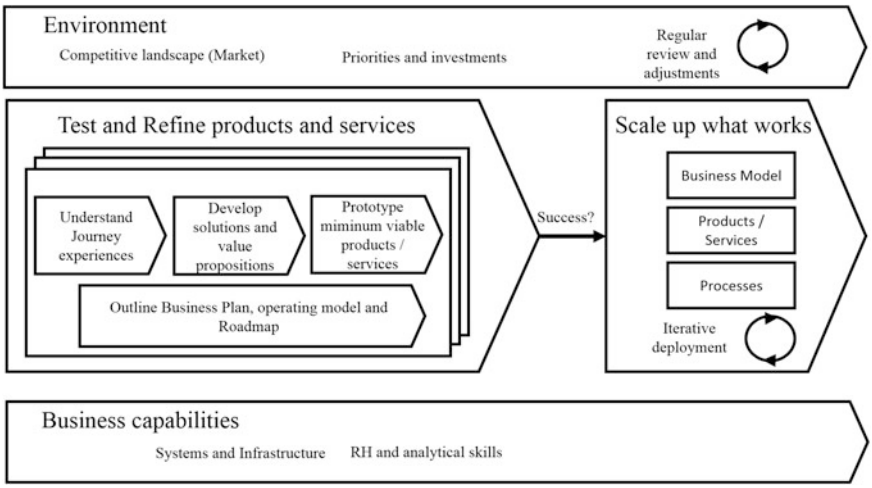


Fig. 2 Agile approach for digital transformation. *Source* Adapted from BCG Analysis (2015)

The Economist Intelligence Unit (2015) identified four categories of companies based on the level of sophistication of their use of corporate data:

- *Strategic data managers*: companies that have well-defined data-management strategies that focus resources on collecting and analyzing the most valuable data;
- *Aspiring data managers*: companies that understand the value of data and are marshalling resources to take better advantage of them;
- *Data collectors*: companies that collect a large amount of data but do not consistently maximize their value; and
- *Data wasters*: companies that collect data, yet severely underuse them.

Regardless of the widespread recognition of the criticality of data for delivering value to customers, most companies are still in the early stages of implementing and adopting a corporate Big Data Strategy[1] [21] which is important due to the data exploiting category on which the organization is found, it is based on its existing business model [12]. Business models are abstractions of real life used to describe all real world businesses, or particular types of business with common characteristics, or a very particular real world business model (like the Apple model) [22].

Big Data driven business model, describes a set of businesses which rely on big data to achieve their key value proposition and to substantially augment their value proposition to differentiate themselves in order to gain competitive advantage [7]. The maturity model describes a course of development/evolution (where to start

[1]Interestingly leading companies' self-perception is different. At the Economist Intelligence Unit research, most of the companies classifying themselves as a strategic data manager indicating that they have developed a well-defined corporate data strategy.

and where to go) that organizations need to know follow when they embark on big data initiatives [23].

IBM has formed a Data Governance Council to define and issue master data quality policies and one of the initiatives from this council is the *data governance maturity model* as shown in Fig. 3 [24].

The maturity models describe the relevant milestones in the process of big data governance and is related to the level of sophistication of how companies use their corporate data as shown in Fig. 3. While this framework present the relationships among the concepts, a methodology is needed by which organizations can move from one level to another and travel up the maturity model. El-Darwiche et al. (2014) resumes this in a Big Data business model as it's shown in Fig. 4.

How organizations will apply its Big Data Strategy depends not just on their current maturity level—recognize it, is necessary but insufficient. It's also essential

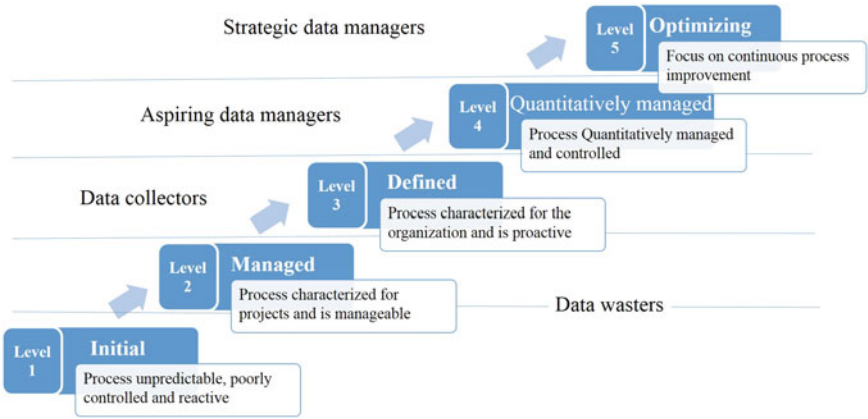


Fig. 3 The IBM Data Governance Council Maturity Model. *Source* Adapted from; The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance (2007)

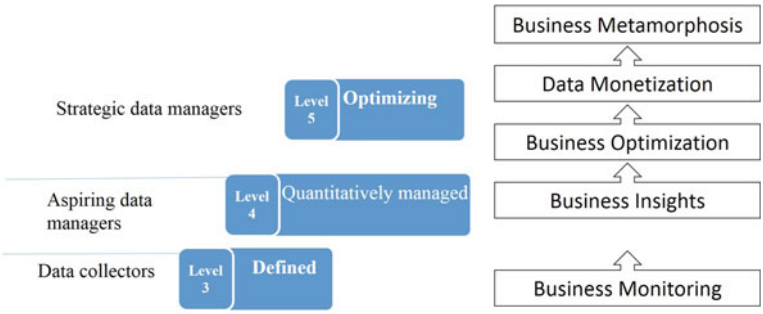


Fig. 4 The maturity phases of Big Data business model. *Source* Adapted from El-Darwiche et al. (2014)

to identify and prioritize strategic choices on how organizations can move from one level to another. An intuitive tool is the digital strategies matrix, which comprises two axes: Business Maturity Model (internal business capabilities) and Market (external environment) See Fig. 5.

A company must take into account that the optimal strategy focus on both; the external environment and internal capabilities of the organization and is always an iterative process that is continually adapting to those external and internal factors as was mentioned above.

The vertical axe of the digital strategies matrix denotes the maturity phase in which the organization is situated. The horizontal axe is a continuum of (1) the existence of an efficient market for ideas and ownership of valuable complementary assets and (2) the role played for an organization in the competitive market, running from organization using data (and technology) in a tacitly and utilitarian way to an organization using data (and technology) in a strategic and disruptive way. This matrix is related to the original Ansoff Matrix [27] that focused on the firm's present and potential products and markets.

Penetration. In mature businesses where technology plays a tactical supporting role, cost containment and risk minimization are priorities. The enterprise uses technology and data as defense mechanisms to maintain an appropriate level to continue providing adequate and reliable service levels in an inefficient market, where companies don´t have incentives to innovate. Firms seek to achieve growth with existing strategies in their current market segments, collecting data but without exploiting it, aiming to increase its market share [25].

Extension. Firms seek growth by targeting to new market segments and they are risk taker. At the end of the continuum (taking the best case) the companies cooperate sharing and buying ideas. However in these enterprises, the core business

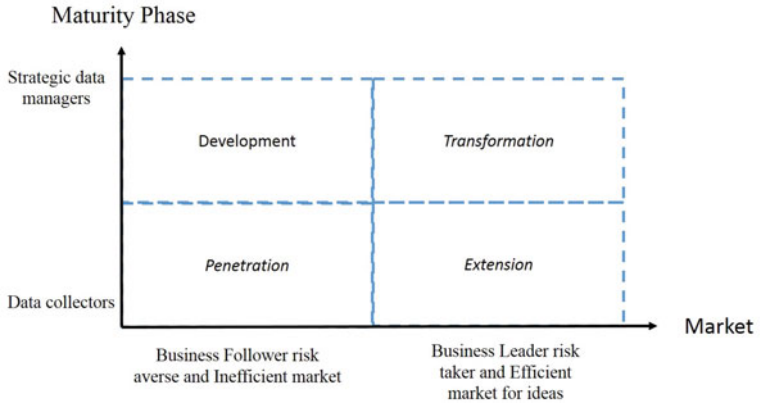


Fig. 5 The digital strategies matrix. *Source* Author's own elaboration based on *IT Value Performance Tools Link to Business-IT Alignment* Gartner Research [25], Growth matrix with Big Data generated options [26] and *the digital opportunity matrix* BCG [10]

processes have often great opportunities for optimizing business performance, due to a lack of use of data [26].

Development. As stated above, most companies are still in the early stages of implementing and adopting a corporate Big Data Strategy. In this quadrant companies have started using data in the strategic layer, using for example predictive maintenance, streamlined digital links to suppliers and customers, and recommendation engines. However firms target to its existing market segments. While this is not the optimal stage of digital opportunities, it provide tremendous and immediate value creation that can fund the broader digital journey [28].

Ecosystem Transformation. Gerbert et al. [10] suggest that transformation is "an all-encompassing strategic move that has the greatest potential to generate competitive advantage, often over several years, but also the greatest risk". Transformative change addresses to the disruptive result of the third point mentioned above on how leading companies typically do three things to understand and map the strategic landscape. Transformation requires major investments and often the development of new partner ecosystems.

4 Decision Making Under Big Data

Organizations used to trust on the managers intuition to make business decisions, due to data were scarce, expensive to obtain, or not available in digital form. Usually those were the top managers in the organization (known as HiPPO—the highest-paid person's opinion).

Nowadays, many companies often make most of their important decisions by relying on the HiPPOs. However relying fully on experience and intuition has proven not be sufficient, a new culture of decision making is emerging where business decisions are data-driven [9].

Data-driven decisions are better decisions, relying on evidence of (an unmatched amount of) facts rather than intuition by HiPPOs or experts [7, 12]. In this sense Big Data is an enabler of improved decision making for enhanced firm performance [17].

Companies characterized themselves as data-driven, have a better performance on objective measures of financial and operational results. For example, Manyika et al. [1] reported that a retailer embracing big data has the potential to increase its operating margin by more than 60 % and McAfee & Brynjolfsson [9] reported that companies data-driven were, on average, 5 % more productive and 6 % more profitable than their competitors. Other research shows that leaders in the use of big data generate 12 % higher revenues than companies that don't experiment with big data [5].

This results evidence that intelligently using their data, organizations are able to improve their decision-making and better realize their objectives. Organizations may lose competitiveness by not systematically analyzing their available information coming from data [2, 5].

In order to take advantage of this competitiveness, companies must have achieved the higher maturity level, those are companies whose execution of key business strategies through utilization of data are considered as better than most other companies in its sector [7].

At lower maturity levels, data tends to be fragmented across the enterprise and out of date, so that a great volume of potentially helpful data—in both structured and unstructured form—is never used [29].

In any case, the availability of accurate, real-time management data is critical to decision making, that's the reason more and more organizations are using Operational intelligence platforms not just to collect and process operational data but also to present it in a clear, consistent, readily available manner throughout the organization—improving the speed and the quality of decision making [30].

The growing number of data-driven businesses is driving increasing interest in Decision Management. Gartner [25] recognize it as a discipline with commercial software products, consulting services, books, formal techniques for decision modeling and extensive academic research. Decision management is particularly relevant to automated decisions and repeatable operational decisions [31].

The realization that fact-based decisions are more critical at every level of the organization has resulted in the emergence of technological tools as Self-Service Data Preparation,[2] and Business analytics platform as a service (baPaaS).[3] Besides, advanced analytics techniques have been incorporated into enterprise systems, namely; Smart Data Discovery,[4] Natural-Language Generation,[5] Hadoop-Based Data Discovery[6] and Event Stream Processing.[7] However Gartner analysis (2011) considers it will take more than five years for those technologies to reach the Plateau of Productivity and become part of the majority of large, new development projects for BI.

As Morabito [6] argued

[2]Self-service data preparation tools enable business users to run machine-learning algorithms on the data to visually highlight the structure, distribution, anomalies and repetitive patterns in data with guided intelligent capabilities to recommend ways for users to improve their data.

[3]In Business analytics platform as a service (baPaaS) solutions are architected with integrated information management and business analytics stacks. These comprise database, integration capabilities and business analytics tools—or solutions that include only business analytic tools (for example, reporting and dashboarding)—leveraging autonomous cloud-based or on-premises data repositories.

[4]Smart data discovery is a next-generation data discovery capability that enables business users or citizen data scientists to find insights from advanced analytics on multistructured data.

[5]Natural-language generation (NLG) combines natural-language processing (NLP) with machine learning and artificial intelligence to dynamically identify the most relevant insights and context in data (trends, relationships, correlation patterns).

[6]Hadoop-based data discovery enables business users to explore and find insights across diverse data (such as clickstreams, social, sensor and transaction data) that is stored and managed in the Hadoop Distributed File System (HDFS).

[7]Event stream processing is a computing technique in which incoming events are processed to generate higher level, more useful summary information (complex events).

> Big data can give companies the ability to target each customer individually based on their preferences and purchasing habits, by integrating personal information about website browsing, purchase histories, physical position, response to incentives, as well as demographic information such as work history, group membership and people's views and opinions based on social influence and sentiment data.

In the Big Data era, organizations make better decisions by analyzing data sets from customers or even sensors embedded in products, and they facilitate access (time and place) of resulting information to the relevant decision maker. People who understand the problems need to be brought together with the right data.

The ultimate objective in this sense is delivering the right information, at the right time, in the right place, in the right way, to the right person [32] in order to improve efficiency and effectiveness in organization to increase the value-added content of products and services [1].

Benefits currently derived or expect to be derived from using Big Data analytics encompass various areas of business, being the top benefit related to improving and speeding up the decision-making process [33] replacing/supporting human decision making with automated algorithms [1].

Big Data benefits of course go besides decision making, as indicated by Manyika et al. [1]

> at the heart of big data lies tremendous potential to transform the way companies operate and creates value in several ways;
>
> - Creating transparency,
> - Enabling experimentation to discover needs, expose variability, and improve performance
> - Segmenting populations to customize actions
> - Innovating new business models, products, and services

From now on, use of big data will become a key basis of competition and growth for organizations.

5 Industry Applications

Big Data is playing a key role in the new business, government and academic environments. Such environments usually encompass a network of companies, individual contributors, institutions, and customers that interact to create mutual value chains operating effectively under the same or similar business models. 5 network of companies (industries) have been identified [1, 34] where Big Data is playing a key role, enhancing or transforming the industries' performance as shown at Table 2.

These environments vary in their sophistication and maturity in the use of big data, together represented 39 % of global GDP in 2010. In lines below, the Big data opportunities and implications related to those five environments are examined.

Table 2 Industries using Big Data

Industry	Main Business functions impacted
Retail	Cross selling
	location-based marketing
	Sentiment analysis
	Supply chain optimization
Manufacturing	Supply chain real time information
	Process analysis using Networked sensors
	Distribution optimization
	Predictive maintenance
Telecommunications	Network optimization
	Churn prevention
Public Sector	Boost productivity
	Create transparency
Healthcare	Bioinformatics
	Prediction/Simulation
	Personalized medicine
	Evaluation drug efficacy

Source Adapted from *Big data: The next frontier for innovation, competition, and productivity. McKinsey (2011) and Big Data Driven Business Models Morabito (2015)*

5.1 Retail

The use of Big Data in the retail industry is boosting the productivity of the entire sector. As the volume of data is growing exponentially, retailers are becoming more sophisticated using data they collect from every customer transaction and online customer behavior and sentiments. Retailers mining customer data to inform decisions they make in order to optimize their supply chain to merchandising and pricing [1]. With that information about customers, retailers improve customer service to increase customer intimacy and loyalty [7].

There are fundamental implications of Big Data on the retail industry. For example customer's data (demographics, purchase history, preferences, real-time locations) are used in cross-selling to increase the average purchase size. In the other hand, PlaceCast [35] reported the use of location-based marketing strategy using personal location data-enabled mobile devices that targets consumers who are close to stores or already in them. Once located the store sends a special offer to the customer's smartphone. Use of this technologies allows store to analyze data on in-store behavior, tracking customers' shopping patterns.

A new and powerful opportunity has rose from the social media data analysis. Consumers are relying increasingly on peer sentiment and recommendations to make purchasing decisions so data generated by consumers in the various forms of social media, helps to inform to the store in its business decisions. A variety of tools

(as IBM Watson Analytics sentiment analysis) has emerged for the real-time monitoring and response to Web-based consumer behavior and choices [1].

At supply chain, retailers are optimizing distribution and logistics by using GPS-enabled big data telematics (i.e., remote reporting of position, etc.) and route optimization to improve their fleet and distribution management. Those actions improve productivity by optimizing fuel efficiency, preventive maintenance, driver behavior, and vehicle routing [1].

The value that players in the retail sector and their customers will actually capture will depend critically on the actions of retailers to overcome barriers related to technology, talent, and organizational culture. While Big Data linked to new technology does squeeze the retail industry in some ways, it also offers significant new opportunities for creating value. Sector retailers and their competitors are in a constant race to identify and implement those Big Data levers that will give them an edge in the market. The volume of data is growing inexorably as retailers not only record every customer transaction and operation but also keep track of emerging data sources such as radio-frequency identification (RFID) chips that track products, and online customer behavior and sentiment [1, 29].

5.2 Manufacturing

According to Manyika et al. [1] "the manufacturing sector was an early and intensive user of data to drive quality and efficiency, adopting information technology and automation to design, build, and distribute products since the dawn of the computer era … getting impressive annual productivity gains because of both operational improvements that increased the efficiency of their manufacturing processes and improvements in the quality of products they manufactured … but despite such advances, manufacturing, faces the challenge of generating significant productivity improvement in industries that have already become relatively efficient".

By using real-time data, companies can manage demand forecasting and supply chain planning, reducing defects and rework within production plants. Manufacturers are embedding real-time data from networked sensors in the production processes. These data allows to reduce waste and maximize in yield or throughput the supply chain, providing a means to achieve dramatic improvements in the management of the complex, global, extended value chains that are becoming prevalent in manufacturing [36].

Analyzing the data reported by such networked sensors embedded in complex products, enables manufacturers of aircraft, elevators, and data-center servers to create proactive smart preventive maintenance service packages transforming in this way the commercial relationship with customers from one in which they sell a product to one in which they sell a service, a technician can be dispatched before that a component is likely to fail [34].

Manyika et al. [1], reported how in the digital oil industry, a single system captures data from well-head flow monitors, seismic sensors, and satellite telemetry systems. The data are transmitted to very large data farms and then relayed to a real-time operations center that monitors and adjusts parameters to optimize production and minimize downtime. This preventive maintenance can cut operational costs by 10 to 25 % even while potentially boosting production by 5 % or more.

Big data can underpin another substantial wave of gains, and such gains will come from improved efficiency in design and production, further improvements in product quality, and better meeting customer needs through more precisely targeted products and effective promotion and distribution [37]. Overall, Big Data provides a means to achieve dramatic improvements in the management of the complex, global, extended value chains that are becoming prevalent in manufacturing and to meet customers' needs in innovative and more precise ways, such as through collaborative product development based on customer data [1].

Some automotive suppliers are rapidly achieving step-change improvements in productivity by embracing a digital transformation based on Big Data for its production processes, ranging from quick wins such as automated material handling to investments in collaborative robotics. Even more, automotive industry newcomers are using new digital tools that facilitate fast prototyping, as Google used Big Data to prototype its new innovative fully autonomous vehicle or Tesla who reconfigured automotive distribution with single-car showrooms, all-online sales, and home delivery of vehicles [10].

Other new offerings are spurring more far-reaching industry changes. Uber, for example, is seeding and shaping the sharing economy for the transport of people (and soon goods), blurring the difference between contractors and employees. None of these developments are fundamentally transforming the core offering of traditional automakers, but they do start affecting demand and supply. Peripheral industries, however, are being transformed by these developments. The new- and used-car-dealer markets are under threat from all-digital pricing platforms such as TrueCar and all-inclusive online consignment services, such as Beepi or Shift.

5.3 Telecommunications

Big Data enable telecommunications industry to use their internal data to optimizing routing and quality of service by analyzing network traffic in real time, analyzing call data records in real time to identify fraudulent behavior immediately and using insights into customer behavior and usage to develop new products and services [38]. But the potential of big data poses a different challenge: how to combine much larger amounts of information to increase revenues and profits across the entire telecom value chain, from network operations to product development to marketing, sales, and customer service—and even to monetize the data itself [39].

Also Big Data provide optimal churn prevention. A customer complaint or service quality problem would trigger a targeted and customized offer that is more attractive to a subscriber, decreasing the propensity of this subscriber churning. Even more, rotational churn identification is propelled identifying and preventing mobiles subscribers that disconnect and reconnect their service in order to take advantage of promotions that only apply to new customers [40].

Other industries are being transformed by developments on the telecommunications industry. Crowdsourced real-time traffic information provided by mobile devices such as service provide by Waze or Uber shaping the sharing economy for the transport of people. Daimler is enhancing both the connected-car and autonomous-driving features of its vehicles through its mobile platform [10].

Mobility industry

Examples of reimagination based on Big Data abound in the mobility industry. Condition monitoring based on car and train sensor data, for instance, offers the opportunity to enhance existing maintenance. So do simple services, such as crowdsourced real-time traffic information provided by Waze or Inrix [10].

Daimler, Audi, BMW and GM for example, are enhancing both the connected-car and autonomous-driving features of its vehicles [41]. Daimler also is exploring car sharing with car2go[8] and has acquired mytaxi,[9] an innovative ride-matching platform in the German cab market. Daimler through its Moovel platform—which integrates car-2go and mytaxi into a broader route-planning ecosystem—and Audi through intermodal route planner are striving to shape intermodal travel based on Big Data technologies [10, 42].

5.4 Public Sector Administration

While productivity in the public sector is not easy to measure, there is evidence that public sector productivity growth has fallen behind that of the private sector in many (or most) economies [6] so governments in many parts of the world are under increasing pressure to boost their productivity and provide a high level of public services at a time of significant budgetary constraint.

In the same vein, many governments are faced with having to continue to provide a high level of public services at a time of significant budgetary constraint as they seek to reduce large budget deficits and national debt levels built up when they spent public money heavily to stimulate growth. Beyond the pressures of reducing debt levels, many countries face medium to long-term budgetary constraints caused by aging populations that will significantly increase demand for medical and social services [6].

[8] https://www.car2go.com/.

[9] https://de.mytaxi.com/index.html.

Thus, the question that arises is: 'Can Big Data help the public sector raise its game on productivity?' public sector could potentially reduce the costs of administrative activities and enabling governments to collect taxes more efficiently. Another Big data lever in the Public sector, is making relevant data more readily accessible. For example having public dashboards measuring the effectiveness of public programs and policies, lead to more informed and better decisions by both citizens and policy makers, while creates improved accountability in public sector agencies and improved public trust [1].

The question that arises is: '*Can Big Data help the public sector raise its game on productivity?*'. Big data levers, such as increasing transparency and applying advanced analytics, offer the public sector a powerful arsenal of strategies and techniques for boosting productivity and achieving higher levels of efficiency and effectiveness. For instance, public sector could potentially reduce the costs of administrative activities by using Big Data, including both efficiency gains and a reduction in the gap between actual and potential collection of tax revenue. These levers could accelerate annual productivity growth in the public sector in a significant manner [6].

5.5 Health Care

Health-related industries are specially boosted by Big Data to create new growth opportunities and entirely new categories of product and services. Particularly interesting for the sector will be the integration of nanotechnology embedded in people, that will be utilized as a monitoring and diagnostics tool [43] integrating and sharing different datasets forms of biological information, from high-resolution imaging such as X-rays, Computed Tomography (CT) scans, and Magnetic Resonance Imaging (MRIs) [44].

However health care has lagged behind other industries in improving operational performance and adopting technology-enabled process improvements. Different stakeholders, including (the pharmaceutical and medical products industries, providers, payors and patients) generates pools of data, very often unconnected from each other. There is a substantial opportunity to create value if these pools of data can be digitized, combined, and used effectively. The magnitude of the problem and potentially long timelines for implementing change make it imperative that decisive measures aimed at increasing productivity begin in the near term to ease escalating cost pressures [1].

It is possible to address these challenges by emulating and implementing best practices in health care. Doing so will often require the analysis of large datasets and Big Data. As stated in the McKinsey report *Big data: The next frontier for innovation, competition, and productivity*, the use of Big Data in the health industry

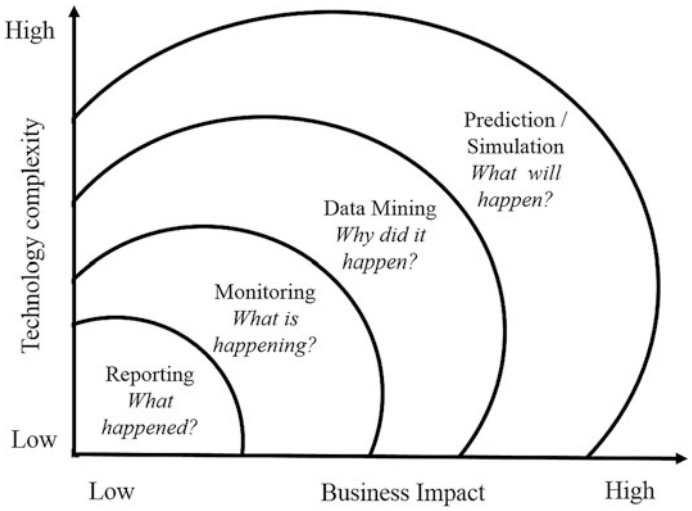


Fig. 6 Company range of Big Data capabilities. *Source* Adapted from The 'big data' revolution in healthcare, accelerating value and innovation. McKinsey (2014)

underlies a set of levers that have the potential to play a major role in more effective and cost-saving care initiatives, the emergence of better health services, and the creation of new business models in health care and its associated industries.

According to Groves et al. [45], one of the such main levers stands in the developing of high tech and high impact capabilities in health industry. Companies must make the journey from using data just to reporting, till use data for making predictions and doing business simulations, as shown in Fig. 6.

In health industry, value is derived from the balance of healthcare spend (cost) and patient impact (outcomes) [45]. Big Data has potential to allow patients to take an active role in their own treatment,—this can refer to anything as simple as visiting webmd.com and self-diagnosing, but increasingly this will take place as a one-on-one service with a qualified professional. This type of service is offered by Healthtap.[10] This role involves a personalized medicine tailoring medicines to a person's unique genetic makeup—and is developed by integrating a person's genetic blueprint and data on their lifestyle and environment, then comparing it alongside thousands of others to predict illness and determine the best treatment [46].

Big Data helps on the evaluation of drug efficacy based on real-world data. Huge amounts of data on applicants will allow researchers the identification of performance evaluation of integrated-care programs and contracts [45, 46].

[10]https://www.healthtap.com/.

6 Conclusions

As stated throughout this chapter, Big Data will profoundly disrupt businesses in nearly every industry. These businesses will tap into better connectivity and customer data to create major new products and services and, eventually, profitable markets that do not yet exist. As a result, competing successfully in the future will require a host of new and different management capabilities related to Big Data and its efficient use, not only for predictive strategies but also for effective decision-making.

First of all, organizations will need to understand how their existing products and services will add value in the new digital ecosystems and how they can create new sources of value. Furthermore, organizations and decision makers must understand the potential for new revenues and profits based on Big Data, as well as the second-order effects that can come from better cross-selling and up-selling of existing offerings. Market dealers must also assess the value at risk, by questioning the implications in term of customer churn or reduced market share as events unfold in this rapidly evolving space.

Although some companies will succeed with totally new digital and Big Data-based products and services, success will come most readily when companies can connect their offerings with what they are already good at doing and where they can add real value to their current or new customers. Such situation will create a need to decide whether an organization's current structure will support change at the pace required or whether a dedicated greenfield organizational unit is required. However, it is also worth it to acknowledge that no company will have the end-to-end capabilities internally to succeed in the new Big Data and digital world. Companies can save themselves an enormous amount of effort and considerable resources if they choose the right partners. Some partners excel at providing digital technologies and Big Data-based solutions, as well as data analysis, or customer service. Other might provide complementary data, products, or services, perhaps from another industry. Forward-thinking companies will need to intentionally select the criteria by which they will work with others, and ensure that those partners will still be delivering value not only today but in the near future.

All of these opportunities, challenges and industry sectors present key challenges for both decision-makers and organizations as a whole. As discussed throughout this chapter, some of them are related to volume, since the amount of data crossing the internet every second nowadays gives organizations an opportunity to work with many petabytes of data in a single data set—and not just from the internet. In the same vein, other opportunities and challenges are associated to velocity, since the speed of data creation may be even more important than the volume. Real-time or nearly real-time information will make it possible for a company to be much more agile than its competitors. Also, variety will be both an opportunity and challenge to tackle, since data will be taking the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Organizations will need to acknowledge that many of the most

important sources of Big Data are already relatively new, as well as the fact that the steadily declining costs of all the elements of computing—storage, memory, processing, bandwidth, and so on—mean that previously expensive data-intensive approaches are quickly becoming economical.

Standing still represents a high-risk option for any organization whose products and services are capable of being connected to the shifting world of Big Data. The business environment is changing and both organizations and decision makers must be prepared to make the most agile moves on the chessboard. Identifying relevant applications is, of course, just the first step in deriving value from Big Data. New capabilities, new organizational structures and mindsets, as well as significant internal change will also be required. But businesses should not underestimate the importance of zeroing in on the right opportunities. Organizations and decision makers will need to think outside the box, embrace new models, and even reimagine how and where they do business. A culture that encourages innovation and experimentation -and even some radical thinking- will serve in this vein, but so will calling in outside help when needed to assess, prioritize, and develop the different routes to value.

References

1. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity
2. Ebner K, Bühnen T, Urbach N (2014) Think big with Big Data: Identifying suitable Big Data strategies in corporate environments. In: Proceedings of 2014 47th Hawaii International Conference on System Sciences. IEEE, pp. 3748–3757
3. Holsapple C, Lee-Post A, Pakath R (2014) A unified foundation for business analytics. Decis Support Syst 64:130–141
4. Acito F, Khatri V (2014) Business analytics: Why now and what next? Bus Horiz 57(5):565–570
5. Dreischmeier R, Close K, Trichet P (2015) The digital imperative. https://www.bcgperspectives.com/content/articles/digital_economy_technology_strategy_digital_imperative/
6. Morabito V (2015) Big data and analytics. In: Strategic and organisational impacts
7. Morabito V (2015) Big data and analytics strategic and organizational impacts. Springer International Publishing, Milan, Italy
8. Saleh T, Brock J, Yousif N, Luers A (2013) The age of digital ecosystems: thriving in a world of big data. BCG Perspective
9. McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D (2012) Big data the management revolution. Harvard Bus Rev 90(10):61–67
10. Gerbert P, Gauger C, Steinhäuser S (2015) The double game of digital strategy. https://www.bcgperspectives.com/content/articles/business-unit-strategy-big-data-advanced-analytics-double-game-digital-strategy/
11. Shapiro C (1989) The theory of business strategy. Rand J Econ 20(1):125–137
12. Statchuk C, Iles M, Thomas F (2013) Big data and analytics. In: Proceedings of the 2013 conference of the center for advanced studies on collaborative research. IBM Corp, pp. 341–343
13. Kurzweil R (2005) Countdown to Singularity. http://www.singularity.com/charts/

14. Bughin J, Chui M, Manyika J (2015) An executive's guide to the Internet of Things. McKinsey Quart, McKinsey&Company
15. Driscoll EM (2010) Metamarkets blog. Metamarkets
16. Factor M (2016) Exabytes, elephants, objects and spark. https://www.ibm.com/blogs/research/2016/02/exabytes-elephants-objects-and-spark/
17. Wamba F, Akter S, Edwards A, Chopin G, Gnanzou D (2015) How big data' can make big impact: findings from a systematic review and a longitudinal case study. Int J Production Econ 2
18. Schumacher R, Lentz A (2008) Dispelling the Myths
19. SAS (2014) SAS® customer intelligence solutions. https://www.sas.com/content/dam/SAS/en_us/doc/overviewbrochure/sas-customer-intelligence-solutions-103116.pdf
20. Catlin T (2015) What it takes to build your digital quotient. In: Seitz B (ed.) McKinsey Quarterly. http://www.mckinsey.com/business-functions/organization/our-insights/what-it-takes-to-build-your-digital-quotient
21. Unit I (2015) Corporate. http://manpowergroup.com/wps/wcm/connect/6ef62c77-77c2-439c-b373-67bfb2732bd8/EIU_SAS_Big+Data+Evolution_PDF.pdf?MOD=AJPERES
22. Osterwalder A, Pigneur Y, Tucci CL (2005) Clarifying business models: origins, present, and future of the concept. Commun Assoc Inf Syst 16(1):1
23. Russom P (2016) TDWI big data maturity model and assessment tool. The Data Warehousing Institute (TDWI) maturity model
24. Council IDG (2007) The IBM data governance council maturity model. http://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf
25. Mahoney J, Gerrard M (2007) IT value performance tools link to business-it alignment. http://www.gartner.com/document/540520ref=solrAll&refval=166845039&qid=cf72dbfc73243dfce3b6ca01a2cbc647
26. Banks G (2014) More growth options up front: Big data enables a new opening step in the growth decision-making process. http://dupress.com/articles/more-growth-options-up-front/
27. Ansoff HI (1957) Strategies for diversification. Harvard Bus Rev 35(5):113–124
28. Pai-Ling Y (2015) Technology Strategy. H. B. S. P. Corporation, pp. 1–48
29. Brock J, Souza, R, Dreischmeier R, Platt J (2013) Big data's five routes to value: opportunity unlocked. BCG Perspectives, Accessed 21 Sept 2014
30. Schulte R, Zaidi E (2015) Market guide for operational intelligence platforms
31. Schlegel K (2015) Hype cycle for business intelligence and analytics. https://www.gartner.com/doc/3106118/hype-cycle-business-intelligence-analytics
32. Fischer G (2012) The right information at the right time, in the right place, in the right way to the right person
33. SAS (2009) Defining business analytics and its impact on organizational
34. Hagen C, Ciobo M, Wall D, Yadav A, Khan K, Miller J, Evans H (2013) Big data and the creative destruction of today's business models. Retrieved 5 Jan 2015
35. Din S (2016) < span lang = "EN-US" style = "font-size:10.0pt;mso-bidi-font-size: pt; font-family:"Times",serif;mso-fareast-font-family:"Times New Roman"; mso-ansi-language: EN-US;mso-fareast-language:DE;mso-bidi-language:AR-SA; mso-no-proof:yes" > Mobile location data management & advertising trends in 2016. http://go.placecast.net/blog/how-mobile-data-will-be-used-by-advertisers-in-2016
36. Bracht U, Masurat T (2005) The digital factory between vision and reality. Comput Ind 56 (4):325–333
37. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big data, analytics and the path from insights to value. MIT sloan ma Manage Rev 52(2):21
38. Van Der Lande J (2014) The future of big data analytics in the telecoms industry. www.analysismason.com
39. Acker O, Blockus A, Pötscher F (2013) Benefiting from big data: a new approach for the telecom industry. Strategy &, Analysis Report
40. Banerjee A (2013) Big data & advanced analytics in telecom: a multi-billion-dollar revenue opportunity. Technical Report, Heavy Reading, New York

41. Putre L (2015) Daimler, Audi, BMW, GM lead on autonomous vehicles: Study, http://www.industryweek.com/emerging-technologies/daimler-audi-bmw-gm-lead-autonomous-vehicles-study
42. Group WEF. a. B. C (2013) Connected world transforming travel, transportation and supply chains. http://www3.weforum.org/docs/WEF_MO_ConnectedWorld_Report_2013.pdf
43. Shah S (2013) Big data will go mainstream when nanotechnology is embedded into humans, says Skype CIO. http://www.computing.co.uk/
44. Jee K, Kim G-H (2013) Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inf Res 19(2):79–85
45. Groves P, Kayyali B, Knott D, Van Kuiken S (2013) The "big data" revolution in healthcare. http://www.images-et-reseaux.com/sites/default/files/medias/blog/2013/12/mckinsey_131204_-_the_big_data_revolution_in_healthcare.pdf
46. Marr B (2015) How Big data is changing healthcare. http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#25c446ca32d9

A Review on Big Data Security and Privacy in Healthcare Applications

Aqeel-ur-Rehman, Iqbal Uddin Khan and Sadiq ur Rehman

Abstract With the increasing use of technologically advanced equipment in medical, biomedical and healthcare fields the collection of patients' data from various hospitals is also getting necessary. The availability of data at the central location is suitable so that it can be used in need of any pharmaceutical feedback, equipment's reporting, analysis and results of any disease and many more. Collected data can also be used for manipulating or predicting and upcoming health crisis due to any disaster, virus or climatically changes. Collection of Data from various health related entities or from any patient raises some serious questions upon leakage, integrity, security and privacy of data. In this chapter the term Big Data and its usage in healthcare applications is discussed. The questions and issues are highlighted and discussed in the last section of the chapter to emphasize on the broad and pre-deployment issues. Available platforms and solutions are also mentioned and detailed to overcome the arising situation and question on usage and deployment of Big Data in healthcare related fields and applications. The available data privacy, data security, users' accessing mechanisms, authentication procedures and privileges are also described.

Keywords Big data · Security · Privacy · Healthcare · Data privacy · Authentication · Verification

1 Big Data

1.1 Introduction

Big data, from the name itself, represent the amount of data in bulk quantity. Big data in medical domain concerned with a useful data that is in big amount, fast in processing and contain high complexity level for process and interpret with the currently available tools.

Aqeel-ur-Rehman (✉) · I.U. Khan · Sadiq ur Rehman
HIET, FEST, Hamdard University, Karachi, Pakistan
e-mail: aqeel.rehman@hamdard.edu

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_4

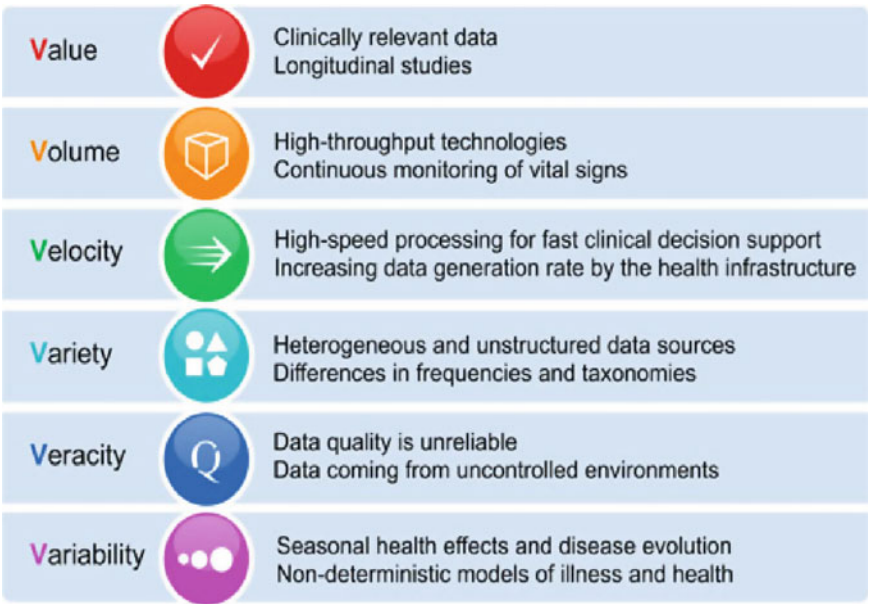


Fig. 1 Five V's of big data (value, volume, velocity, variety and veracity) [2]

At the beginning, the term Big Data was usually associated with a META Group report by Doug Laney entitled "3-D Data Management: Controlling Data Volume, Velocity, and Variety" published in 2001 [1]. After further research and developments, it is now suggested to identify big data problems by the so-called "5 V" (refer to Fig. 1): volume (quantity of data), variety (data from different categories), and velocity (fast generation of new data), veracity (quality of the data), and value (in the data) [2] which also apply to health data.

Recent studies [3–6] indicate that the Big Data technologies are commonly used for the improvement of quality and efficiency of data delivery in healthcare system. The impact of Big Data application can be observed when data is to be integrated from several healthcare areas, such as from clinical side, commercial or administrative side etc.

Big data contain huge amount of volume with the data type of unstructured, semi-structured or structured ones. Due to the massive volume, processing of data is very challenging by using conventional methods. The size of data for heavily populated countries like China and India was recorded to be in zetta-byte (10^{21}) and yotta-byte (10^{24}) but due to increase of multi-scale data along with the upcoming high data throughput sequencing, instantaneous imaging etc. the US healthcare system alone already reached 150 EB (10^{18}) five years ago [7].

Big data has now been the most popular search engine term with in few years. Huge amount of publications can be seen in Fig. 2 under the heading of "Big Data" nonetheless of disciplines, as well for those within the healthcare field [8].

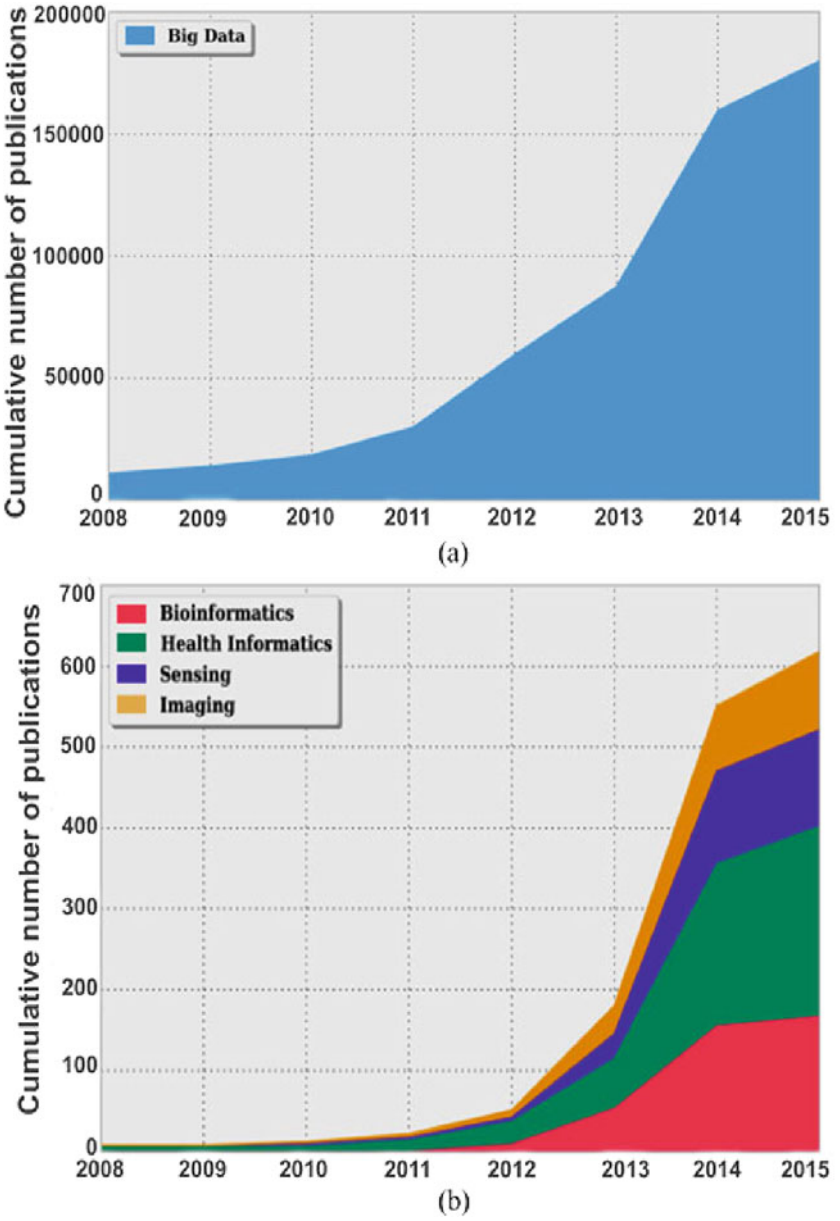


Fig. 2 **a** Accumulative number of publications mentioning to "big data" indexed by Google Scholar. **b** Accumulative number of publications per health research area referring to "big data," as indexed in IEEE Xplore, ACM Digital library, PubMed (National Library of Medicine, Bethesda, MD), Web of Science, and Scopus [8].

Nowadays big data has become the most important source of information that feed organizations and effect economies and international decisions [9]. Governments and organizations who are gaining and harnessing big data brought many advantages for their own such as [10]:

- Increasing operational efficiency and operating margins by taking advantage of detailed customer data [11]
- Improving performance by collecting more accurate and detailed performance data
- Targeting users based on their needs
- Improving decision making and minimizing risks
- Invention of new business area, services and products
- Improving threat detection capabilities of governments

1.2 Big Data Technologies

As we are looking forward to use Big Data applications (e.g. Electronic Healthcare Records) on large scale in the domain of healthcare, we must keep in mind the technical requirements for Big Data. Data digitization, data sharing, data quality, data security and privacy are the technologies that creates the technical foundation for subsequent Big Data applications and cover most of the health- specific data management technologies [12].

2 E-Health or Medical and Health Informatics

E-health is a theme for the improvement of healthcare quality by using technology. E-health facilitates both patients and medical professionals to enlarge their circle of knowledge by having access to different resources, this act makes healthcare cost effective and efficient. *Gunther Eysenbach* spelled out 10 E's of eHealth in an article publishes in Journal of Medical Internet Research which reflect the true meaning of E-healthcare concept, includes [13]:

1. Efficiency
2. Enhancing Quality
3. Evidence Based
4. Empowerment
5. Encouragement
6. Education
7. Enabling
8. Extending
9. Ethics
10. Equity

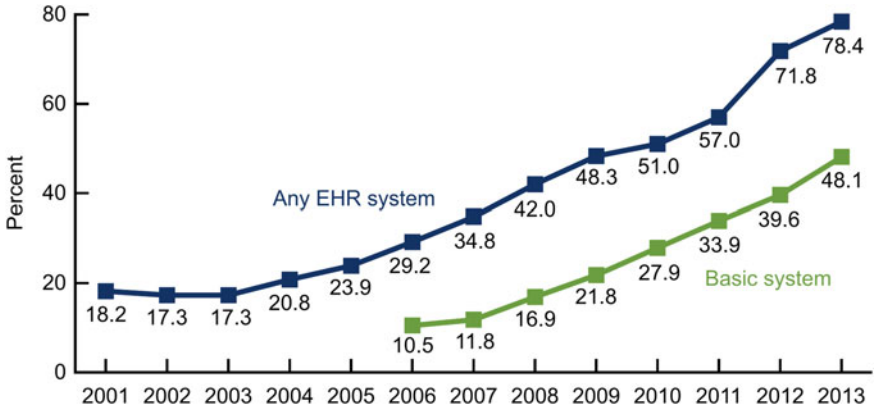


Fig. 3 Percentage of office-based physicians with EHR systems: United States, 2001–2013 [14]

The main concerns of health informatics is to focus on developing tools that can advance healthcare, on the other hand, E-health is responsible to spread this useful information. E-health and health informatics works together to give patients best quality of care in highly efficient way.

2.1 Electronic Health Records (EHRs)

We are living in a world of digital technology and the use of electronic equipment make our life easier. Internet and web-enables devices such as mobile phone, tablets, laptops etc. have dramatically changed our lives and the way of communication. Medicine is an enterprise filled with huge information, in digital healthcare where seamless flow of information is required, electronic health records (EHRs) are used. By using EHRs, patient's medical information is available all the time (refer to Fig. 3).

Electronic Health records are digital form of medical paper chart. EHRs are based on real-time technology. In EHRs, patient's information is securely accessible to authorized concerns. EHRs contain history (medical and treatment) of patients which is an important apparatus for scientific knowledge and clinical exploration, e.g., to discover physical information [15]. Local information mining including in EHRs data is useful for a large variety of healthcare related challenges, like supporting disease management [16, 17], pharmacovigilance [18], predicting model building for health risk assessment [19, 20], observing more about survival rates [21, 22], therapeutic recommendation [21, 23], discovering comorbidities, and developing support coordination for the enrollment of patients for new clinical trials [24].

2.2 Social Health

Compare to last decades, with the easy access and availability of internet and modern technologies, there is no such problem to get in touch within a friction of time with those who are at long distances. This also had a great impact in the field of medicine, especially on telemedicine which facilitate people from all over the world to seek help, communicate and get treated even though they are far away. Thanks to modern technologies that now the communication in telemedicine has been expended by adding the feature of social networks (social interaction). According to the latest research, one out of every fourth patients with enduring diseases, such as diabetes, cancer, and heart conditions, now utilizes social network for experience sharing with patients with similar symptoms and conditions. By these means they are providing one more potential big data's source [25].

3 Data Collection via Bio Informatics

The volume rate of data is rapidly increasing in all fields of research. There is now no restriction on the limit of big data resources to perform experiments on particle physics. Due to the availability of high throughput devices at cheap rates and with the help of digitization technology, data volume is dramatically increasing in bioinformatics research. For example, the size of a single sequenced human genome is approximately 200 GB [26]. With the advancement in new technologies, biologist are no-more using classical laboratory method to discover unique biomarker for disease. Instead of this, biologist used to rely on enormous and progressively expending genomic data which is provided by different researchers.

As the data size in bioinformatics is keep on growing, the European Bioinformatics Institute (EBI), which is among the largest biology-data repositories, had approximately 40 PB of data about genes, proteins, and small molecules in 2014, in comparison to 18 PB in 2013 [27]. This institute has hinxton data center cluster which contain 17,000 cores and 74 terabytes of RAM to process bioinformatics data. Furthermore, National Center for Biotechnology Information (NCBI), USA and National Institute of Genetics, Japan are the organizations who are also storing and processing huge collections of biological databases and distributing them around the world [28]. There are primarily five types of data that are massive in size and used heavily in bioinformatics research [28]: (i) gene expression data, (ii) DNA, RNA, and protein sequence data, (iii) protein-protein interaction (PPI) data, (iv) pathway data, and (v) gene ontology (GO).

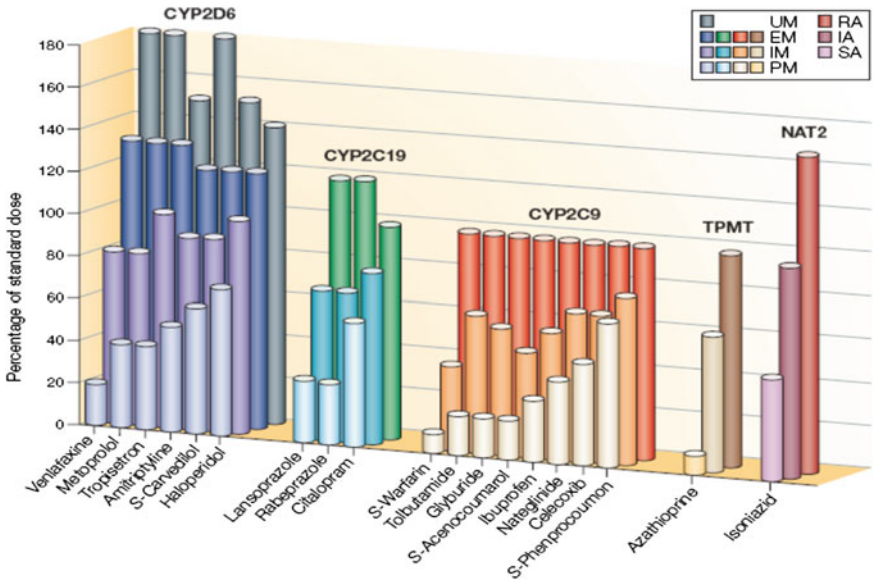


Fig. 4 Pharmacogenomics variants and dose [30]

3.1 Pharmacogenomics

Pharmacogenomics is a way to analyze and study the genes effect on a person's response to a drug. Pharmacogenomics is a combination of pharmacology which is related to drugs and genomics that is related to genes and their functions. Due to massive research in the field of Pharmacogenomics, safe medicine and dosage can be found in medical field (refer to Fig. 4). Pharmacogenomics are aiming to develop coherent means to optimize drug treatment, with respect to the patients' genotype, to make sure maximum effectiveness with minimal amount of contrary effects [29].

If we want to relate Pharmacogenomics with Big Data, typically about 3 GB of human genome can be obtained by using next-generation sequencing (NGS) which can vary up to 200 GB depending on average depth of coverage. It is estimated that about only 0.1 % of genome is unique amongst individuals. Normally, in data communication and transmission, data compressing is a common practice to save space of memory of resources, however, at present, compressed genotyping is not preferred to be compressed. One of its reasons could be, in order to examine complex medical diseases like cancer, studying complete genome sequencing by NGS is of high value.

4 Security and Privacy Issues in Big Data

4.1 Overview

In this segment we will discuss concerns of security and privacy in big data domain. The security and privacy concerns encompasses big data management and cyber security analytics. As it is understood that big data is now collated with many upcoming and advanced technologies dealing with data, so the database management is in the core of it. In this segment we will going to discuss that how the database and its management has changed and after that our focus will be on the Big Data's security and privacy issues.

Systems having databases has advanced very much during the past few decades, which not only includes systems as well as database management software and techniques. This also includes hierarchical models to relational and object database systems as well as network based legacy systems. Accessing databases via web and data management services are now been implemented and available as web services. As rapid growth of web based services emerged along with the amorphous type data management along with social media and mobile computing simultaneously, the requirement of handling collected data has now been increased from terabytes to petabytes and zeta bytes within few decades. The increment in the amount of complex data in now known as Big Data. The term big data is not only responsible to efficiently manage the huge amount of data but also it has to analyze it and extract meaning full or required tuples for business means and social engineering. The fore said data analytic is now known as Big Data Analytics.

The storage, analysis and management of such huge amount of data also lead towards volition of security and privacy. Sometime data is kept for a number of reasons including rules and monitoring as per agreements on webs. The data kept by the entities may violate the rules of privacy of a user. In addition, manipulating such a huge amount of data, including linking tuples/sets of diverse forms of data may results in violation security and privacy of a user. For case in point, although the collected raw data may removes individually detectable information, but the resulting data may enclose private and sensitive material of any individual. Likely, the collected raw data of any individual may be combined with his/her address, which may become adequate enough to classify the specific person.

There are multiple groups of people who are working on the challenges of Big Data. Multiple groups are working to develop platforms and standards for big data to manage massive networked and data storage, as well as solutions for effective and efficient management and analyses of huge amount of data groups. To contribute for Big Data, many research groups are working not only in industries but also in academic fields and in some countries groups are organized on government level also. On the other hand, very slight consideration has been given yet to the matters of security and privacy within Big Data. The recently done work on Big Data by various groups is acknowledgeable but the concerns of security and privacy

are not yet been considered by Network oriented or storage/system oriented research for big data.

Up till now research in data management and analytics for big data are proceeded in three directions:

1. To develop a standard and infrastructure with high performing computing techniques for mass storage.
2. To develop management technique for data, like integrating various data sources, indexing and querying techniques.
3. Efficient analytic techniques that can analyze big data and extract required tuples of data.

Moving towards security and privacy issues of Big Data, while gathering, storage, manipulation and withholding of enormous amounts of data sets, have given rise to some severe security and privacy concerns. Several procedures are now being proposed to grip the Big Data in such a manner that the privacy of any individual is not despoiled.

For understanding, if individually distinguishable material is filtered out from the collected data and then data is joined with other data sets for further manipulation, the removed information can be retrieved and individual can be identified.

It is also observed in several circumstances, regulations may perhaps grounds privacy to be despoiled. Just assume that the email data is gathered together and it has to be reserved for a definite period which is usually around five years. As long as any exchange server keep the data it could lead to the violation of privacy. A lot of regulations be able to suppress improvement. Consider a case that if by any means implementation takes place that raw collected data can be kept but it could not be manipulated or simulations cannot be constructed out of that data, at that point organizations will not be able analyze the data in state-of-the-art ways to boost their business so in this way the revolution may be barely audible.

Consequently, unique among the leading trials for warranting security and privacy as soon as allocating with the big data is to arise a well-adjusted method towards the procedures and analytics. This is the general way of an organization that can carry out beneficial analytics and still they be able to guarantee the privacy of any individual. Various procedures are adopted for privacy maintaining data mining, privacy-protective data mixing and privacy-protective statistics recovery. The core challenge is to outspread these techniques for control enormous volumes of over and over again networked data. One more security risk for Big Data management and analytics is to develop secure infrastructures. Several of the available tools that are now currently been developed comprising Hadoop, MapReduce, Hive, Cassandra, PigLatin, Mahout and Storm but the said before do not have suitable security mechanisms. Now the question arises that how these tools considered secured, simultaneously providing high performance computing?

Another issue is secure accessing mechanisms, indexing and query processing in data management. Again, a question arises is that how can procedures for diverse

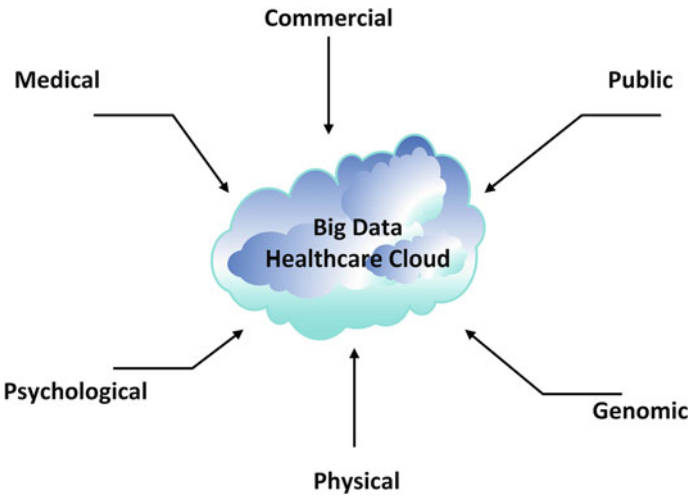


Fig. 5 Illustration of big data healthcare cloud [32]

forms of data like structured, semi structured, unstructured and graph data be combined?

Lastly, the complete zone of security, privacy, integrity, data quality and trust strategies must be scrutinized inside the framework of Big Data security.

Above, we have discussed just a few concerns and challenges in the light of security and privacy in Big Data domain. Still Big Data requires a comprehensive research and the results from academic and industrial units, to identify or pin point the challenges and develop the appropriate solutions for security and privacy. In future researchers from various disciplines needs to come forward and determine the solution. The disciplines may include policy management, high performance computing, and network science as well as data management [31].

Embracing the platforms available for big data in healthcare related applications, concerns of patient's privacy and security considerably increases. As from the beginning, statistics of any patient are stored in their data centers through variable intensities of applied security. Besides, many of the healthcare related data centers may have Health Insurance Portability and Accountability Act's (HIPAA) certification, but these type of certification does not give any assurance of patient's record security and integrity. The purpose being HIPAA, is additionally driven for safeguarding security guidelines and measures only rather employing them. Additionally, the entry of the huge data sets obtained from various cradles have an additional load on the storage section, processing units and as well as on communicating platforms. Figure 5 broadly illustrates a big data's healthcare related cloud that is able to hosts medical, commercial, public, genomic, physical and psychological data which belongs to the patients.

Conventionally security resolutions cannot be openly practically deployed to large and naturally dissimilar data collections. As there is growth in acceptance of

healthcare cloud based solutions, convolution in safeguarding enormous distributed, Software as a Service (SaaS) solutions rises along with variable data sources and their organizations. Henceforth, the governance of Big Data is compulsory, before exposing the collected data to analytics [32].

4.2 Data Control

Now the healthcare is emerged as an industry, moving on the way to an assessment's grounded business model, borrowing healthcare related analytics, data control must be the principal phase in modifying and handling healthcare related data. The main objective is to have a collective data demonstration that covers all industry related standards like Logical Observation Identifiers Names and Codes (LOINC), International Classification of Diseases (ICD), Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), Current Procedural Terminology (CPT) and other native and district standards. At present, data is generated by Body Sensing Networks is assorted in nature and would need to be normalized, standardized and controlled earlier to analysis.

4.3 Instantaneous Security Analytics

Examining the possible security threats and at the same time calculating the threat sources instantaneously is the ultimate necessity in the rapidly increasing healthcare industry. So at the current scenarios, the healthcare industry is observing a surge of classy assaults stretching from Distributed Denial of Service (DDoS) to the sneaky malwares. What's more, societal engineered attacks are rising and the hazards linked with such type of attacks are problematic to predict without bearing in mind the human's rational actions. Cognitive bias, be able to originate into action, mainly in the advanced years patients' cases. Cognitive bias is an arrangement of abnormality in conclusion, whereby impacts roughly other people and state of affairs may be pinched in an inconsistent manner. Taking an example, it is quite simple that man in the middle attack can be produced by persuading an old and aged patient to receive a digital X.509 certificate. The said type of situations necessarily to be taken of explanation while scheming some concrete end to end authenticating mechanisms or solution. As Internet of Things (IoT) platform is emerging, in its environment, employing security in less resourced networks has been always a challenging task and it will carry on to propagate more composite as the number of IoT enabled devices increases. For reference let us take orthodox symmetric and asymmetric key sharing and withdrawal structures cannot be long-drawn-out to a billion IoT enabled devices. Henceforth, there is requirement of an innovative scalable key management schemes, most important to unified inter-operability among unrelated networks, for example, IoT and other traditional IP based networks are critical for combination of IoT and big data within any cloud oriented environment.

Now the healthcare related industries are influencing on evolving big data technologies to create updated resolutions and also security analytics are the indispensable component of any recent design depending on the Software as a Service (SaaS) oriented solutions for accommodating Protected Health Information (PHI). In addition, the implementation of real time security and intelligence within any system will construct new standards in risk management. As a result, IT based healthcare benefactors may observe threats within any system instantaneously and they will be able to take defensive procedures before the healthcare business goes under distress.

4.4 Privacy-Maintaining Analytics

Incursion of patient's privacy is the major apprehension on the rise under the domain of big data analytics. Forbes magazine published an occurrence of privacy compromised, which gives rise to an alarming situation to be considered for patient privacy. In that report, it is mentioned that a well-known corporation referred some baby care product's coupons to a teenaged girl and her parents were unaware of it, resultantly creating panic with in family. This has increases the concerns and urges big data researchers to think and canvas the strategies to deal with privacy analytics on prior bases. As a case in point, data must become anonymous, earlier to make it available analytics, may possibly safeguard patient's identification. Additionally, privacy-maintaining encryption patterns that allows successively forecast algorithms on any encrypted data even though guarding the characteristics of any patient is necessary for conducting the healthcare related analytics. With the passage of time industry will influence on IoT enabled devices to send health related vital statistics to dedicated healthcare related clouds, so some prerequisites are implemented for dealing out and analyzing the collected data in any functional ad hoc distributed method. On the other hand, carrying out resource-fatiguing computing operations, mostly for analytics, even though preserving privacy of data is a gigantic challenge while having limited resources' environment. In addition, as healthcare oriented analytics will increases acceptance, innovative privacy protecting laws will be the requirement to safeguard privacy of any patient. Consider an example, from many patients "informed consent" is mandatory, before undergoing any type of analytics on data, keeping this considered, first-hand laws will be required to be enlisted, in which all the involved process of undergoing big data analytics on any patient's data must be explicitly mentioned and listed.

4.5 Data Leakage

In digital world, data leakage is one of the most important challenges under security and privacy in line to make available the enormous volumes of data, the ability to

tie unrelated data generating sources, the upsurge in the collected data's distribution, and the nonappearance of guidelines and techniques that reveal growing technological know-hows.

4.5.1 Current Keys to Counter Data Leakage

Oddly, the communal technique to defend against possible data crack and its leakage and confirm acquiescence with applied security, privacy procedures and techniques. The National Security Agency also count on the oral pledges to defend in contradiction of deliberate data leakages, but in some cases only oral pledges considered useless too if the resulting incentive to leak data is higher than the encouragement to protect it.

Organized Access Privilege of entry through authorizations till now remains the most communal technological methods to have protection against illegal right to use collected data. Passwords have been used during the course of history to validate individual's identity. Considering digital ecosphere, strings contains passwords or passcodes are made up of typographical letterings, comes under verification process to accept the access toward a specific computing environment or some various kinds of digital equipment and platforms. Although passcodes may able increase data's security, but still they have limits. The limits consist of a fact that passcodes can be certainly conveyed from any person to another without having consent of the possessor of the data. Most of the people are unaware that how important passcodes and passwords are, so they are not good their password management and carry on to depend on the passwords which they can easily remember, they type of passwords are able to hacked, like rotation of some fixed passwords, involving some simple additions or changes to an fix string of password, also includes the use of surnames and family names or birthdates or even place names. Old-styled methodologies to reset a forgotten password is the use of temporary allotted passwords which may bring together vulnerabilities and threads. Without a doubt, in 2012 there was a case where client's service staff members at major companies mistakably aided in the hacking of one of famous individual's digital files via password security vulnerabilities like asking for an answers to reset password aka a security question, that an stranger may possibly easily find out and the old one too, it was a solution to reset password by sending a temporary password to email. New safe and sound solutions to reset passwords like blocked access, requirement of postal address to reset passwords are considered to be weighty on the customer end and as a result very unlikely to be embraced [33].

Multi-Factor Verification At least two-factor or multifactor verification symbolizes an enhancement over the use of simple password too. Although it is not only limited to use within the digital domain, two-factor authentication usually requires any user to submit any two of three authentication factors in advance for gaining access to collected data or any other computing resource. Generally these factors are: Passwords, any smart or banking card and Fingerprints [34]. The most commonly use of said example is security methodology used at the ATM machine,

which not only requires a bank's ATM card but also a unique Personal Identification Number (PIN), usually of 4 digits.

4.5.2 Data Leakage Prevention

DLP Technology is comparatively an innovative methodology to secure data and it was announced in the middle of 2000s, it was also aimed to counter the outflow of delicate or personal data, mainly by insiders with malicious intent. DLP technology consist of the a process in which data packets are assessed locality and groups are made to share file and data's movement control within department from an its internal network, over and done with the execution of procedures that are based on data locality and file organization.

4.6 Data Privacy

The word "privacy" involves not only evading the observation, or else keeping any individual's problems and associations stealthy, on the other hand also have the capability to share required info very selectively but not openly. Unrecognizability correspondences with secrecy, but the two are not alike. Voting is acknowledged as a private, but not unidentified, however composition of a politically aware tract can be anonymous, but it is not private. Similarly, the capability to sort familiar individual judgments without government interfering is considered to be a privacy right, as is guarded from perception on the base of convinced personal individualities like as an individual's race, its gender, or other medial and physical characteristics. Consequently, now we can say that privacy is not only about keeping secrets.

The assurance of big-data gathering and exploration is that the consequential data can be used for resolutions that subsidy both persons and humanity. Fears to confidentiality stalk from the measured or careless revelation of composed or resultant individual's data, the mistreatment of the data, and the fact that resultant data might be imprecise or fabricated. The battle between privacy and innovative technology is not new-fangled, apart from it possibly now in its larger scope, amount of understanding, and ubiquity. For more than two centuries, morals and prospect linking to privacy have been repetitively reinterpreted and reshaped in the light of the impression of new technologies in our society [35].

4.7 Data Sharing

By way of the swift progress of information digitization, enormous volumes of structured, semi-structured, and unstructured data are produced rapidly. By means of gathering, organizing, analyzing, and excavating these collection of data, any

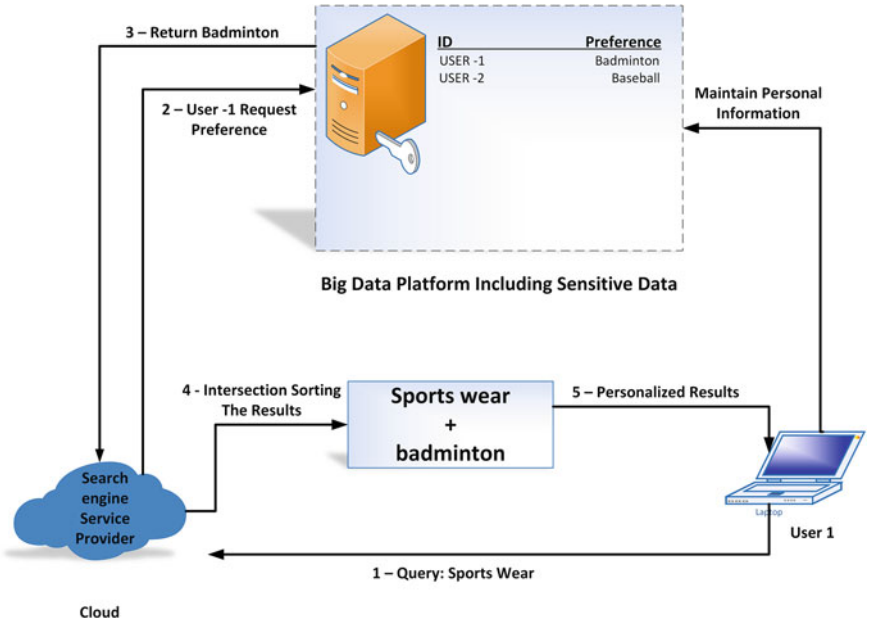


Fig. 6 Application of sensitive data (user's preferences) [36]

enterprise be able to acquire large volumes of discrete users' complex data. These data not only bump into the requirements of the enterprises itself, but also be responsible for services to other businesses if the collected data are stored through the platform of Big Data. Conventional approach of any cloud storage is that they stores plain text or encrypted data inactively. The data stored normally is well-thought-out as "dead", because there is no involvement of any calculation. On the other hand, the big data platform permits the interchange of data which may include any private and sensitive data. The aforementioned provides bulk data storage along with computing facilities. Computing facilities mainly referred as the calculating and analytical processes including encryption of data, transformation, or functional encryption on data to be used by various contributors and members, which be able to stimulate the "dead" data. A specimen of such type of an application is illustrated in Fig. 6 to demonstrate the flow process of sensitive data. As you can observe in Fig. 6, User's favorites or preferences are considered as sensitive data. When User-1 submits a query of a sportswear, the Search Engine Service Provider first search for the user's preference available within the big data platform, and if the information in big data platform has previously collected or have it via shared mechanism, "badminton", then the search engine returns personalized results like illustrated "sportswear + badminton" to user. As the user observes his/her favorite badminton sportswear, it will experiences a satisfying and pleasant buying. On the other hand, even though data sharing rises enterprise's resources, the threats like Internet insecurity and the leakage of sensitive data also craft the possible security issues for sharing of sensitive data [36].

5 Open Questions

5.1 Who Will Own the Collected Data?

One of the primary questions that could be asked is who has the power to erase, oversee, and add statistics to health's data along with implementation of regulations adjoining it?-Dose the data is individually owned by every person? Or will their physicians own their data? Does the insurance company will own the collected data? Or all will be the all combined owners? Thus the question 'Who will owns the collected data?' is practically worrying and also uncomfortable. In answer to aforementioned question that, if data is collected and stored via third party, than do that entity have the equivalent access right and power as same as the owner of the data, or their rights will be much constricted? It is still undecided that till which level the privacy and security layer necessarily be kept when the data is transported to any third party.

5.2 Which Type of Data to Be Collected and What Will Be the Amount, of Data to Be Stored?

Notes, Magnetic resonance imaging (MRIs), and the results by test labs are models of data that may be required to be stored in any patient's physical record at any hospital. Concerning Electronic Patient Records (EPRs), should all before mentioned models of data to be electronically stored or the subset of this data will be adequate enough for healthcare dedications? The asked question may also relates to the cases of patients who are remotely monitored via sensing network. In afore mentioned cases the amount of data to be collected and needs to be stored should be minimized to a certain extend that it should be able to meet the level of healthcare requirement.

5.3 What Will Be the Storage Location?

Now this question will lead us to centralize verses decentralized storage competition. Considering the example of EPRs, would data reside in the local databases which are linked to each other, or it will be better to be stored in a centralized database? Considering the case of patient monitoring via sensing network, will it be better that raw data from sensors be stored only on local database or it should be stored at any central monitoring location? What type of both data storages will be a better choice to accommodate the privacy and security requirements?

5.4 To Whom Patient's Medical Record Should Be Visible?

There are two broad categories of EPR users, one is the user with the maximum privileges i.e. read and write, as an example the doctors and nurses, who can view a patient's EPR, but can also manage the records. Second is the user who have read only privileges, contingent on which user is get into the EPR, there might requirement of further limitations that on which portion of the data their privileges apply to.

5.5 Is Disclosure of Information Without Patient Permission Allowed?

In many situations in which it is necessary to disclose the health information of any specific patient in front of people other than who were previously authorized to see. As an example, considering the case of remote patient monitoring, in any emergency situation it might be necessary to disclose the health record of the patient's without his/her consent to receive necessary care [37].

6 Conclusions

In this chapter, we have disused the security and privacy issues emerging with the fast growing need and deployment of Big Data for Healthcare needs. We have also discussed some recent progress in the said field, which includes Data storage, security, sharing, Patient records, healthcare related databases and real time data protection.

References

1. Laney D (2001) 3D data management: controlling data volume, velocity, and variety, META Group
2. Terzo O, Ruiu P, Bucci E, Xhafa F (2013) Data as a service (DaaS) for sharing and processing of large data collections in the cloud. In: Proceedings international conference complex intelligent software intensive system, pp 475–480
3. Frost and Sullivan, U.S. (2012) Hospital Health Data Analytics Market
4. McKinsey and Company. Big data (2011) The next frontier for innovation, competition, and productivity
5. Groves P, Kayyali B, Knott D, Van Kuiken S (2013) The 'big data' revolution in healthcare. McKinsey & Company
6. Porter M, OlmsteadTeisberg E (2006) Redefining health care: creating value-based competition on results. Harvard Business ReviewPress, Boston

7. Cottle M, Hoover W, Kanwal S, Kohn M, Strome T, Treister NW (2013) Transforming health care through big data. Institute for Health Technology Transformation, Washington DC, USA
8. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang G-Z (2015) Big data for health. IEEE J Biomed Health Inf, 19(4)
9. AlMutairi AM, AlBukhary RAT, Kar J (2015) Security and privacy of big data in various applications. Int J Big Data Secur Intell 2(1):19–24
10. Tankard C (2012) Big data security. Netw Secur 2012, 5–8. Devakunchari R, Handling big data with Hadoop toolkit. In: 2014 International conference on information communication and embedded systems (ICICES), pp 1–5
11. Moon, H, Chou HS, Jeong SH, Park J (2014) Policy design based on risk at big data era: case study of privacy invasion in South Korea. In IEEE International congress on big data (BigData Congress), pp 756–759
12. Zillner S, Oberkampf H, Bretschneider C, Zaveri A, Neururer S (2014) Towards a technology roadmap for big data applications in the healthcare domain. In: 2014 IEEE 15th international conference information reuse and integration (IRI), pp 291–296
13. Eysenbach G (2001) What is E-Health. J Med Internet Res, 3
14. Hsiao C-J, Ph D, Esther Hing MPH (2014) Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2013. NCHS Data Brief, No. 143, January 2014
15. Friedman C, Shagina L, Lussier Y, Hripcsak G (2004) Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 11:392–402
16. Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, Kirby J, Lasko T, Saip A, Malin BA (2014) Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc 21:337–344
17. Forrest GN, Van Schooneveld TC, Kullar R, Schulz LT, Duong P, Postelnick M (2014) Use of electronic health records and clinical decision support systems for antimicrobial stewardship. Clin Infect Dis 59:122–133
18. Eriksson R, Werge T, Jensen LJ, Brunak S (2014) Dose-specific adverse drug reaction identification in electronic patient records: Temporal data mining in an inpatient psychiatric population. Drug Saf 37:237–247
19. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: Using analytics to identify and manage highrisk and high-cost patients. Health Aff 33:1123–1131
20. Boland MR, Hripcsak G, Albers DJ, Wei Y, Wilcox AB, Wei J, Li J, Lin S, Breene M, Myers R (2013) Discovering medical conditions associated with periodontitis using linked electronic health records. J Clin Periodontol 40:474–482
21. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, Levy M, Shah A, Han X, Ruan X (2014) Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc, pp 1–10
22. Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N (2014) Survival analysis with electronic health record data: Experiments with chronic kidney disease. Stat Anal Data Min ASA Data Sci J 7:385–403
23. Cars T, Wettermark B, Malmstr¨om RE, Ekeving G, Vikstr¨om B, Bergman U, Neovius M, Ringertz B, Gustafsson LL (2013) Extraction of electronic health record data in a hospital setting: Comparison of automatic and semi-automatic methods using anti TNF therapy as model. Basic Clin Pharmacol Toxicol 112:392–400
24. Marcos M, Maldonado JA, Mart´ınez-Salvador B, Bosc´a D, Robles M (2013) Interoperability of clinical decision-support systems and electronic health records using archetypes: A case study in clinical trial eligibility. J Biomed Inform 46:676–689
25. Andreu-Perez J, Leff D, IP HMD, Yang G-Z (2015) From wearable sensors to smart implants towards pervasive and personalised healthcare. IEEE Trans Biomed Eng, pp 1–13 (submitted for publication)
26. Robison RJ (2014) How big is the human genome?. Precis Med

27. EMBL-European Bioinformatics Institute (2014) EMBL-EBI annual scientific report 2013
28. Kashyap H, Ahmed HA, Hoque N, Roy S, Bhattacharyya DK (2014) Big data analytics in bioinformatics: a machine learning perspective. J Latex Class Files, 13(9)
29. Aneesh TP, Sekhar SM, Jose A, Chandran L, Zachariah SM (2009) Pharmacogenomics: the right drug to the right person. J Clin Med Res, pp 191–194
30. Papaluca M (2013), Introduction to Pharmacogenomics, ENCePP Plenary meeting, 12 November 2013. http://www.encepp.eu/publications/documents/6.1_Pharmacogenomics.pdf
31. Thuraisingham B, Bertino E, Kantarcioglu M (2014) Workshop report big data security and privacy, NSF Sponsored
32. Patil HK, Seshadri R, (2014) Big data security and privacy issues in healthcare. IEEE Int Congr Big Data, pp 762–765
33. Schmitt C, Shoffer M, Owen P, Wang X, Lamm B, Mostafa J, Barker M, Krishnamurthy A, Wilhelmsen K, Ahalt S, Fecho K (2013) Security and privacy in the era of big data: the SMW, a technological solution to the challenge of data leakage. RENCI, University of North Carolina at Chapel Hill. Text. www.renci.org/wp-content/uploads/2014/02/0213WhitePaper-SMW.pdf
34. Federal Fiancial Institutions Examination Council. Authentication in an internet banking environment. http://federalreserve.gov/boarddocs/srletters/2005/SR0519a1.pdf
35. Big Data and Privacy: A Technological Perspective, May 2014
36. Dong X, Li R, He H (2015) Secure sensitive data sharing on a big data platform, Tsinghua Sci Technol I S S Nll1007-0214ll08/11 ll, 20(1), 72–80
37. Meingast M, Roosta T, Sastry S (2006) Security and privacy issues with health care information technology. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE

What Is Big Data

Eizo Kinoshita and Takafumi Mizuno

Abstract This chapter consists of three parts. In first section, we describe what Big Data are. We explain concepts of Big Data, and how it arose. Big Data affect scientific schemes. We mention limitations of predictions by use of Big Data, and a relation between Big Data and hypotheses. And we note that electric power of Big Data systems. In next section, we describe necessity of Big Data. This is a view from aspects of macroeconomics. In service science capitalism, measurements of values of products need Big Data. Service products are classified into stock, flow, and rate-of-flow-change. Immediacy of Big Data implements and makes sense of each classification. And we provide a macroeconomic model with behavioral principles of economic agents. The principles have mathematical representation with high affinity of correlation deduced from Big Data. In last section, we provide an explanation of macroeconomic phenomena in Japan since 1980 as an example of use of the model.

Keywords Big Data · Macroeconomics · Thetical and Antithetical economics

1 What Is Big Data

People have found a vague uneasiness in every field. This is that we will not be able to treat floods of data made by information communication technologies. Big Data is a concept of data which is "impossible to treat by use of traditional technologies", and is "provided by information communication technologies." There are some definitions of Big Data, we adapt 3 V definition in this chapter; Big Data are date which are

E. Kinoshita (✉) · T. Mizuno
Meijo University, Kani, Gifu, Japan
e-mail: kinoshit@meijo-u.ac.jp

T. Mizuno
e-mail: tmizuno@meijo-u.ac.jp

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_5

1. large in quantity (Volume),
2. traded high speed (Velocity),
3. formed in many style (Variety).

Big Data is impossible data for traditional technologies, but it would be resources if we can overcome impossibility. A term "data" has two different meanings

- information which are obtained experiment or survey for problem solving or decision making,
- information stored in computers.

Big Data have these two meanings simultaneously.

Computer scientists have predicted arising of Big Data earlier than 2001. In those days, broadband wires were spreading ordinary homes, number of portable phones was increasing, and B2C commerce had just realized. Technological innovations miniaturized computer terminals, reduced costs of communications, accelerated communication speed, and released computer networks to open societies. Computer networks became our common property. People can use Web, search information, and send their message to the world. Corporations find values in searching information. Google added a value which is PageRank to the searching information, and Google monopolize portals on the Web. Around 2005, people want environment which they can access to information whenever or wherever they are. This demand is realized by cloud infrastructures. Corporations provide services on the cloud infrastructures at free price. And now, people use Web searching, mail, social networks, and e-commerce. It has become quite common. But people perceive a danger that information become bigger as they can treat. Around 2010, a term Big Data arose to societies.

Big Data represent projections of things on real world, thinking of people, results of calculations of computer. To say concretely, they are numerical values, texts, images, movies, sounds, programs, and so on. They are coded by any rules, and stored in storages of computers. Computers can proceed any procedures to the coded things. The procedures are operations on binary digits essentially. So, infrastructures of Big Data are computer systems, and theoretical backgrounds of Big Data is computer sciences and information communication technologies.

Computer is a system consists of four parts: arithmetic units, memories, I/O interfaces, and storages. Data are stored in storages. When a computer processes some procedures, computers load data and programs from storages to its memories. Arithmetic unites process data in the memories as calculation on binary digits. Results of the processing are displayed to operators via I/O interfaces, or stored into storages.

Storage accesses occupy almost all times of processing with large volume data. Access times between arithmetic units and memories are nanoseconds order. While access times between memories and storages are micro seconds or milliseconds order. Or the speed of data transfer between memories and storages is about gigabits per seconds. Now, the speed of data transfer on Web is about gigabits per seconds or hundreds gigabits per seconds. It means that storage access times and network access

times are same order. We do not mention which data are on storages or beyond networks. Network speeds give us permeability of data accesses on networks and realize using Big Data.

Infrastructures of Big Data consist of many computer nodes. When we process Big Data by use of more than ten thousand nodes, we need specific software technologies and hardware technologies. For examples, Google constructs Big Data systems which consist of about million nodes. And Google also provides Big Data infrastructures: Google File System, MapReduce, BigTable.

1.1 Predictions by Use of Big Data

Theories used to be our experiences which are arranged economically. Because costs of memorizing are higher than costs thinking for humans, we generalize our experiences and avoid memorizing all experiences. Searching of information from vast experiences has high costs for humans, too. So, memorizing has high status in every examination for humans. Until about twenty years ago, there are experts who treat data and sources in every corporation, administrations, and universities. But now, we will change drastically the situations.

- Costs of memory close to zero for person.
- We can test a hypothesis with large data.
- We can use all data at statistics operations.
- We can downloads data which used to be offered by experts.
- We can ask any questions, and we can obtain their answers immediately.
- We can memorize our ideas anytime and anywhere, and we can edit the ideas anytime and anywhere.

A position of the memory in our lives falls extraordinary. Thinking used to supplement to memory, but purposes of the thinking will change. A role of thinking will be that Big Data systems cannot do.

Big Data are large volume data and exist with systems which treat the large volume data. Generally, increasing data volume improves the precision of predictions in scientific field and social sciences. When we introduce Big Data to people, we often tell successful stories of Big Data, too. A prediction flu from social messages on Web by Google, a recommendation system of Amazon, a retail link system of Walmart, and so on. They are successful case studies of improvement of predictions by use of Big Data. There are some factors which lead the cases their successes.

- Attributes or variables which are used in the predictions are concentrated by experts in advance.
- Users do not control their environments by the predictions.
- Environments do not drastically change.

In other words, the cases are predicted well by correlations.

Increasing variables increases the number of data which are required for precise prediction. The increase is proportion to exponential scale of the number of the variables. Although Big Data has enormous data, the volume of data are less than required data by the increasing variables. It is often referred to as curse of dimension.

Generally, predictions by use of Big Data are based on correlations. It means that the predictions does not say causality. The prediction does not tell us input-output relations; variables of the predictions are not causes, and output of the predictions are not results. So, we cannot control outputs of a system by adjusting inputs of the system with the predictions.

Big Data fill environments with enormous data. Predictions by use of Big Data suitable for the environments, and Big Data explain the environments well. But, in areas beyond the environments, the predictions are not suitable. We must common premise between the environments and the areas when we use the predictions. It is difficult to abstract the premise from correlations. The difficulty occurs when the environments drastically change.

1.2 Big Data and Hypotheses

Big Data are information which are obtained from systems. The information corresponds results of experiments or observations in scientific fields. A system consists of minute parts, and each parts have relations mutually. If we analyze each parts, we cannot understand the system because of mutual actions. Big Data are obtained from such systems. Because systems are complex, we must collect large volume data. Urban, human, weather, traffic, software, financial trade, and other economic actions are systems. When we combine parts, the complexity of a system appears as new characteristics of the system.

We cannot understand a system by just viewing the system. So we ignore or simplify complex parts of the system for understanding or study of the system. The simplified representations of the system is referred to as model. The model is the abstracted of the system. The model, of course, is differ from the system. But we can understand the system via the model. The model represented by artificial languages or mathematical symbols is referred to as mathematical model. Mathematical model can represent actions of the system by algebraic operations.

Let us consider a mathematical model with any parameters. We can specialize the mathematical model by fill the parameters with any values. The specialized mathematical model is referred to as hypothesis. In statistical fields, targets of tests are the hypothesis.

$$\text{model} \quad + \quad \text{specific parameters} \quad = \quad \text{hypothesis} \tag{1}$$

We can test hypotheses by concrete procedures. A hypothesis which tolerate against various tests are called theory. We can use Big Data for constructing hypotheses, and

for testing the hypotheses. In other words, Big Data can make hypotheses. And the hypotheses are representations of system that we want to understand.

We can understand a system by combining data and model, not only seeing data. Arising Big Data and Big Data systems reduces costs of data, and reduces values of data itself. We treat Big Data as assets. It means, to be exact, our assets are Big Data with model, and hypotheses made by Big Data.

1.3 Big Data and Electric Power

On physical aspects, we can grasp Big Data by electric power. Big Data systems consist of thousands computer nodes or million computer nodes. Required electric power is million times as large as one personal computer. Heat produced by the systems are also large. Big Data system is often in data centers. Computer nodes arranged into the data centers. Each data center has huge buildings to cool down computer nodes. Electric power which one computer nodes consumes is hundred watt order. If there are ten thousands computer nodes, required electric power is megawatt order.

If we see services provided by Big Data systems as merchandises, raw materials of the merchandises are electric power. Big Data systems have unique cost structure; we can raise profit rates by reducing electric power for produce one merchandises. Electric power is most important for Big Data systems, because consuming electric power by the system is very large, and because using electric power effectively rises profits directly.

2 Why We Need Big Data

We need appropriate use of Big Data to manage our society. If we solve an urban complex problem, the solution makes new issues. Strategic solutions for the issues need large data and data which are obtained without delay. Urbanization and appearance of Big Data change our approaches for the issues. Before appearance of Big Data, we search data for causal relationships. But after Big Data, because of its huge size, we can acquire sufficient correlation to solve the issues. It means that correlation substitutes for causation (Fig. 1).

Information communication technologies reduce sectionalism of governments and give us solutions for urban problems by collaboration of mutual sections. Impor-

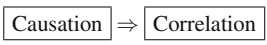


Fig. 1 Relationship in our societies dominated by Big Data

tance of leader who plays role of control tower of each section increases in such mutual society. The leader is CEO or president in corporations, and is prime minister in government. Decision makings design our society. Grand design of society consists of designs of each sector of the society. Each design must be verified that the design is in accordance with the grand design. We need Big Data to construct a strategy by integrating designs and to verify accordance between the strategy and the designs. The society is dominated by economics. Big Data from economic activities are important us for our decision makings or for the construction of the strategy. In this section, we describe necessity of Big Data from aspects of macroeconomics.

In 1980s, researchers of macroeconomics recognized difference between goods products and service products, and they have tried to define what service products are. Now, service products are defined as products that have properties: intangibility, immediacy, variability, perishability, and customer's high satisfaction.

A major premise of macroeconomics is that our world is capitalism. If the world is not capitalism, then every theory of macroeconomics will lost its senses. So, researchers of macroeconomics, managers of companies, or government administrators must consider whether we are in the world with capitalism.

The most important concept of capitalism is fixed price sales. Fixed price sales enable us to run our planned business and guarantee value of capitals.

To enforce fixed price sales without any contradictions on our business, we must measure values of our products precisely. In a word, precise measurements of products provide bases of every index about economics and managements in the world of capitalism; the measurement of values of products is an element forming economics and managements.

For any goods products, we can measure its values relatively easily. Because the goods have physical entities and properties, we can reduce eventually their values to their length, weight, temperature, velocity, or entropy.

On the other hand, we cannot measure values of service products easily. Service products often stand on relations between goods and goods, or between services and services. Relationship is combinations of products, and increasing the number of the combinations makes measurements of values of the products complex. As service products consist of some lower level services, they are developed in high abstraction level far from physical goods products. To overcome the complexity and the distance abstraction level, we need much knowledge of many fields.

In early 2000s, IBM researchers advocated a necessity of "service science" which is a new research filed to construct knowledge systems for service products. We need accumulation of knowledge. It means that we must collect Big Data and extract new theories form Big Data.

We refer to a society in which almost all employees work for service industry as service science capitalism society. In the society, every price value has large amount of information in the background of the value, and the value is detected in high abstraction level far from its physical entity. To fill the gap between abstraction levels, we must learn techniques which reduce from Big Data to a value through experience.

Big Data provide us new measurements for service products, and enable us to classify service products into three services: stock service, flow service, and rate-of-

flow-change service. Stock service is construction of social infrastructures or information infrastructures. Flow service is ordinary everyday service which provided by government administrators and private companies. Rate-of-flow-change service is unusual service.

There is an analogy between physics and economics. In physics, a phenomenon is described in distance, velocity, and acceleration. Establishing the three concepts makes modern physics since 17th century. While economics was made by establishing three concepts: stock, income, and growth rate. In economics, a product is described in the three concepts. Distance, velocity, and acceleration in physics correspond to stock, income, and growth rate in economics, respectively. Distance and stock are measured by some accumulations. Velocity and income are represented in time differentiations. Acceleration and growth rate are represented in twice differentiations. The classification of service products corresponds to the concepts of physics and economics.

The classification presumes that we can trace changes of values of service products every times. It corresponds to time derivative in physics. Immediacy of Big Data provides us feasibility the classification.

When we use Big Data sufficiently, correlation plays important roles in any analyses of economics. So we must build macroeconomic models which we can construct by detecting parameters from correlation deduced from Big Data.

Kinoshita provides a macroeconomic model which is referred to as "Thetical economics and Antithetical economics" [1–4]. That is a rearrangement of theories of macroeconomics into two set; a set of them is Thetical economics and another set is Antithetical economics. If Say's law is valid in an economic phase in an economic cycle, then the Thetical economics dominates the phase. We feel that we are in normal economy and economic growth in the phase. While if the Keynes's effective demand is effective in an economic phase, then the Antithetical economics dominates the phase. We feel that we are in depressed economy in the phase. Economic phases dominated by Thetical economy and economic phases dominated by Antithetical economy are illustrated in Fig. 2. Easy to say, Thetical economics represents what prosperity is, while Antithetical economics represents what recession is.

With the macroeconomic model, we can provide behavioral principles of economic agents such as corporations and governments as follows [5, 6]:

- A principle of corporations under Thetical economics
  - Objective function (maximize profits)

$$\max \sum_{j=1}^{n} c_j x_j \tag{2}$$

  - Constraint condition

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad i = 1, \ldots, m \tag{3}$$

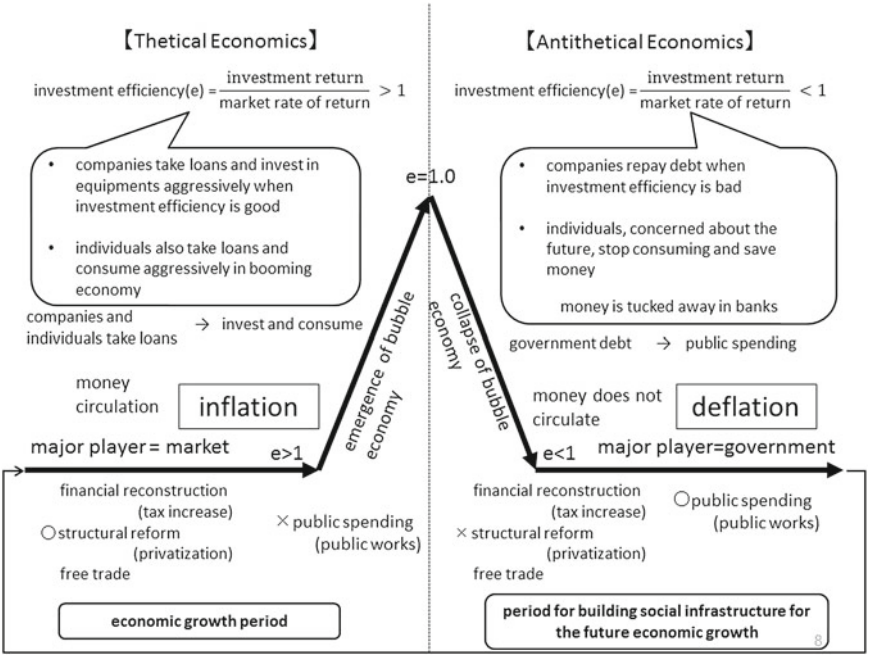


Fig. 2 A economic cycle. The figure is reproduced from a figure which was made by Kinoshita [5, 6]

- A principle of corporations under Antithetical economics
  - Objective function (minimize debts)

$$\min \sum_{i=1}^{m} u_i b_i \tag{4}$$

  - Constraint condition

$$\sum_{i=1}^{m} u_i a_{ij} \leq c_i, \quad j = 1, \ldots, n \tag{5}$$

Following list is correspondence of variables and its meanings.

x_j The number of units of a product j made by the corporation.
c_j The amount of profits of one unit of a product j; $P_j - (1 + r)h_j$, where P_j is price of the product j, r is interest rate, and h_j is cost of the product j.
a_{ij} Costs in an account subject i to produce the product j for one unit.
b_i The amount debts of an account subject i.
u_i Unpaid balance rate for the accounting subject i; $u_i = 1 -$ amortization_rate.

- A principle of governments under Thetical economics
  - Objective function (fiscal reconstruction)

$$\min \sum_{j=1}^{N} G_j K_j \tag{6}$$

  - Constraint condition

$$\sum_{j=1}^{N} A_{ij} K_j \leq B_i, \quad i = 1, \ldots, M \tag{7}$$

- A principle of governments under Antithetical economics
  - Objective function (fiscal stimulus)

$$\max \sum_{i=1}^{M} Y_i B_i \tag{8}$$

  - Constraint condition

$$\sum_{i=1}^{M} Y_i A_{ij} \leq c_i, \quad j = 1, \ldots, N \tag{9}$$

Following list is correspondence of variables and its meanings.

K_j A rate of the remainder of national loans for an administrative service j. Increasing the rate increases expenses of the service j.
G_j Demand for funds as national loans for an administrative service j.
A_{ij} Satisfaction of a resident i when the government gives the resident one unit of costs of a service j.
B_i A desiring level of total services of the government for a resident i.
Y_i The amount of public money to increase satisfaction by one unit for a resident i.

In usual studies of the macroeconomics, economic agents, such as customer, corporations, and governments, are modeled simply. All agents expand their profits, they are well-disciplined, they can acquire all information of markets, and their behavior is rational. The principles, which we provide, give a concrete mathematical model of the rationality.

The behavioral principle is linear equation system. Construction the principle is detecting parameters of the equations. So, the model has high affinity with correlation obtained from Big Data [7].

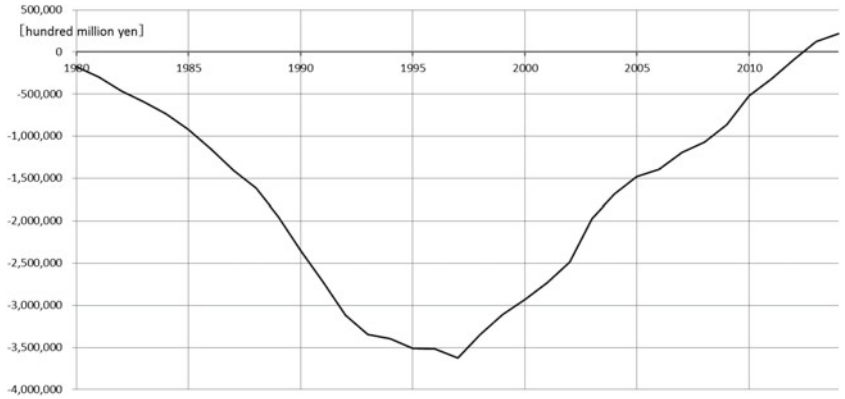


Fig. 3 Financial net worth of non-financial enterprises (total) in Japan from 1980 to 2015. The data is provided by the Bank of Japan

3 The Example of Big Data

As an example, we provide an explanation of macroeconomic phenomena in Japan since 1980 with the model. Let us see Fig. 3, which represents transition of financial net worth of corporations (non-financial enterprises) in Japan. Japan is dominated by Thetical economics before 1995, and is dominated by Antithetical economics after 1995.

Before 1995, corporations increase investments. It is an evidence of behavior of maximization of their profits; the Japanese economy was dominated by Thetical economics. In Japan, Heisei bubble collapse at February 1990. Five years later, Japanese economy was into recession in 1995. Since the year, corporations decrease their debts and increase their savings. It shows a change of behavioral principle of them; the economy is dominated by Antithetical economics.

GDP (Gross Domestic Products) is a macroeconomic index which represents business conditions of the nation. GDP (often denoted in Y) is sum of national consumption (C), national investment (I), governmental fiscal stimulus (G), and trade gap (E).

$$Y = C + I + G + E. \tag{10}$$

Transition of GDP of Japan is shown in Fig. 4. From the change of the index, we can confirm that Japanese corporations do not expand their profits since 1995.

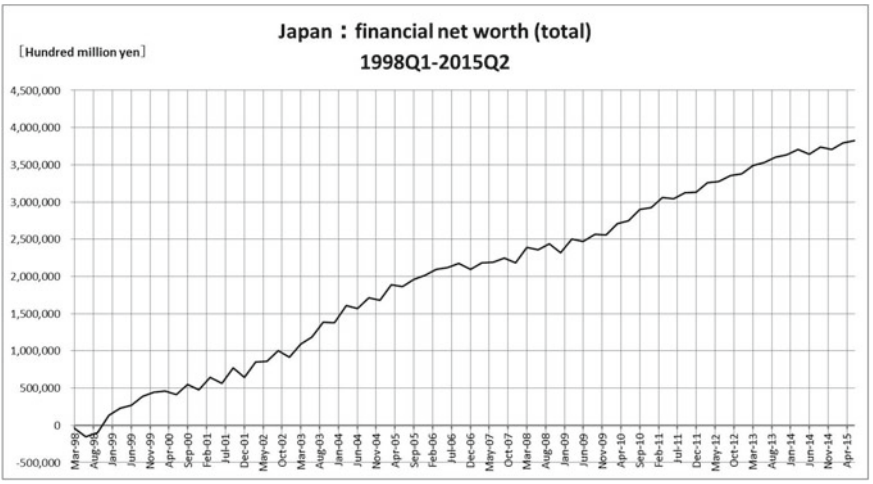


Fig. 4 Nominal GDP of Japan since 1980. The data is provided by the World Bank

4 Conclusions

In this chapter, we describe what Big Data are, and limitations use of Big Data. Experts, of course, use Big Data while considering the properties and limitations of them. We can grasp knowledges deduced from Big Data by paying attention to how a Big Data system treats the properties and overcomes the limitations.

We describe necessity of Big Data with a view from aspects of macroeconomics, and we provide a macroeconomic model with behavioral principles of economic agents. The principles have mathematical representation with high affinity of correlation deduced from Big Data. And we provide an explanation of macroeconomic phenomena in Japan since 1980 as an example of use of the model.

References

1. Kinoshita E (2011) A proposal of primal and dual problems in macro-economics. China-USA Bus Rev 10(2):115–124
2. Kinoshita E (2011) Why bubble economy occurs and crashes?-Repeated history of economic growth and collapse. Chin Bus Rev 10(2):102–111
3. Kinoshita E (2012) A proposal of thetical economy and antithetical economy-mechanism of occurrence and collapse of bubble economy. J Bus Econ (Acad Star Publ Co) 3(2):117–130
4. Kinoshita E, Mizuno T (2015) Trap of economics the world has fallen in a survey of Kinoshita theory in macro-economics. Eur Sci J:75–80 (2015). SPECIAL edition
5. Kinoshita E (2015) A proposal of thetical economy and antithetical economy by using operations research techniques. Eur Sci J 11(19):29–48 (2015). July 2015 edition
6. Kinoshita E (2015) Thetical and antithetical business management. J Bus Econ 6(6):1086–1096 (2015). June 2015
7. Kinoshita E, Mizuno T (2016, in printing) Why we need big data? In: Proceeding of the 3rd international conference on advances in big data analytics

Big Data for Conversational Interfaces: Current Opportunities and Prospects

David Griol, Jose M. Molina and Zoraida Callejas

Abstract As conversational technologies develop, we demand more from them. For instance, we want our conversational assistants to be able to solve our queries in multiple domains, to manage information from different usually unstructured sources, to be able to perform a variety of tasks, and understand open conversational language. However, developing the resources necessary to develop systems with such capabilities demands much time and effort, as for each domain, task or language, data must be collected, annotated following an schema that is usually not portable, the models must be trained over the annotated data, and their accuracy must be evaluated. In recent years, there has been a growing interest in investigating alternatives to manual effort that allow exploiting automatically the huge amount of resources available in the web. In this chapter we describe the main initiatives to extract, process and contextualize information from these rich and heterogeneous sources for the various tasks involved in dialog systems, including speech processing, natural language understanding and dialog management.

Keywords Conversational interfaces · Big Data · Spoken interaction · Automatic speech recognition · Dialog management · Speech synthesis

D. Griol (✉) · J.M. Molina
Department of Computer Science, Carlos III University of Madrid, Avda, de la Universidad, 30, 28911 Leganés, Spain
e-mail: david.griol@uc3m.es

J.M. Molina
e-mail: josemanuel.molina@uc3m.es

Z. Callejas
Department of Languages and Computer Systems, University of Granada, CITIC-UGR, C/ Pdta. Daniel Saucedo Aranda S/n, 18071 Granada, Spain
e-mail: zoraida@ugr.es

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_6

1 Introduction

Speech and natural language technologies allow users to communicate in a flexible and efficient way, also enabling the access to applications when traditional input and output interfaces cannot be used (e.g. in-car applications, access for disabled persons, etc.). Also speech-based interfaces work seamlessly with small devices (e.g., smarthphones and tablets PCs) and allow users to easily invoke local applications or access remote information. For this reason, spoken dialog systems [22, 48, 59] are becoming a strong alternative to traditional graphical interfaces which might not be appropriate for all users and/or applications.

These systems are computer programs that receive speech as input and generate as output synthesized speech, engaging the user in a dialog that aims to be similar to that between humans [48, 59]. Thus, these interfaces make technologies more usable, as they ease interaction [23], allow integration in different environments [22], and make technologies more accessible, especially for disabled people and the elderly [80].

In a dialog system of this kind, several modules cooperate to perform the interaction with the user: the Automatic Speech Recognizer (ASR), the Spoken Language Understanding Module (SLU), the Dialog Manager (DM), the Natural Language Generation module (NLG), and the Text-To-Speech Synthesizer (TTS). Each one of them has its own characteristics and the selection of the most convenient model varies depending on certain factors: the goal of each module, the possibility of manually defining the behavior of the module, or the capability of automatically obtaining models from training samples. Figure 1 shows the set of actions and main modules in the architecture of a Spoken Dialog System.

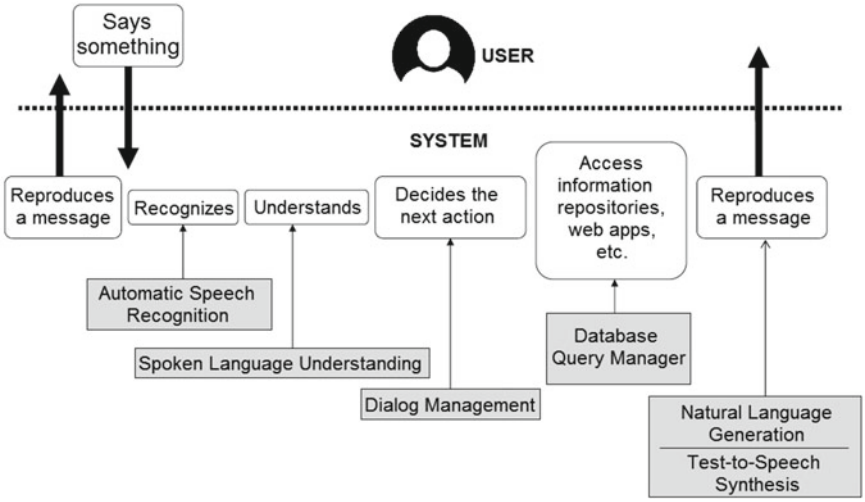


Fig. 1 Set of actions and modules in a spoken dialog system

The goal of speech recognition is to obtain the sequence of words uttered by a speaker. Once the speech recognizer has provided an output, the system must understand what the user said. The goal of spoken language understanding is to obtain the semantics from the recognized sentence. This process generally requires morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge.

The dialog manager decides the next action of the system, interpreting the incoming semantic representation of the user input in the context of the dialog. In addition, it resolves ellipsis and anaphora, evaluates the relevance and completeness of user requests, identifies and recovers from recognition and understanding errors, retrieves information from data repositories, and decides about the next system's response. Natural language generation is the process of obtaining sentences in natural language from the non-linguistic, internal representation of information handled by the dialog system. Finally, the TTS module transforms the generated sentences into synthesized speech.

In order to enable rapid deployment of these systems, markup languages such as VoiceXML[1] have been widely adopted as they reduce the time and effort required for system implementation. However, system development with this approach involves a very costly engineering cycle [62]. As an alternative, data-based models try to reduce the effort and time required to develop a new dialog system or to adapt them to deal with a new task. This kind of models are usually based on modeling the different processes probabilistically and learning the parameters of the different statistical models from a dialog corpus. This approach has been widely used for speech recognition and also for language understanding [14, 20, 37, 51, 67]. Even though in the literature there are models for dialog managers that are manually designed, over the last few years, approaches using statistical models to represent the behavior of the dialog manager have also been developed [36, 38, 74, 83].

As described by [58], there are three main categories of elements of the spoken dialog interaction where the availability of vast amounts of data (known as Big Data [2, 19, 47]) can potentially improve automation rate, and ultimately, the penetration and acceptance of speech interfaces in the wider consumer market. They are task-independent behaviors (e.g., error correction and confirmation behavior), task-specific behaviors (e.g., logic associated with certain customer-care practices), and task-interface behaviors (e.g., prompt selection). However, these three categories have in common today the lack of robust guiding principles validated by empirical evidence.

The following sections of this chapter describe the current uses of Big Data to develop conversational interfaces including speech recognition, natural language understanding, dialog management and optimization, context-awareness, emotion recognition, user adaptation and service personalization, multi-domain and multilingual services, proactiveness, and spoken language generation and synthesis.

[1] http://www.w3.org/TR/voicexml20/.

2 Spoken Language Recognition

As described in the introduction section, speech recognition is the process of obtaining the text string corresponding to an acoustic input [45, 55, 78]. It is a highly complex task, as there is a great deal of variation in input characteristics, which can differ according to the linguistics of the utterance, the speaker, the interaction context and the transmission channel. Different aspects that are usually taken into account when classifying ASR systems are the kind of users supported (user-independent or user-dependent systems), style of speech supported (recognizers isolated words, connected words or continuous speech), or vocabulary size (small, medium, or large vocabulary).

The complexity of the recognition task lies in several problems: the acoustic variability (each person pronounces sounds differently when speaking), acoustic confusion (many words sound similar), the coarticulation problem (the characteristics of spoken sounds may vary depending on neighboring sounds), out of vocabulary words and spontaneous speech (interjections, pauses, doubts, false starts, repetitions of words, self-corrections, etc.), and environmental conditions (noise, channel distortion, bandwidth limitations, etc.). For these reasons, it is very important to try to detect and correct errors generated during the ASR process, since the output of the ASR is the starting point of the other modules in a spoken dialog system.

During the last decades, the field of automatic speech recognition has progressed from the recognition of isolated words in reduced vocabularies to continuous speech recognition with increasing vocabulary sets. These advances have made the communication with dialog systems increasingly more natural. Among the variety of techniques used to develop ASR systems, the data-based approach is currently the most widely used. In this approach, the speech recognition problem can be understood as finding the word sequence W uttered by the user given a sequence of acoustic data A. This sequence can be determined by means of the following expression:

$$W = \max_{W} P(W|A) \tag{1}$$

Using the Bayes rule, the previous equation can be rewritten as follows:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)} \tag{2}$$

where $P(A|W)$ is called the acoustic model (probability of the sequence A the word sequence W has been uttered) and $P(W)$ is provided by the language model (probabilities of sequences of words.). The probabilities of the rules in these models are learned from training data. The acoustic model is created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for the different words. Learning a language model requires the transcriptions of sentences related to the application domain of the system. Since

the probability of the acoustic sequence is independent of the sequence of words, the previous expression can be written as follows:

$$W = \max_{W} P(A|W)P(W) \tag{3}$$

For the practical implementation of this approach, the most widely used solution consists of modeling the acoustic units by means of Hidden Markov Models (HMM), as it is the case of speech recognizers widely used by the scientific community as HTK (Hidden Markov Model Toolkit)[2] or CMU Sphinx.[3]

The success of the HMM is mainly based on the use of machine learning algorithms to learn the parameters of the model [61], as well as in their ability to represent speech as a sequential phenomenon over time. Multiple models have been studied, such as discrete models, semicontinuous or continuous, as well as a variety of topologies models.

The language model is one of the essential components required to develop a recognizer of continuous speech. The most used language models are based on N-grams [3, 28] and regular or context-free grammars [29, 67]. Grammars are usually suitable for small tasks, providing more precision based on the type of restrictions. However, they are not able to represent the great variability of natural speech processes.

N-grams models allow to collect more easily the different concatenations among words when a sufficient number of training samples is available. In an n-gram model, the probability $P(w_1, \ldots, w_m)$ of observing the sentence $w_1, \ldots, w_m$ is approximated as

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \tag{4}$$

This equation assumes that the probability of observing the i-th word w_i in the context history of the preceding $i-1$ words can be approximated by the probability of observing it in the shortened context history of the preceding $n-1$ words (n-th order Markov property). The conditional probability can be calculated from n-gram model frequency counts:

$$P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1}, w_i)}{\text{count}(w_{i-(n-1)}, \ldots, w_{i-1})} \tag{5}$$

Figure 2 shows an example of the estimation of the bigram probabilities using the Maximum Likelihood Estimate.[4] Typically, the probabilities are not derived directly from the frequency counts. Instead, some form of smoothing is necessary, assigning

[2] http://htk.eng.cam.ac.uk/.

[3] http://cmusphinx.sourceforge.net/.

[4] https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf.

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$
$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

Fig. 2 Estimating bigram probabilities by means of the maximum likelihood estimate

some of the total probability mass to unseen words or n-grams. Various methods are then used, from simple "add-one" smoothing (assign a count of 1 to unseen n-grams) to more sophisticated models, such as Good-Turing discounting or back-off models.

From around 2010, Deep Neural Networks (DNNs) have replaced HMM models. DNNs are now used extensively in industrial and academic research as well as in most commercially deployed ASR systems. Various studies have shown that DNNs outperform HMM models in terms of increased recognition accuracy [24, 68]. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. As described in [53], a key benefit of Deep Learning is the analysis and learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabeled and uncategorized.

3 Spoken Language Understanding

Once the spoken dialog system has recognized what the user uttered, it is necessary to understand what he said [46, 50, 85]. Natural language processing is a method of obtaining the semantics of a text string and generally involves morphological, lexical, syntactical, semantic, discourse and pragmatical knowledge. In the first stage, lexical and morphological knowledge divide the words into their constituents by distinguishing between lexemes and morphemes: lexemes are parts of words that indicate their semantics and morphemes are the different infixes and suffixes that provide different word classes.

Syntactic analysis yields the hierarchical structure of the sentences. However, in spoken language, phrases are frequently affected by difficulties associated with the so-called disfluency phenomena: filled pauses, repetitions, syntactic incompleteness and repairs [18]. Semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituent parts. In the pragmatic and discourse-processing stage, the sentences are interpreted in the context of the whole dialog, the main complexity of this stage is the resolution of anaphora, and ambiguities derived from phenomena such as irony, sarcasm or double entendre.

The process of understanding can be understood as a change in language representation, from natural language to a semantic language, so that the meaning of the message is not changed. As in the speech recognizer, the spoken language understanding module can work with several hypotheses (both for recognition and understanding) and confidence measures. There are currently two major approaches to tackling the problem of understanding: rule-based approaches and statistical models learned from data corpus.

Rule-based approaches extract semantic information based on a syntactic-semantic analysis of the sentences, using grammars defined for the task, or by means of the detection of keywords with semantic meanings. Some analyzers, in order to improve the robustness of the analysis, combine syntactic and semantic aspects of the specific task. Other techniques are based on an analysis at two levels, in which grammars are used to carry out a detailed analysis of the sentence and extract relevant semantic information. In addition, there are systems that use rule-based analyzers automatically learned from a training corpus using natural language processing techniques.

In the case of statistical methods, the process is based on the definition of linguistic units with semantic content and obtaining models from labeled samples. This type of analysis [50, 67] uses a probabilistic model to identify concepts, markers and values of cases, to represent the relationship between markers of cases and their values and to decode semantically pronunciations of the user. The model is generated during a training phase (learning), where its parameters capture the correspondences between text entries and semantic representation. Once the training model has been learned, it is used as a decoder to generate the best representation.

The semantic definition is usually based on the concept of frame in most of the current dialog systems. In this approach, the representation generated by the spoken language understanding module contains concepts (different types of queries that users can require to the system) and attributes (information to be provided by the user to complete or modify the queries). Thus, every message sent by the spoken language understanding module to the dialog manager after each user utterance consists of a frame structure.

4 Dialog Management

Although dialog management is only a part of the development cycle of spoken dialog systems, it can be considered one of the most demanding tasks given that this module encapsulates the logic of the speech application [81]. [77] state that dialog management involves four main tasks: (i) updating the dialog context, (ii) providing a context for sentence interpretation, (iii) coordinating other modules and (iv) deciding the information to convey to the user and when to do it. Thus, the selection of a specific system action depends on multiple factors, such as the output of the speech recognizer (e.g., measures that define the reliability of the recognized information), the dialog interaction and previous dialog history (e.g., the number of

repairs carried out so far), the application domain (e.g., guidelines for customer service), knowledge about the users, and the responses and status of external back-ends, devices, and data repositories. Given that the actions of the system directly impact users, the dialog manager is largely responsible for user satisfaction. This way, the design of an appropriate dialog management strategy is at the core of dialog system engineering.

Statistical approaches for dialog management present several important advantages with regard traditional rule-based methodologies. Rather than maintaining a single hypothesis for the dialog state, they maintain a distribution over many hypotheses for the correct dialog state. In addition, statistical methodologies choose actions using an optimization process, in which a developer specifies high-level goals and the optimization works out the detailed dialog plan. For instance, Hoxha and Weng have very recently proposed a mixed-initiative dialog-based approach to support autonomous clinical data access and recommend needed technology development and communication study for accelerating clinical research [26].

Automating dialog management is useful for developing, deploying and re-deploying applications and also reducing the time-consuming process of handcrafted design. In fact, the application of machine learning approaches to dialog management strategy design is a rapidly growing research area. Machine-learning approaches to dialog management attempt to learn optimal strategies from corpora of real human-computer dialog data using automated "trial-and-error" methods instead of relying on empirical design principles [86]. The main trend in this area is an increased use of data for automatically improving the performance of the system.

Statistical models can be trained with corpora of human-computer dialogs with the goal of explicitly modeling the variance in user behavior that can be difficult to address by means of hand-written rules [66]. Additionally, if it is necessary to satisfy certain deterministic behaviors, it is possible to extend the strategy learned from the training corpus with handcrafted rules that include expert knowledge or specifications about the task [32, 72, 75, 87].

The goal is to build systems that exhibit more robust performance, improved portability, better scalability and easier adaptation to other tasks. However, model construction and parameterization is dependent on expert knowledge, and the success of statistical approaches is dependent on the quality and coverage of the models and data used for training [66]. Moreover, the training data must be correctly labeled for the learning process. The size of currently available annotated dialog corpora is usually too small to sufficiently explore the vast space of possible dialog states and strategies. Collecting a corpus with real users and annotating it requires considerable time and effort.

To address these problems, researchers have proposed alternative techniques that facilitate the acquisition and labeling of corpora, such as Wizard of Oz [16, 31], bootstrapping [1, 15], active learning [9, 41], automatic dialog act classification and labeling [56, 79], and user simulation [43, 66].

Another relevant problem is how to deal with unseen situations, that is, situations that may occur during the dialog and that were not considered during training. To address this point it is necessary to employ generalizable models in order to obtain appropriate system responses that enable to continue with the dialog in a satisfactory way.

Another difficulty is in the design of a good dialog strategy, which in many cases is far from being trivial. In fact, there is no clear definition of what constitutes a good dialog strategy [35, 66]. Users are diverse, which makes it difficult to foresee which form of system behavior will lead to quick and successful dialog completion, and speech recognition errors may introduce uncertainty about their intention.

The most widespread methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Markov Decision Processes (MDP) and reinforcement methods [38, 70]. The main drawback of this approach is that the large state space of practical spoken dialog systems makes its direct representation intractable [89]. Partially Observable MDPs (POMDPs) outperform MDP-based dialog strategies since they provide an explicit representation of uncertainty [63]. This enables the dialog manager to avoid and recover from recognition errors by sharing and shifting probability mass between multiple hypotheses of the current dialog state.

An approach that scales the POMDP framework for implementing practical spoken dialog systems by the definition of two state spaces is presented in [88]. Approximate algorithms have also been developed to overcome the intractability of exact algorithms but even the most efficient of these techniques such as Point-Based Value Iteration (PBVI) cannot scale to the many thousand states required by a statistical dialog manager [82]. Composite Summary Point-Based Value Iteration (CSPBVI) has suggested the use of a small summary space for each slot where PBVI policy optimization can be applied. However, policy learning in this technique can only be performed offline, i.e. at design time, because policy training requires an existing accurate model of user behavior. An alternative technique for online training based on Q-learning is presented in [73], which allows the system to adapt to real users as new dialogs are recorded. This technique does not require any model of user behavior, so user simulation techniques are proposed to iteratively learn the dialog model.

Other authors have combined conventional dialog managers with a fully-observable Markov decision process [21, 71], or proposed using multiple POMDPs and selecting actions using hand-crafted rules [82]. In [84], the authors combine the robustness of the POMDP with the developer control afforded in conventional approaches: the (conventional) dialog manager and POMDP run in parallel, but the dialog manager is augmented so that it outputs one or more allowed actions at each time-step. The POMDP then chooses the best action from this limited set. Results from a real voice dialer application show that adding the POMDP machinery to a standard dialog system can yield a significant improvement [84].

Other interesting approaches for statistical dialog management are based on modeling the system by means of Hidden Markov Models [10], stochastic Finite-State Transducers [25, 27, 60], or using Bayesian Networks [49, 57]. Also [33] proposed

a different hybrid approach to dialog modeling in which n-best recognition hypotheses are weighted using a mixture of expert knowledge and data-driven measures by using an agenda and an example-based machine translation approach respectively.

5 Natural Language Generation

Natural language generation is the process of obtaining texts in natural language from a non-linguistic representation [34, 42]. It is usually carried out in 5 steps: content organization, content distribution in sentences, lexicalization, generation of referential expressions and linguistic realization. It is important to obtain legible messages, optimizing the text using referring expressions and linking words and adapting the vocabulary and the complexity of the syntactic structures to the users linguistic expertise.

The simplest approach consists of using predefined text messages (e.g. error messages and warnings). Although intuitive, this approach completely lacks from flexibility. The next level of sophistication is template-based generation, in which the same message structure is produced with slight alterations. The template approach is used mainly for multi-sentence generation, particularly in applications whose texts are fairly regular in structure, such as business reports.

Phrase-based systems employ what can be considered as generalized templates at the sentence level (in which case the phrases resemble phrase structure grammar rules), or at the discourse level (in which case they are often called text plans). In such systems, a pattern is first selected to match the top level of the input, and then each part of the pattern is expanded into a more specific one that matches some portion of the input. The cascading process stops when every pattern has been replaced by one or more words.

Finally, feature-based systems represent the maximum level of generalization and flexibility. In feature-based systems, each possible minimal alternative of expression is represented by a single feature; for example, whether the sentence is either positive or negative, if it is a question or an imperative or a statement, or its tense. To arrange the features it is necessary to employ linguistic knowledge. Another alternative is to use corpus-based natural language generation [54], which stochastically generates system utterances.

6 Text-To-Speech Synthesis

Text-to-speech synthesizers transform a text into an acoustic signal [11]. A text-to-speech system is composed of two parts: a front-end and a back-end. The front-end carries out two major tasks. Firstly, it converts raw text containing symbols such as numbers and abbreviations into their equivalent words. This process is often called text normalization, pre-processing, or tokenization. Secondly, it assigns a phonetic

transcriptions to each word, and divides and marks the text into prosodic units, i.e. phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. The output of the front-end is the symbolic representation constituted by the phonetic transcriptions and prosody information.

The back-end (often referred to as the synthesizer) converts the symbolic linguistic representation into sound. On the one hand, speech synthesis can be based on human speech production. This is the case of parametric synthesis which simulates the physiological parameters of the vocal tract, and formant-based synthesis, which models the vibration of vocal chords. In this technique, parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. Another approach based on physiological models is articulatory synthesis, which refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes.

On the other hand, concatenative synthesis employs pre-recorded units of human voice. Concatenative synthesis is based on stringing together segments of recorded speech. It generally produces the most natural-sounding synthesized speech; however, differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output. The quality of the synthesized speech depends on the size of the synthesis unit employed.

Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. Unit selection provides the greatest naturalness, because it applies only a small amount of digital signal processing to the recorded speech. There is a balance between intelligibility and naturalness of the voice output or the automatization of the synthesis procedure. For example, synthesis based on whole words is more intelligible than the phone-based but for each new word it is necessary to obtain a new recording, whereas the phones allow building any new word. In one extreme, domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. It is used in applications in which the variety of texts the system will produce is limited to a particular domain, like transit schedule announcements or weather reports.

At the other extreme, diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language. The number of diphones depends on the phonotactics of the language: for example, Spanish has about 800 diphones and German about 2,500. In diphone synthesis, only one example of each diphone is contained in the speech database. Finally, HMM-based synthesis is a method in which the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves, based on the maximum likelihood criterion.

7 User Modeling and Evaluation of the System

Research in techniques for user modeling has a long history within the fields of language processing and conversational agents. The main purpose of a simulated user in this field is to improve the usability of a conversational agent through the generation of corpora of interactions between the system and simulated users [52], reducing time and effort required for collecting large samples of interactions with real users. Moreover, each time changes are made to the system it is necessary to collect more data in order to evaluate the changes. Thus, the availability of large corpora acquired with a user simulator should contribute positively to the development of the system.

User simulators can be used to evaluate different aspects of a conversational agent, particularly at the earlier stages of development, or to determine the effects of changes to the system's functionalities (e.g., evaluate confirmation strategies or introduce of errors or unpredicted answers in order to evaluate the capacity of the dialog manager to react to unexpected situations). A second usage, in which we are mainly interested in this contribution, is to support the automatic learning of optimal dialog strategies using statistical methodologies. Large amounts of data are required for a systematic exploration of the dialog state space and corpora acquired with simulated users are extremely valuable for this purpose.

Two main approaches can be distinguished to the creation of simulated users: rule based and data or corpus based. In a rule-based simulated user the researcher can create different rules that determine the behavior of the system [8, 39, 44]. This approach is particularly useful when the purpose of the research is to evaluate the effects of different dialog management strategies. In this way the researcher has complete control over the design of the evaluation study.

Data-based user models are based on probabilistic methods to generate the user input, with the advantage that this uncertainty can better reflect the unexpected behaviors of users interacting with the system. Statistical models for modeling user behavior have been suggested as the solution to the lack of the data that is required for training and evaluating dialog strategies. Using this approach, the dialog manager can explore the space of possible dialog situations and learn new potentially better strategies. Methodologies based on learning user intentions have the purpose of optimizing dialog strategies. A summary of user simulation techniques for reinforcement learning of the dialog strategy can be found in [66].

The most extended methodology for machine-learning of dialog strategies consists of modeling human-computer interaction as an optimization problem using Markov Decision Process (MDP) and reinforcement methods [38]. The main drawback of this approach is the large state space of practical spoken dialog systems, whose representation is intractable if represented directly. Although Partially Observable MDPs (POMDPs) outperform MDP-based dialog strategies, they are limited to small-scale problems, since the state space would be huge and exact POMDP optimization is again intractable [83].

In [12, 13], Eckert, Levin and Pieraccini introduced the use of statistical models to predict the next user action by means of a n-gram model. The proposed model has the advantage of being both statistical and task-independent. Its weak point consists of approximating the complete history of the dialog by a bigram model. In [38], the bigram model is modified by considering only a set of possible user answers following a given system action (the Levin model). Both models have the drawback of considering that every user response depends only on the previous system turn. Therefore, the simulated user can change objectives continuously or repeat information previously provided.

Georgila, Henderson and Lemon propose the use of HMMs, defining a more detailed description of the states and considering an extended representation of the history of the dialog [17]. Dialog is described as a sequence of *Information States* [7]. Two different methodologies are described to select the next user action given a history of information states. The first method uses n-grams [12], but with values of n from 2 to 5 to consider a longer history of the dialog. The best results are obtained with 4-grams. The second methodology is based on the use of a linear combination of 290 characteristics to calculate the probability of every action for a specific state.

Cuayáhuitl et al. present a method for dialog simulation based on HMMs in which both user and system behaviors are simulated [10]. Instead of training only a generic HMM model to simulate any type of dialog, the dialogs of an initial corpus are grouped according to the different objectives. A submodel is trained for each one of the objectives, and a bigram model is used to predict the sequence of objectives.

In [64], a new technique for user simulation based on explicit representations of the user goal and the user agenda is presented. The user agenda is a structure that contains the pending user dialog acts that are needed to elicit the information specified in the goal. This model formalizes human-machine dialogs at a semantic level as a sequence of states and dialog acts. An EM-based algorithm is used to estimate optimal parameter values iteratively. In [65], the agenda-based simulator is used to train a statistical POMDP-based dialog manager.

A data-driven user intention simulation method that integrates diverse user discourse knowledge (cooperative, corrective, and self-directing) is presented in [30]. User intention is modeled based on logistic regression and Markov logic framework. Human dialog knowledge is designed into two layers, domain and discourse knowledge, and integrated with the data-driven model in generation time. A methodology of user simulation applied to the evaluation and refinement of stochastic dialog systems is presented in [76]. The proposed user simulator incorporates several knowledge sources, combining statistical and heuristic information to enhance the dialog models by an automatic strategy learning. As it is described in the following section, our proposed user simulation technique is based on a classification process that considers the complete dialog history by incorporating several knowledge sources, combining statistical and heuristic information to enhance dialog models by an automatic strategy learning.

In the area of user modeling and dialog systems, emotion has been used for several purposes, as summarized in the taxonomy of applications proposed in [5]. In some application domains, it is fundamental to recognize the affective state of the user to

adapt the systems behavior. For example, in emergency services [6] or intelligent tutors [40], it is necessary to know the user emotional state to calm them down, or to encourage them in learning activities. For other applications domains, it can also play an important role in order to solve stages of the dialog that cause negative emotional states, avoid them and foster positive ones in future interactions. Bain have recently presented a proposal to extract emotional information using cloud-based Big Data infrastructure and mobile devices [4]. Hosain et al. has also very recently proposed an infrastructure that combines the potential of emotion-aware Big Data and cloud technology towards the future generation mobile communication technologies (5G) [69].

8 Future Research and Challenges

Throughout the last years, some experts have dared to envision what the future research guidelines in the application of multimodal dialog systems for educative purposes would be based on the advances in Big Data research. These objectives have gradually changed towards ever more complex goals, such as providing the system with advanced reasoning, problem solving capabilities, adaptiveness, proactiveness, affective intelligence, and multilinguality. All these concepts are not mutually exclusive, as for example the system's intelligence can also be involved in the degree to which it can adapt to new situations, and this adaptiveness can result in better portability for use in different environments.

As can be observed, these new objectives refer to the system as a whole, and represent major trends that in practice are achieved through joint work in different areas and components of the dialog system. Thus, current research trends are characterized by large-scale objectives which are shared out between the different researchers in different areas.

Proactiveness is necessary for computers to stop being considered a tool and becoming real conversational partners. Proactive systems have the capability of engaging in a conversation with the user even when he has not explicitly requested the system's intervention. This is a key aspect in the development of ubiquitous computing architectures in which the system is embedded in the user's environment, and thus the user is not aware that he is interacting with a computer, but rather he perceives he is interacting with the environment. To achieve this goal, it is necessary to provide the systems with problem-solving capabilities and context-awareness.

Adaptivity may also refer to other aspects in speech applications. There are different levels in which the system can adapt to the user. The simplest one is through personal profiles in which the users have static choices to customize the interaction. Systems can also adapt to the users' environment, for example ambient intelligence applications such as the ubiquitous proactive systems described. A more sophisticated approach is to adapt to the user's knowledge and expertise. This is especially important in educative systems to adapt the system taking into account the specific

evolution of each of the students, the previous uses of the system, and the errors that they have made during the previous interactions.

There is also an increasing interest in the development of multimodal conversational systems that dynamically adapt their conversational behaviors to the users' affective state. The empathetic educative agent can thus indeed contribute to a more positive perception of the interaction.

Portability is currently addressed from very different perspectives, the three main ones being domain, language and technological independence. Ideally, systems should be able to work over different educative application domains, or at least be easily adaptable between them. Current studies on domain independence center on how to merge lexical, syntactic and semantic structures from different contexts and how to develop dialog managers that deal with different domains.

Finally, technological independence deals with the possibility of using multimodal systems with different hardware configurations. Computer processing power will continue to increase, with lower costs for both processor and memory components. The systems that support even the most sophisticated multimodal applications will move from centralized architectures to distributed configurations and thus must be able to work with different underlying technologies.

9 Conclusions

Dialog systems appeared as a technology aimed at sustaining conversations with their users that could be considered natural and human-like. However, to achieve this long pursued objective, these systems must be able to operate in a wide range of domains and tasks, some of them difficult to process and complex [19], for which being able to learn from massive amounts of data becomes crucial to show appropriate behaviors.

We have addressed Big Data as (1) new sources of huge amounts of data, and (2) as the novel machine learning and information extraction algorithms that have appeared to process them. On the one hand, web pages and searches, social network, blog posts, and emails, they all provide an invaluable source for natural language resources. Similarly, voice calls, recorded dialogs and conversations have a huge potential to provide insights into human conversational behavior. However, manually examining such Big Data is laborious and error-prone. On the other hand, the emergence of different statistical approaches has enabled to accurately analyze unstructured data with a double benefit: to be less dependent on intensive manual annotation, and to gain a better understanding of human conversation by learning more accurate models from a more representative amount of data.

In this chapter we have discussed the tremendous potential of Big Data to improve several aspects of dialog system research and development, including speech processing, natural language understanding and dialog management.

References

1. Abdennadher S, Aly M, Bhler D, Minker W, Pittermann J (2007) Becam tool - a semi-automatic tool for bootstrapping emotion corpus annotation and management. In: Proceedings of the international conference on spoken language processing (Interspeech'2007), pp 946–949
2. Agerri R, Artola X, Beloki Z, Rigau G, Soroa A (2015) Big data for natural language processing: a streaming approach. Knowl-Based Syst 79:36–42
3. Bahl L, Jelinek F, Mercer R (1990) A maximum likelihood approach to continuous speech recognition. Readings in Speech recognition, pp 308–319
4. Baimbetov Y, Khalil I, Steinbauer M, Anderst-Kotsis G (2015) Using Big Data for emotionally intelligent mobile services through multi-modal emotion recognition. Springer, pp 127–138
5. Batliner A, Burkhardt F, van Ballegooy M, Noth E (2006) A taxonomy of applications that utilize emotional awareness. In: Proceedings of 1st international language technologies conference (IS-LTC 06), pp 246–250
6. Bickmore T, Giorgino T (2004) Some novel aspects of health communication from a dialogue systems perspective. In: Proceedings of AAAI fall symposium on dialogue systems for health communication, pp 275–291
7. Bos J, Klein E, Lemon O, Oka T (2003) DIPPER: description and formalisation of an information-state update dialogue system architecture. In: Proceedings of the SIGdial, pp 115–124
8. Chung G (2004) Developing a flexible spoken dialog system using simulation. In: Proceedings of ACL, pp 63–70
9. Cohn DA, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2):201–221
10. Cuayhuitl H, Renals S, Lemon O, Shimodaira H (2005) Human-computer dialogue simulation using hidden Markov models. In: Proceedings of ASRU, pp 290–295
11. Dutoit T (1996) An introduction to text-to-speech synthesis. Kluwer Academic Publishers
12. Eckert W, Levin E, Pieraccini R (1997) User modeling for spoken dialogue system evaluation. In: Proceedings of ASRU, pp 80–87
13. Eckert W, Levin E, Pieraccini R (1998) Automatic evaluation of spoken dialogue systems. Technical report, TR98.9.1, ATT Labs Research
14. Esteve Y, Raymond C, Bechet F, Mori RD (2003) Conceptual decoding for spoken dialog systems. In: Proceedings of European conference on speech communications and technology (Eurospeech'03). vol 1, pp 617–620
15. Fabbrizio GD, Tur G, Hakkani-Tr D, Gilbert M, Renger B, Gibbon D, Liu Z, Shahraray B (2008) Bootstrapping spoken dialogue systems by exploiting reusable libraries. Nat Lang Eng 14(3):313–335
16. Fraser M, Gilbert G (1991) Simulating speech systems. Comput Speech Lang 5:81–99
17. Georgila K, Henderson J, Lemon O (2005) Learning user simulations for information state update dialogue systems. In: Proceedings of Eurospeech'05, pp. 893–896
18. Gibbon D, Mertins I (Eds.), R.M.: Handbook of multimodal and spoken dialogue systems: resources, terminology and product evaluation. Kluwer Academic Publishers (2000)
19. Gudivada VN, Rao D, Raghavan VV (2015) Big data driven natural language processing research and applications, vol 33. Elsevier, pp 203–238
20. He Y, Young S (2003) A data-driven spoken language understanding system. In: Proceedings of IEEE Automatic speech recognition and understanding workshop (ASRU'03), pp 583–588
21. Heeman P (2007) Combining reinforcement learning with information-state update rules. In: Proceedings of the 8th Annual conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07), pp 268–275
22. Heinroth T, Minker W (2012) Introducing spoken dialogue systems into intelligent environments. Kluwer Academic Publishers, Springer
23. Hempel T (2008) Usability of speech dialog systems: listening to the target audience. Springer

24. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
25. Hori C, Ohtake K, Misu T, Kashioka H, Nakamura S (2009) Recent advances in WFST-based dialog system. In: Proceedings of the international conference on spoken language processing (Interspeech'2009), pp 268–271
26. Hoxha J, Weng C (2016) Leveraging dialog systems research to assist biomedical researchers interrogation of big clinical data. J Biomed Inf 61:176–184
27. Hurtado L, Planells J, Segarra E, Sanchis E, Griol D (2010) A stochastic finite-state transducer approach to spoken dialog management. In: Proceedings of the international conference on spoken language processing (Interspeech'2010), pp 3002–3005
28. Jelinek F (1990) Self-organized language modeling for speech recognition. Readings in Speech recognition, pp 450–506
29. Jelinek F, Lafferty MR (1992) Basic methods of probabilistic context free grammars. Springer, pp 345–360
30. Jung S, Lee C, Kim K, Lee D, Lee G (2011) Hybrid user intention modeling to diversify dialog simulations. Comput Speech Lang 25(2):307–326
31. Lane I, Ueno S, Kawahara T (2004) Cooperative dialogue planning with user and situation models via example-based training. In: Proceedings of workshop on man-machine symbiotic systems, pp 2837–2840, Kyoto, Japan
32. Laroche R, Putois G, Bretier P, Young S, Lemon O (2008) Requirements analysis and theory for statistical learning approaches in automaton-based dialogue management. Technical report, School of Informatics, Edinburgh University, Edinburgh, UK
33. Lee C, Jung S, Kim K, Lee GG (2010) Hybrid approach to robust dialog management using agenda and dialog examples. Comput Speech Lang 24(4):609–631
34. Lemon O (2011) Learning what to say and how to say it: joint optimisation of spoken dialogue management and natural language generation. Comput Speech Lang 25(2):210–221
35. Lemon O, Pietquin O (2012) Data-Driven methods for adaptive spoken dialogue systems. Computational learning for conversational interfaces. Springer, Berlin
36. Lemon O, Georgila K, Henderson J (2006) Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the TALK TownInfo evaluation. In: Proceedings of IEEE-ACL workshop on spoken language technology (SLT'06), pp 178–181
37. Levin E, Pieraccini R (1995) Concept-based spontaneous speech understanding system. In: Proceedings of European conference on speech communications and technology (Eurospeech'95). pp. 555–558 (1995)
38. Levin E, Pieraccini R, Eckert W (2000) A stochastic model of human-machine interaction for learning dialog strategies. IEEE Trans Speech Audio Process 8(1):11–23
39. Lin B, Lee L (2001) Computer aided analysis and design for spoken dialogue systems based on quantitative simulations. IEEE Trans Speech Audio Process 9(5):534–548
40. Litman D, Forbes-Riley K (2006) Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun 48(5):559–590
41. Liu Y, Shriberg E (2005) Does active learning help automatic dialog act tagging in meeting data. In: Proceedings of the international conference on spoken language processing (Interspeech'2005), pp 2777–2780, Lisbon, Portugal
42. López V, Eisman E, Castro J, Zurita J (2011) A case based reasoning model for multilingual language generation in dialogues. Expert Syst Appl 39(8):7330–7337
43. López-Cózar R, Callejas Z, McTear M (2006) Testing the performance of spoken dialogue systems by means of an artificially simulated user. Artif Intell Rev 26:291–323
44. López-Cózar R, la Torre AD, Segura J, Rubio A, Sánchez V (2003) Assessment of dialogue systems by means of a new simulation technique. Speech Commun 40(3):387–407
45. López-Cózar R, Callejas Z (2008) ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. Comput Speech Lang 50(8–9):745–766

46. López-Cózar R, Callejas Z, Griol D (2010) ASR post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. Knowl-Based Syst 23(5):471–485
47. Mayer-Schonberger V (2003) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan-Houghton Mifflin Harcourt
48. McTear MF, Callejas Z, Griol D (2016) The conversational interface. Springer
49. Meng HH, Wai C, Pieraccini R (2003) The use of belief networks for mixed-initiative dialog modeling. IEEE Trans Speech Audio Process 11(6):757–773
50. Minker W (1999) Design considerations for knowledge source representations of a stochastically-based natural language understanding component. Speech Commun 28(2):141–154
51. Minker W, Waibel A, Mariani J (1999) Stochastically-based semantic analysis. Kluwer Academic Publishers, Dordrecht (Holland)
52. Mller S, Englert R, Engelbrecht K, Hafner V, Jameson A, Oulasvirta A, Raake A, Reithinger N (2006) MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In: Proceedings of the Interspeech, pp 1786–1789
53. Najafabadi M, Villanustre F, Khoshgoftaar T, Seliya N, WaldEmail R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1)
54. Oh AH, Rudnicky AI (2000) Stochastic language generation for spoken dialogue systems. In: Proceedings of ANLP/NAACL workshop on conversational systems, pp 27–32
55. O'Shaughnessy D (2008) Automatic speech recognition: history, methods and challenges. Pattern Recogn 41(10):2965–2979
56. O'Shea J, Bandar Z, Crockett K (2012) A multi-classifier approach to dialogue act classification using function words. Lecture notes in computer science, vol 7270, pp 119–143
57. Paek T, Horvitz E (2000) Conversation as action under uncertainty. In: Proceedings of the 16th conference on uncertainty in artificial intelligence, pp 455–464
58. Paek T, Pieraccini R (2008) Automating spoken dialogue management design using machine learning: an industry perspective. Speech Commun 50(8–9):716–729
59. Pieraccini R (2012) The voice in the machine: building computers that understand speech. MIT Press
60. Planells J, Hurtado L, Sanchis E, Segarra E (2012) An online generated transducer to increase dialog manager coverage. In: Proceedings of the international conference on spoken language processing (Interspeech'2012)
61. Rabiner L, Juang B, Lee C (1996) An overview of automatic speech recognition. Kluwer Academic Publishers, pp 1–30
62. Rojas-Barahona L, Giorgino T (2009) Adaptable dialog architecture and runtime engine (adarte): a framework for rapid prototyping of health dialog systems. Int J Med Inf 78:56–68
63. Roy N, Pineau J, Thrun S (2000) Spoken dialogue management using probabilistic reasoning. In: Proceedings of the 38th Annual meeting of the association for computational linguistics (ACL'00), pp 93–100
64. Schatzmann J, Thomson B, Weilhammer K, Ye H, Young S (2007) Agenda-based user simulation for bootstrapping a POMDP dialogue system. In: Proceedings of HLT/NAACL, pp 149–152
65. Schatzmann J, Thomson B, Young S (2007) Statistical user simulation with a hidden agenda. In: Proceedings of SIGdial, pp 273–282
66. Schatzmann J, Weilhammer K, Stuttle M, Young S (2006) A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowl Eng Rev 21(2):97–126
67. Segarra E et al (2002) Extracting semantic information through automatic learning techniques. Int J Pattern Recogn Artif Intell 16(3):301–307
68. Seide F, Li G, Yu D (2011) Conversational speech transcription using context-dependent deep neural networks. In: Proceedings of the 12th annual conference of the international speech communication association (InterSpeech 2011), pp 437–440. Florence, Italy

69. Shamim-Hossain M, Muhammad G, Alhamid MF, Song B, Al-Mutib K (2016) Audio-visual emotion recognition using big data towards 5G. Mobile Netw Appl 1:1–11
70. Singh S, Kearns M, Litman D, Walker M (1999) Reinforcement learning for spoken dialogue systems. In: Proceedings of neural information processing systems (NIPS'99), pp 956–962
71. Singh S, Litman D, Kearns M, Walker M (2002) Optimizing dialogue management with reinforcement leaning: experiments with the NJFun system. J Artif Intell 16:105–133
72. Suendermann D, Pieraccini R (2012) One year of contender: what have we learned about assessing and tuning industrial spoken dialog systems? In: Proceedings of NAACL-HLT workshop on future directions and needs in the spoken dialog community: tools and data (SDCTD'12), pp 45–48
73. Thomson B, Schatzmann J, Weilhammer K, Ye H, Young S (2007) Training a real-world POMDP-based Dialog System. In: Proceedings of NAACL-HLT-Dialog'07 workshop on bridging the gap: academic and industrial research in dialog technologies, pp 9–16
74. Torres F, Sanchis E, Segarra E (2003) Development of a stochastic dialog manager driven by semantics. In: Proceedings of European conference on speech communications and technology (Eurospeech'03), pp 605–608
75. Torres F, Sanchis E, Segarra E (2008) User simulation in a stochastic dialog system. Comput Speech Lang 22:230–255
76. Torres F, Sanchis E, Segarra E (2008) User simulation in a stochastic dialog system. Comput Speech Lang 22(3):230–255
77. Traum D, Larsson S (2003) The information state approach to dialogue management. Kluwer, pp 325–353
78. Tsilfidis A, Mporas I, Mourjopoulos J, Fakotakis N (2013) Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. Comput Speech Lang 27(1):380–395
79. Venkataraman A, Stolcke A, Shriberg E (2002) Automatic dialog act labeling with minimal supervision. In: Proceedings of the 9th Australian international conference on speech science & technology
80. Vipperla R, Wolters M, Renals S (2012) Spoken dialogue interfaces for older people. IOS Press, pp 118–137
81. Wilks Y, Catizone R, Worgan S, Turunen M (2011) Some background on dialogue management and conversational speech for dialogue systems. Comput Speech Lang 25:128–139
82. Williams J, Poupart P, Young S (2006) Partially Observable Markov decision processes with continuous observations for dialogue management. Springer, pp 191–217
83. Williams J, Young S (2007) Partially observable Markov decision processes for spoken dialog systems. Comput Speech Lang 21(2):393–422
84. Williams J (2009) The best of both worlds: unifying conventional dialog systems and pomdps. In: Proceedings of Interspeech, pp 1173–1176
85. Wu WL, Lu RZ, Duan JY, Liu H, Gao F, Chen YQ (2010) Spoken language understanding using weakly supervised learning. Comput Speech Lang 24(2):358–382
86. Young S (2002) The statistical approach to the design of spoken dialogue systems. Technical report, CUED/F-INFENG/TR.433, Cambridge University Engineering Department, Cambridge, UK
87. Young S, Gasic M, Thomson B, Williams J (2013) Pomdp-based statistical spoken dialogue systems: a review. In: Proceedings of the IEEE, pp 1–18, Montreal, Canada
88. Young S, Williams J, Schatzmann J, Stuttle M, Weilhammer K (2005) The hidden information state approach to dialogue management. Technical report, Department of Engineering, University of Cambridge, Cambridge, UK
89. Young S, Schatzmann J, Weilhammer K, Ye H (2007) The hidden information state approach to dialogue management. In: Proceedings of the 32nd IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 149–152

Big Data Analytics in Telemedicine: A Role of Medical Image Compression

Vinayak K. Bairagi

Abstract Big data analytics which is one of most rapidly expanding field has started to play a vital role in the field of healthcare. A major goal of telemedicine is to eliminate unnecessary travelling of patients and their escorts. Data acquisition, data storage, data display and processing, and data transfer represent the basis of telemedicine. Telemedicine hinges on transfer of text, reports, voice, images and video, between geographically separated locations. Out of these, the simplest and easiest is through text, as it is quick and simple to use, since sending text requires very little bandwidth. The problem with images and videos is that they require a large amount of bandwidth, for transmission and reception. Therefore, there is a need to reduce the size of the image that is to be sent or received i.e. data compression is necessary. This chapter deals with employing prediction as a method for compression of biomedical images. The approach presented in this chapter offers great potential in complete lossless compression of the medical image under consideration, without degrading the diagnostic ability of the image.

Keywords Medical image compression · Predictor based compression · Lossless compression · Volumetric image compression · Telemedicine

1 Telemedicine

Telemedicine is a method, by which patients can be examined, investigated, monitored and treated, with the doctor and patient located at different places. In telemedicine one transfers the expertise, not the patient. Diagnostic details and treatment is made mobile, without disturbing the patient. The World Health Organization defines telemedicine as "The delivery of healthcare services, where distance is a critical factor, by all healthcare professionals using information and communication technologies for the exchange of valid information for diagnosis,

V.K. Bairagi (✉)
E&TC Department, AISSMS-Institute of Information Technology, Pune, India
e-mail: vbairagi@yahoo.co.in

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_7

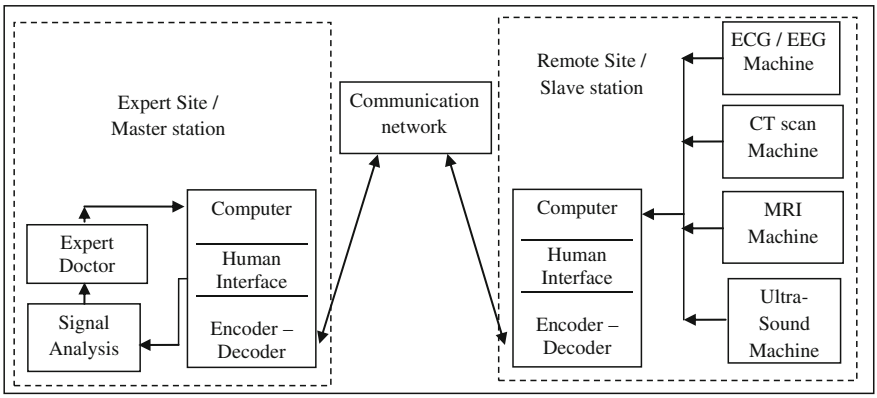


Fig. 1 Schematic view of telemedicine system [5]

treatment and prevention of disease and injuries, research and evaluation, and for the continuing education of healthcare providers, all in the interests of advancing the health of individuals and their communities" [1].

Studies affirm that, due to inexperienced medical help and poor medical facilities in the rural areas, the rural population afflicted with the same disease face twice the risk of death as their urban counterparts [2]. The government and private sectors had taken many initiatives, but still, the rural and remote areas continue to suffer due to the absence of quality healthcare. Telemedicine attempts to narrow the gap underlying urban and rural counterparts, in terms of quality health care. A major goal of telemedicine is to eliminate unnecessary travelling of patients and their escorts. Data acquisition, data storage, data display and processing, and data transfer represent the basis of telemedicine. Telemedicine is becoming an integral part of health-care services in several countries [3, 4].

Figure 1 shows the schematic view of a telemedicine system in which master station and remote area slave station are connected by a wireless link.

1.1 Types of Telemedicine

Depending on the requirement and connectivity available, two types of telemedicine systems can be possible, namely real-time and store forward technique.

1.1.1 Real-Time

In a real-time telemedicine clinic, a patient is seen in real time via an interactive video-conference. It involves both the parties, i.e. the patient and medical experts, at

the same time and a communication link which will allow real-time interaction between them. In this type of interaction, depending on the location of the clinic and the on-site personnel available, the patient may be accompanied by a medical assistant, a nurse, a primary care physician, or in some instances by a tele-health site coordinator, without medical experience or certification. A real-time system can be as simple as a telephonic call or as complex as robotic surgery [6].

The benefits of real-time telemedicine are:

1. It is similar in nature to an "in-person" visit;
2. Direct patient education and medical information can be delivered by the doctor.

1.1.2 Store-and-Forward

Store-and-forward (SF) applications use the same technologies, examining peripherals and clinical protocols, to acquire data regarding the patient. It does not require both the parties at same time. The pertinent information is stored in a specific format and sent to a consultation provider for diagnosis, interpretation, confirmatory opinion, second opinion, or for any reason that the input of the consulting provider is requested. SF requests can be as simple as a question posed in an email or as complex as a multi-media file containing narrative history and physical examination data, digital pictures, X-Rays, streaming video clips, and other imaging data sent electronically to a consultation provider [6].

1.2 Applications of Telemedicine

The various types of possible applications of Telemedicine are [7]:

- Disease Surveillance
- Disaster and Disease management
- Remote Consultation
- Second opinion
- Telementored procedures
- Home care
- Medical education and public awareness

1.3 Use of ICT in Medical Field

Today the use of computers for handling image data in the healthcare field is growing. The use of computers and a network makes it possible to distribute the image data among the staff efficiently [8]. The Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scan are modern image generating techniques. ECG produces one dimensional data whereas, MR and CT produce sequences of

images (image stacks), each a cross-section of an object [9, 10]. Large amount of image data is produced in the field of medical imaging in the form of CT, MRI, Positron Emission Tomography (PET) and Ultrasound Images, which can be stored in Picture Archiving and Communication System (PACS) or hospital information system [11]. American College of Radiology (ACR) recommendations for large matrix size radiographs (including those from digital radiography and digitized radiographic films, i.e., scanned X-rays) provides a minimum of 2.51 pixel/mm special resolution (approximately 4.0 megapixels) at a minimum of 10-bit pixel depth [12, 13]. A medium scale hospital with the above facilities produces on an average 5–15 GB of data per day [14–16]. Managing the storing facilities for the same can prove to be cumbersome for the hospitals. Moreover, such high data demands high-end network capabilities, especially for transmitting it over the network, such as in telemedicine. This is an essential requirement in the field of telemedicine, due to limitations of bandwidth in transmission media for Information and Communication Technology (ICT), especially for rural area [14].

2 Telemedicine in Rural Area

2.1 Application to Rural Areas

Rural areas have many difficulties which are inherent in nature, in addition to those experienced by populated urban areas. Some of the more obvious ones are as follows:

1. Part time staff: Many times because of less population concentration at rural parts, number of patients coming at clinic is less. Because of which doctors visits such clinics 1 or 2 days per week. This causes delay to some patients, not forgetting the doctor's time that is spent whilst travelling.
2. Few local specialists: In smaller hospitals of rural areas, it is not feasible or economic to employ many specialist consultants. It can take days or weeks for diagnosis to be accurately determined.
3. When it is necessary to refer a patient to another hospital, the relevant images can be made available prior to his/her arrival [17].

2.2 Connectivity and Bandwidth in Rural Area

The communication channel plays a vital role in telemedicine. If the channel bandwidth is not sufficient for handling the signal bandwidth, the signal will get distorted at the receiving end and there will be loss of information. Hence, bandwidth of the communication link is a key technical feature determining its usage for data transmission. For ECG, the signal bandwidth is approximately 100 Hz. The minimum sampling rate for digitizing signal is 200 samples per seconds and if one

Table 1 Transmission bandwidth requirement for various data [16], [18]

Sr. no	Type of multimedia data	Bandwidth required for transmission
1	Usual data	100 bps–2 kbps
2	Voice	4 kbps–150 kbps
3	Image	40 kbps–150 kbps
4	3D medical images	6 Mbps–120 Mbps

maps the analog values (in millivolts) to 256 discrete levels, it requires 8 bits per sample to represent the data. Subsequently, the minimum data rate at which ECG signal could be acquired is 1.6 kbps for one channel and 19.2 kbps will required for 12 channel ECG. Bandwidth requirement (expressed in terms of bits per second) of different multimedia data is shown in Table 1.

Traditional phone service sometimes called POTS for "plain old telephone service" with maximum speed of 56 kbps is not an optimal solution for transfer of images because of its connectivity issues. In rural areas, connectivity with expert stations can be done with ISDN, leased line or Very Small Aperture Terminal (VSAT). For Telehealth services a bandwidth of 384 kbps or higher is required [1, 19]. Leased line (~1 Mbit/s) is not suitable for the rural sector where, the main city switching center is far away. For such cases, satellite based VSAT connectivity (~512 kbps) is the best option. For a VSAT connection, the running cost is very high which can be minimized by governmental support (In India, DIT with ISRO are supporting this activity).

2.3 Role of Compression in Telemedicine

In rural areas, where connectivity is a major concern, store and forward technique is preferable. Transmitting hundreds of images daily (~80 MB) would be a very time-consuming process. Data compression is a viable solution in this scenario. Data compression is the process of converting an input file into another file having smaller size. Telemedicine hinges on transfer of text, reports, voice, images and video, between geographically separated locations. Out of these, the simplest and easiest is through text, as it is quick and simple to use, since sending text requires very little bandwidth. The problem with images and videos is that they require a large amount of bandwidth, for transmission and reception [20]. Therefore, there is a need to reduce the size of the image that is to be sent or received i.e. data compression is necessary. The aim of data compression techniques is to reduce the amount of data needed to accurately represent an image, such that this data can be economically transmitted or archived [21]. The medical image data mainly includes MRI scan, CT scan, Ultra-Sound images and Computed Radiography (*CR*). The primary health setup would include an X-Ray unit and Pathological Unit [18, 6].

As an example, a certain sequence of MRI, known as magnetic resonance cholangiopancreatography (MRCP) for the pancreatobiliary structures in the liver

and abdominal regions, will produce 100–200 images per patient during a single examination. With just 50 examinations a day, over 250 GB of image data would be generated per year [22].

Image compression is useful in reducing the storage and transmission bandwidth requirements of medical images [14]. To make this fact clear, let's see an example. An image, 1024 pixel × 1024 pixel × 24 bit, without compression, would require 3 MB of storage and 7 min for transmission, utilizing a high speed, 64 kbit/s, ISDN line. If the image is compressed at a 10:1 compression, the storage requirement is reduced to 300 KB and the transmission time drops to under 6 s [23, 24]. Therefore, there is a need to decrease the size of the image that is to be sent or received. Increasing the bandwidth is another method, but the cost involved to achieve this, sometimes makes this a less attractive solution. Also, the number of channels gets reduced if bandwidth is increased. By shrinking the size of the image, fewer pixels need to be stored and consequently, the file will take less time to load.

A generalized block diagram of any communication system where compression is used is shown in Fig. 2. The data having very less meaning or the data by removing which the quality of image will not be affected much is removed from the image. Then, the compression coding is applied over the image. Different algorithms such as Run Length Encoding (RLE), Lempel Ziv Welch (LZW) Compressors, Huffman Encoding, DCT, JPEG, JPEG 2000 are widely used for compression [14]. The compressed image data is then transmitted over the channel. At the receiving station, exactly the reverse procedure as that of the transmitting station is carried out. First, the decoding algorithm is used and the received data is then reconstructed to form the original image [25].

Compression methods are classified into lossless and lossy methods [21]. In the medical imaging scenario, lossy compression schemes, despite having up to 10 % compression ratios, are not generally used. This is due to possible loss of useful clinical information which may influence diagnosis. In addition to these reasons, the legal issues may arise. Storage of medical images is characterised by the need to

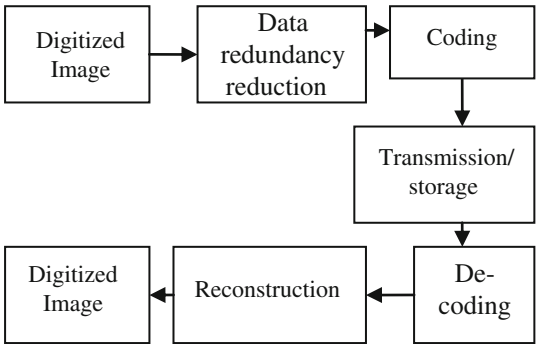


Fig. 2 A generalized block diagram of any image compression technique

preserve the best possible image quality which is usually interpreted as a requirement for lossless compression [23, 24]. 3D MRI containing multiple slices representing a part of the body requires all information of that part. The storage size for such 3D images is huge.

3 DICOM Format

Different file formats are available like BMP, TIFF and PNG etc. which are widely used in the field of image processing. Each of these has special properties. Medical Imaging uses a special kind of file format to represent data, called Digital Imaging and Communications in Medicine (DICOM) format. It was developed by the National Electrical Manufacturers Association (NEMA) in conjunction with the ACR. It covers most image formats for all of medicine.

DICOM differs from other data formats in that it groups information into data sets. That means that a file of a chest X-Ray image, for example, actually contains the patient ID within the file, so that the image can never be separated from this information by mistake. DICOM is the most comprehensive and accepted version of an imaging communications standard. DICOM format has a header which contains the information about the image, imaging modality and information about patient [1]. One must preserve header of DICOM files. This is because of the legal questions raised and the regulatory policies set by agencies such as the Food and Drug Administration (FDA). Additionally, DICOM defined Information Objects not only for images but also for patients, studies, reports, and other data groupings. To compress such DICOM files, special attention should be given to header information.

DICOM format contains the stack of images along with header carrying some text information as shown in Fig. 3. During compression, there is need to carefully consider header information; hence hybrid combination of image compression, with text compression is required as shown in Fig. 4.

A lot of data storage is required when it comes to medical images [26]. The memory requirements table for the mid-size or large clinic is surveyed. Every day it produces nearly 65 GB of data. It is seen that the storage requirements increase tremendously as the instruments as well as the number of patients are more in count.

Some of the most desirable properties of any medical image compression method include: (1) fidelity Criteria, (2) high lossless compression ratios, (3) resolution scalability, which refers to the ability to decode the compressed image data at various resolutions, and (4) Quality scalability, which refers to the ability to decode the compressed image at various qualities up to lossless reconstruction [27].

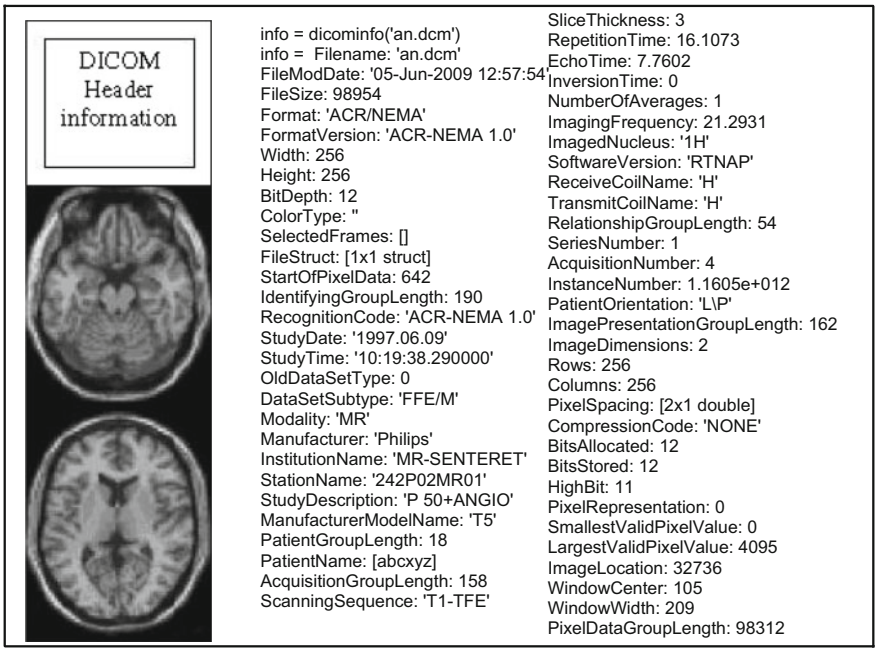


Fig. 3 Typical DICOM file along with header information

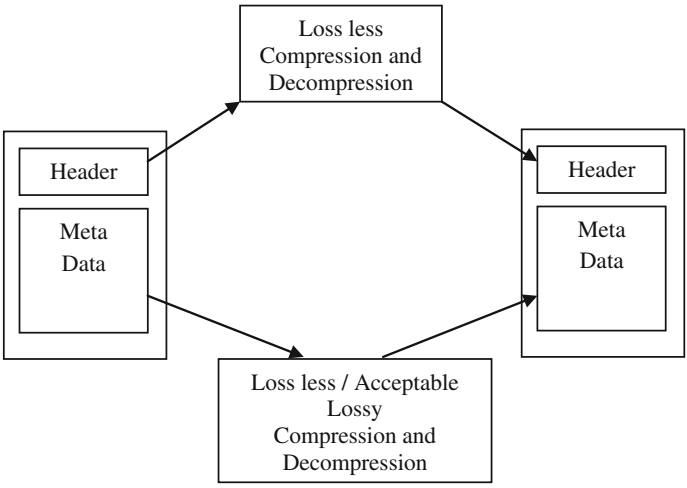


Fig. 4 DICOM file compression system

4 Image Compression

This section deals with employing prediction as a method for compression of biomedical images. Image compression algorithms can be broadly classified as Entropy based compressors (e.g. Huffman coding), Transform based compressors and Prediction based compressors. Many a times for natural images special redundancy is main type of redundancy. Transformation-based and prediction-based approaches are used to remove special redundancy. Researchers have proved that single algorithm alone will not give high compression but in combination with other algorithms it will give an improved result [20]. Most of time entropy based compressors are used after transform based or prediction based compressors in cascaded way.

Prediction-based techniques use a simple assumption that pixel value can be partially or totally represented as a linear combination of neighbour pixels. Transformation-based techniques transform data in to transform domain, which allows better exploitation of spectral components, which are present in image. The transform may required complex mathematical operation. For lossless image compression, prediction based algorithms are given preferences before transform based algorithms, because they are simple, fast, and most important predictors can guarantee a lossless data recovery [28], [21].

5 Static Predictors

Prediction removes most of the spatial redundancy, and the choice of an optimal predictor is essential for improving the efficiency of compression methods [29, 21]. It is assumed that an image to be encoded is scanned using a raster-scan technique that is the pixels are scanned from left to right, top to bottom. Therefore as shown in Fig. 5, at any time instance 't' (close prediction) of encoding it is assumed that all the previous pixels W, WW, N, NE, and NW etc. have been scanned and encoded. Similar is the case at decoder side. For static predictor, the prediction condition is always fixed [28]–[30].

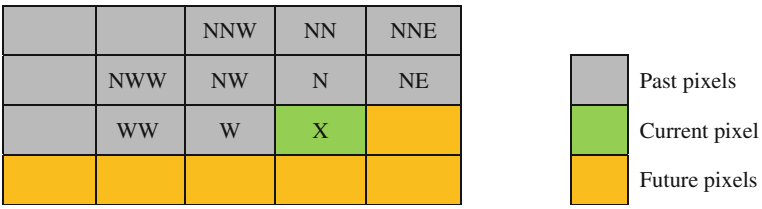


Fig. 5 Locations of other pixels relative to the current pixel X [31]

5.1 Differential Coding

The simplest prediction coding is differential coding, where difference between the two pixels is saved in the form of error values. Here predicted value of 'X_P' is assumed to be 'W' for encoding, then only the error value 'e' [=(Xa − Xp)] is preserved for decoding processes as shown in Fig. 6. In the process of prediction based encoding, the encoder saves only error value instead of actual pixel values, hence low bits are required to save error (assuming that prediction is more closer to actual value and error is small). The success of predictor depends on closeness of prediction with actual value [31, 32].

Differential coding prediction works well on smooth or continuously varying images. But the prediction error is high at the edge of image. Hence if there is large variation in image it will not be efficient. For image having large variation in pixel intensity, closer prediction is more effective. Here the image that is transmitted over communication channel will be error image, which have small size compared to original image. Variable length coding is suitable in case of entropy coding after prediction.

5.2 Lossless JPEG

To meet its requirement for a lossless mode of operation, of JPEG algorithm, Joint Photographic Experts Group (JPEG), has chosen a simple predictive method which

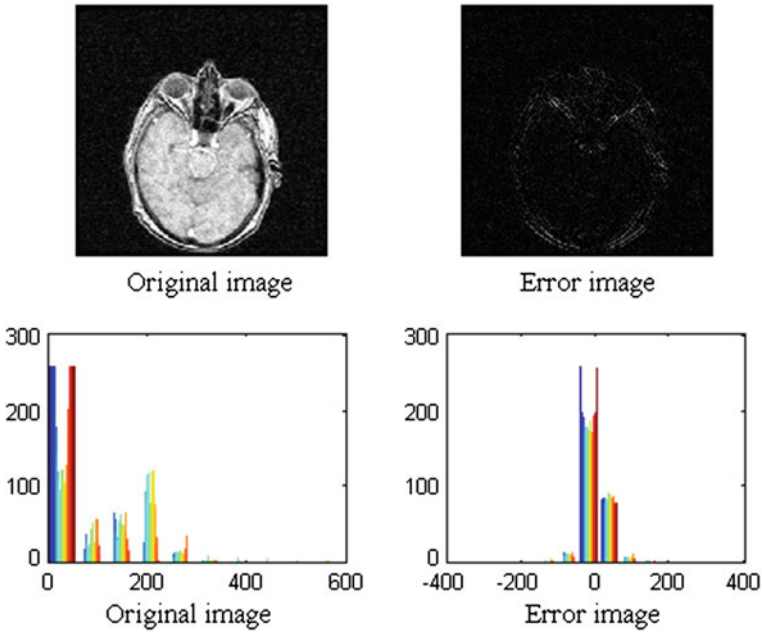


Fig. 6 Result of static prediction with condition $X = W \begin{matrix} a, b \\ c, d \end{matrix}$

Table 2 Predictor conditions for lossless JPEG coding [33, 34]

Selection value	Prediction condition
0	No prediction
1	W
2	N
3	NW
4	W + N − NW
5	W + ((N − NW)/2)
6	N + ((W − NW)/2)
7	(W + N)/2

is wholly independent of the DCT processing. It consists of two blocks in cascaded manner. Former is predictive coding followed by entropy coding [33]. A predictor combines the values of up to three neighbouring samples (W, N and NW) to form a prediction of the sample indicated by X in Fig. 5. Table 2 represents the selection condition with their predictors. Selections 1, 2, and 3 are one-dimensional predictors and selections 4, 5, 6 and 7 are two-dimensional predictors. For given "selection-value" Any one of the eight predictors listed in Table 2 can be used. Entropy coder can be any one of Huffman or Arithmetic coding.

5.3 EDPCM

New transform named Enhanced DPCM Transformation (EDT) is proposed by Farshid Sepehrband and his team. The method uses improved version prediction scheme of Lossless JPEG [35]. In EDT, redundancy reduction is improved and complexity is less as compared with JPEG-LS.

Pixel values in image are divided by numbers like 2, 3 or 4. Due to which pixel intensity values in image are separated in two numbers; where one is quotient and second is remainder as shown in Fig. 7. In other words the small variation in image is separated; values in quotient are closely related.

Then, quotients of division are predicted by one of the prediction equations of Fig. 8, the neighbouring pixels position and prediction equations.

The predicted image is again up-scaled by same factor as that of at the time of division. The remainder is added with predicted part to get final corrected predicted image. The error image is generated by subtracting predicted image from original image. Huffman coding is used as a second level compressor.

The experiments and their validations are carried out using MATLAB as a software platform. The hardware computing platform used for experiments is Intel C2D processor, 2 GHz with 3 GB RAM. The images are grey scale with resolution of 8 bit and 16 bit ranging from size 200 × 200 till 1024 × 1024 (pixel × pixel). Input image is in DICOM format. The input image database consists of more than 1800 images available from private domain database as well as from public domain images [36, 30, 37]. The sample test image data base is shown in Fig. 9.

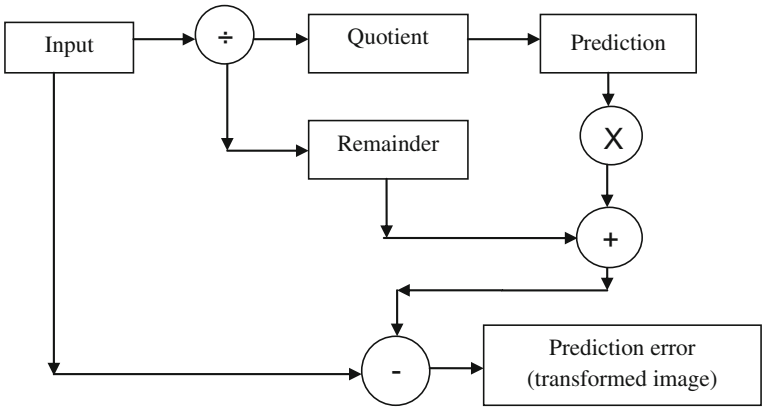


Fig. 7 Block diagram of EDT [35]

A	B	C
D	X	

1. X=(B+D)/2	4. X=D	7. X=B+(D-A)/2
2. X=B+D-A	5. X=B	8. X=(A+B+C+D)/4
3. X=D+(B-A)/2	6. X=A	9. X=(B+C)/4 + D/2

Fig. 8 Neighbouring pixels and prediction equations used in EDT [31, 33]

Observations

- From the Table 3 it is seen that out of the nine prediction conditions, second and ninth prediction conditions gives good results as compared with other condition for most of the images. The method is analogous with removing of the LSB from pixel values in image.
- It is observed that, the prediction at particular pixel position is away from original value if their exist edge in image.
- Also the negative values of error are another problem, because one extra bit is required to store that sign bit.
- The pixel intensity values in original image are divided by factor of 2, due to which the Quotient information is closely related because of removal of high variation information. The prediction conditions will produce less error values if the pixels are closely related.

To overcome the drawbacks of static predictor, dynamic predictors are preferred. Now we will discuss dynamic predictors.

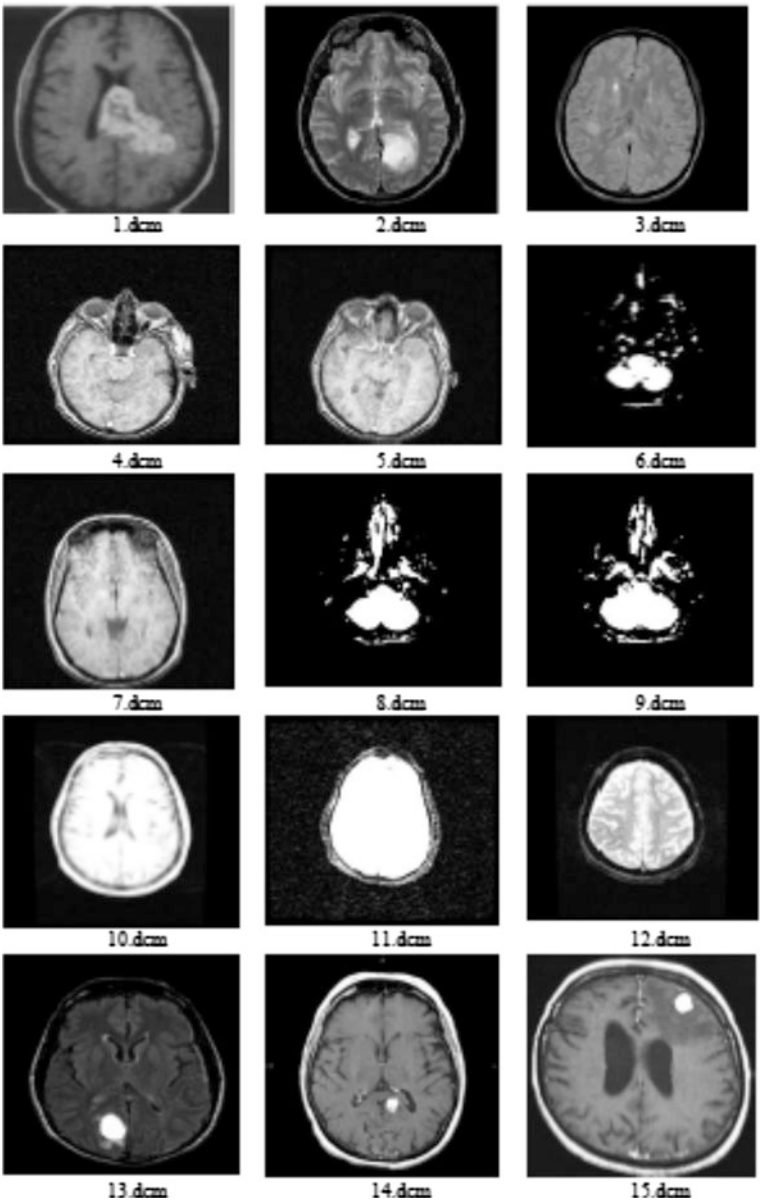


Fig. 9 Sample images in database

Table 3 Compression ratio result table for different conditions of EDPCM

Filename	Conditions								
	1	2	3	4	5	6	**7**	8	9
1.dcm	0.24693	***0.21097***	0.24159	0.27585	0.23090	0.29043	0.22022	0.24873	0.24018
2.dcm	0.21162	0.20427	0.21241	0.22744	0.20988	0.23671	***0.20295***	0.21809	0.21303
3.dcm	0.17272	0.17291	0.17421	0.17708	***0.1709***	0.18748	0.17143	0.177	0.17304
4.dcm	0.36131	0.38129	0.36860	0.37466	0.38133	0.39332	0.39608	0.3643	***0.35933***
5.dcm	0.3553	0.37435	0.36144	0.36701	0.37378	0.38850	0.38872	0.35799	***0.35235***
6.dcm	0.05846	0.05874	0.05981	***0.05563***	0.05666	0.05971	0.06062	0.06234	0.06021
7.dcm	0.35373	0.37163	0.35946	0.36557	0.37254	0.38769	0.38661	0.35671	0.3514
8.dcm	0.08494	0.08455	0.08618	***0.08167***	0.08197	0.08722	0.08563	0.09135	0.0885
9.dcm	0.09572	0.09450	0.09660	0.09265	***0.09251***	0.09822	0.09606	0.10204	0.09912
10.dcm	0.23872	***0.2319***	0.23762	0.25579	0.24626	0.27196	0.23246	0.24284	0.2361
11.dcm	0.36448	0.35856	0.36181	0.38546	0.37833	0.43758	0.35827	0.36873	***0.35808***
12.dcm	0.21835	***0.2133***	0.21558	0.23300	0.23152	0.25266	0.2160	0.22289	0.21481
13.dcm	0.19217	***0.18229***	0.19042	0.20763	0.19386	0.21884	0.18362	0.19541	0.18972
14.dcm	0.21699	0.21629	0.21828	0.23344	0.22134	0.24632	***0.21327***	0.2212	0.21649
15.dcm	0.30116	***0.29205***	0.2994	0.32559	0.31005	0.34533	0.29469	0.3058	0.29745

Bold and Italic indicates lowest value in particular Row

6 Dynamic Predictor

6.1 Median Edge Detector

Median Edge Detector (MED) takes into account N, and NW pixel along with W. In dynamic predictor the prediction condition is dependent on past pixels [38].

$$X[i,j] = \begin{cases} \min(\mathrm{N},\mathrm{W}), & \text{if } \mathrm{NW} \geq \max(\mathrm{N},\mathrm{W}) \\ \max(\mathrm{N},\mathrm{W}), & \text{if } \mathrm{NW} \leq \min(\mathrm{N},\mathrm{W}) \\ \mathrm{N}+\mathrm{W}-\mathrm{NW}, & \text{Otherwise} \end{cases} \quad (1)$$

MED is illustrated below

(1) Original image (X) =	102	49	22	15
	19	***19***	26	17
	20	20	15	12
	14	12	13	11
(2) Predicted image =	168	102	49	22
	25	19	19	19
	14	20	26	15
	20	14	12	12
(3) Error image (Y) = (Original image – Predicted image)	−66	−53	−27	−7
	−6	0	7	−2
	6	0	−11	−3
	−6	−2	1	−1

Bold and Italic indicates lowest value in particular Row

The predicted image using MED prediction is shown in Fig. 10b. The error image obtained after subtraction of predicted image from original image is shown in Fig. 10c. The respective histograms of all images in Fig. 10 are shown in Fig. 11.

Observations

- The original and predicted images are similar up to certain extent, and it can be seen in Fig. 10b and consecutively in Fig. 11b. One needs to store only error image for decoding purpose.
- In error image most of values are concentrating near to zero as seen in Fig. 11c. Thus instead of storing many higher values (near to 255 for 8 bit image), one can store many lesser (near to zero) by using predictors.
- The performance of MED is much higher than static predictor like differential coding, because of its dynamicness (the predicted pixel values are dependent on more than one conditions).
- Practically MED cannot give same performance for different types of edges like strong and soft.

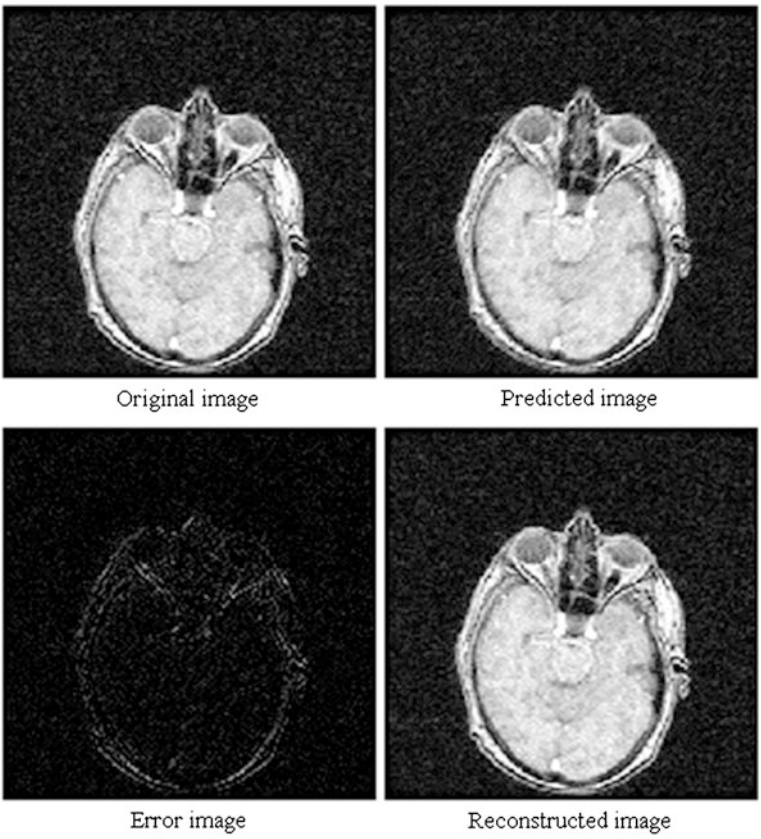


Fig. 10 Result of MED predictor $\begin{matrix} a, b \\ c, d \end{matrix}$

To overcome this difficulty, Gradient Adjusted Predictor (GAP) based on gradient estimation around the current pixel [39].

6.2 Gradient Adjusted Predictor

GAP distinguishes three types of edges, strong, simple and a soft edge, and is characterized by high flexibility to different regions. Gradient estimation is done using [38, 39]:

$$\begin{aligned} dh &= |w - ww| + |n - nw| + |n - ne| \\ dv &= |w - nw| + |n - nn| + |ne - nne| \end{aligned} \tag{2}$$

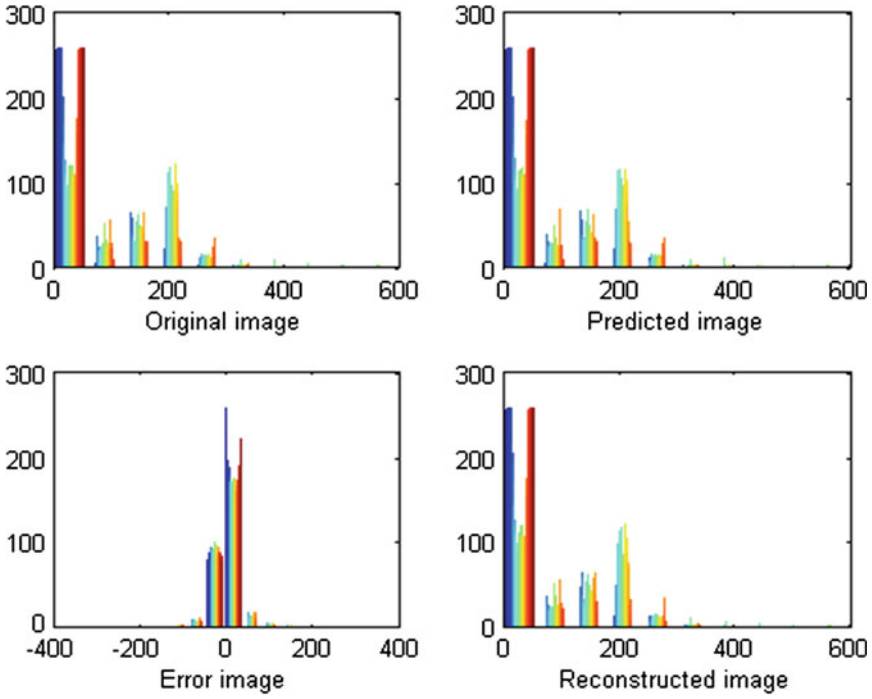


Fig. 11 Probability distribution of pixel intensity in MED predictor at different stages $\begin{matrix} a,b \\ c,d \end{matrix}$

and the prediction is made by the algorithm:

$$
\begin{aligned}
&if\ dv - dh > 80, p = w \\
&elseif\ dv - dh < -80, p = n \\
&else\{\ \ p = [(w + n) \div 2] + [(ne - nw) \div 4] \\
&\qquad if\ dv - dh > 32, p = (p + w) \div 2 \\
&\qquad else\ if\ dv - dh > 8, p = (3p + w) \div 4 \\
&\qquad else\ if\ dv - dh < -32, p = (3p + w) \div 4 \\
&\qquad else\ if\ dv - dh < -8, p = (3p + n) \div 4 \\
&\}
\end{aligned}
\tag{3}
$$

GAP is improved version of MED prediction, where predictor takes care of edges in side image. This extra modelling gives GAP better performance than MED, although typically not by a large margin. Unlike MED prediction is dependent on pixels values in last two column and rows [39].

6.3 Calic

Context-Based, Adaptive, Lossless Image Coding (Calic) [40] uses advance error mapping features which results in lower values of errors at the cost of more computational complexity. It has following major component.

1. Linear prediction: Calic uses GAP predictor to predict current pixel position.
2. Context selection and quantization followed by Context modelling of prediction errors depending on textural features.
3. Entropy coding of prediction errors for further compression.

The error re-modelling base on textural features makes the process more complicated. Calic is powerful predictor for continuous-tone image. It uses non-linear prediction followed by entropy coding. JPEG-LS is somewhat same in prediction as that of Calic but in second level of compression, it uses fast Golomb rice code. The computational complexity of JPEG-LS is much less than Calic, due to which it is preferred and has become more popular [41].

6.4 JPEG Lossless

The drawbacks of static predictors in lossless JPEG are removed by using Medial Edge Detector as a dynamic predictor in JPEG-LS. The algorithm first investigates whether current region to be coded is flat; if so it uses RLE code otherwise it uses MED predictor with error context modeller, followed by Golomb–Rice coder as an entropy coder at the end. The computational complexity of Golomb–Rice code is less as compared with arithmetic coder, hence it is preferred in JPEG-LS standard [42, 43].

Tables 4, 5 and 6 shows the comparison of Differential Coding, MED predictor and JPEG-LS standard for CT, X-Ray and MRI images from database respectively [36, 30, 37].

Observations

- From Tables 4, 5, 6, it is observed that the dynamic predictor (i.e. MED) works much better than static predictor like Differential Coding.
- From Table 4, it is observed that the dynamic predictor, MED works better (∼15 %) than JPEG–LS standards for most of images in database of CT image except for Chest images.
- For MRI and also for X-Ray images, the performance in terms of average compression ratio for MED and JPEG-LS is nearly same.

Table 4 Comparative analysis of prediction based compression techniques for CT images

Sr. no	Image description	No of images	Compression ratio (Avg)		
			DC	MED	JPEG-LS
1	CT abdomen	102	0.4508	***0.272***	0.3465
2	Angiography thoracic	19	0.1979	***0.1311***	0.22
3	Brain axial plan	25	0.3223	***0.1997***	0.339
4	Head brain bone	60	0.6024	***0.355***	0.4485
5	Head brain bone -plan	29	0.2967	***0.1911***	0.3274
6	Head brain	75	0.6344	***0.3751***	0.4781
7	Brain seg plan	30	0.4527	***0.2735***	0.3888
8	Dental scan	72	0.6758	***0.3915***	0.5274
9	Neck	30	0.3278	***0.2047***	0.3079
10	Spine	59	0.5668	***0.3407***	0.4274
11	Chest 4	71	0.4113	0.4134	***0.3558***
12	Chest 1	67	1.254	1.2316	***0.9855***
13	Inner ear	5	0.5512	***0.3382***	0.4612
14	Thorax	76	0.3688	***0.3797***	0.402
	Average		0.5081	***0.3640929***	0.429679

Bold and Italic indicates lowest value in particular Row

Table 5 Comparative analysis of prediction based compression techniques for x-ray images

Sr. no	Image description	Size		No of images	Compression ratio (Avg)		
		Row	Col		DC	MED	JPEG-LS
1	Angiography	1024	1024	46	0.5454	***0.3094***	0.3722
2	C. Spine with head	1755	1462	95	0.7627	***0.5502***	0.6111
3	Spine	2487	2048	51	0.6007	***0.4004***	0.4126
4	Chest	372	616	12	0.5402	0.4082	***0.3624***
5	Thoracic spine	372	616	15	0.5232	0.4072	***0.3481***
6	Spinal cord	372	616	21	0.5497	0.4192	***0.3727***
7	Lumber spine	528	828	19	0.5001	0.3461	***0.3327***
8	Lumber	528	828	27	0.5033	0.3561	***0.3354***
	Average				0.56566	0.3996	0.3934

Bold and Italic indicates lowest value in particular Row

6.5 Edge Enhanced Predictor

To address the issue of high error values at edges, new edge enhanced method is needed to be developed. A canny edge predictor is used to find out the edges within image. The edge matrix is maintained separately consisting of binary values. Depending on the '1' followed by '0' or vice versa a prediction condition is selected. After a complete performance evaluation for compression ratio, it is clear that the GAP and GED predictors give better result than MED and differential prediction. These two predictors are used in combination but using one at a time

Table 6 Comparative analysis of prediction based compression techniques for MRI images

Sr. no	Image description	Size		No of images	Compression ratio (Avg)		
		Row	Col		DC	MED	JPEG-LS
1	Knee	512	512	40	1.1577	1.1629	***0.9918***
2	Abdomen	256	256	45	0.532	***0.3086***	0.4174
3	Rectum	256	256	43	0.6637	***0.4103***	0.4966
4	Heart	256	256	45	0.5224	***0.3262***	0.3977
5	Pelvis	256	256	52	0.5986	***0.3847***	0.4966
6	Rectum 1	256	256	35	0.5434	***0.3244***	0.383
7	Rectum 2	256	256	30	0.5415	***0.3224***	0.3826
8	Pelvis hip	640	576	41	0.4821	***0.3108***	0.4142
9	MR knee T1	256	256	32	0.4389	0.3358	***0.3301***
10	Lumbar-axial	512	512	43	0.4899	0.5104	***0.41***
11	Lumbar-axial T2	512	512	46	0.4797	0.4906	***0.39***
12	Lumbar sagittal	512	512	32	0.5343	0.5223	***0.4316***
13	Lumbar-FOV	512	512	29	0.4046	0.3965	***0.344***
14	Wrist	512	512	28	0.4018	***0.2726***	0.3291
15	Brain-T2	256	256	41	0.4586	0.4474	***0.3301***
16	Brain-COR T1	256	256	35	0.5684	0.5637	***0.4172***
17	Brain-sagittal T2	256	256	35	0.5835	0.577	***0.43***
18	Spinal cord	512	512	33	0.5012	***0.328***	0.4288
19	Brain ven	256	256	40	0.4564	0.4859	***0.3648***
	Average				0.54519	0.446342105	0.430821053

Bold and Italic indicates lowest value in particular Row

and the selection of any one of them is dependent on the values in edge matrix. At the edges the GED is applied and for rest of the part in image GAP is applied.

Algorithm

Encoding:

1. Input the image.
2. Compute the edge matrix of 0's and 1's by using canny edge function.
3. Check the two pixel of edge matrix.
4. If the combination is '01' or '10' then apply GED prediction technique else GAP.
5. Calculate the error matrix by subtracting predicted value from original.
6. Save error matrix and edge matrix.

Observations

- The original and predicted images are similar to certain extent.
- One needs to store error image along with edge matrix for decoding purpose.
- From the result Table 7, we observed that edge enhanced predictor gives better result for images having smooth region as well as for images which have drastic

Table 7 Performance of edge enhanced predictor on sample images

Sr. no	Image name	Rows	Columns	Orig_size in B	CR	Encoding time
1	1.dcm	256	256	131984	0.259054	33.73438
2	2.dcm	256	256	131984	0.281511	31.29688
3	3.dcm	256	256	131984	0.225542	32.00000
4	4.dcm	256	256	140774	0.515976	32.45313
5	5.dcm	256	256	140774	0.505711	31.8125
6	6.dcm	256	256	134816	0.06563	31.1875
7	7.dcm	256	256	140758	0.501598	32.73438
8	8.dcm	256	256	134816	0.095241	30.65625
9	9.dcm	256	256	134816	0.106167	31.26563
10	10.dcm	512	512	534210	0.330241	519.2656

variation at edges which is possible because of maintaining separated records of edges.

- The only bottleneck of the system is that, edge matrix is also needs to be preserved for decoding purpose. This is as good as adding one extra bit in pixel.
- Due to preservation of edge matrix, the compression performance is limited.

To make efficient compression of image data multi-level processing is required. Hence most of the cases RLE, Huffman or Arithmetic coders are used as last stage of compression process. Again the problem of negative sign in the error values arises. Because these negative values are considered as a separate entity in entropy coding.

6.6 GAP with Positive Error Modelling

After GAP prediction, some of the predicted values are negative. To convert these negative values into positive one, error sign flipping is used. The working is explained as follows,

Error $= Xa - X_P$

where,

Xa original pixel value

X_P predicted pixel value

Suppose, for 3 bit image, range of Xa: 0 to 7. Let $Xa = 4$ and $X_P = 6$. error $= 4 - 6 = -2$.

$$\begin{aligned} \text{Now, range of error} &= (\text{Range of Xa}) - X_P \\ &= (0 \text{ to } 7) - (6) \end{aligned} \tag{4}$$

Range of error $= (0 - 6)$ to $(7 - 6)$ i.e. from -6 to 1.

All possible values of error in between −6 to 1 are

$$[-6\ -5\ -4\ -3\ -2\ -1\ 0\ 1]$$

Now Sort the values on both side of zero

$$0,\ 1,\ -1,\ 2,\ -2,\ 3,\ -3,\ 4,\ -4 \ldots \text{so on..}$$

After sorting pass, assign index values

	[0	1	−1	−2	−3	−4	−5	−6]
Error index	−	0	1	2	3	4	5	6	7

Thus, instead of passing −2 error value its index value i.e. 3 is stored for further operation.

Thus, negative values are mapped to positive values.

Figure 12 represents different stages of GAP predicted image with Positive error (GAPPE). The negative errors in Fig. 12c are mapped into positive error index values as shown in Fig. 12d; and its histogram in Fig. 13d. In error image most of values are concentrating near to zero but on both side of zero, as seen in Fig. 13c.

The negative values of errors are transferred to positive values. The spectrum of pixel values is on both the sides in error image, due to which sign bit was required to be saved as extra bit. But in positive error modelling, since all errors are positive there is no need of sign bit. In other words (1/n) compression is already achieved in this process (n is bit depth of pixel). It is also seen that, most of the intensity values in positive error image have low value, where as very small number of values are of high intensity. A variable length coding is effective in this case.

Observations

- Table 8 represents the compression ratio performance of the presented approach with other discussed algorithms. The presented approach is 21.68 % better JPEG-LS, 33.68 % better than Calic.
- The computational complexity of presented GAPPE algorithm is much reduced than JPEG 2000 as it works on simple principal of prediction, whereas JPEG 2000 uses complex EBCOT algorithm with bit level encoding.
- Average time required for encoding by CALIC is 74.13 s, whereas for presented GAPPE approach it is 60.93 s which is almost 17.80 % less as compared with CALIC. Both algorithms can be optimized for their time complexity by using machine level coding. The complex algorithm like JPEG 2000 when implemented for its optimal performance takes only 0.09 s as average time for CT images in database.

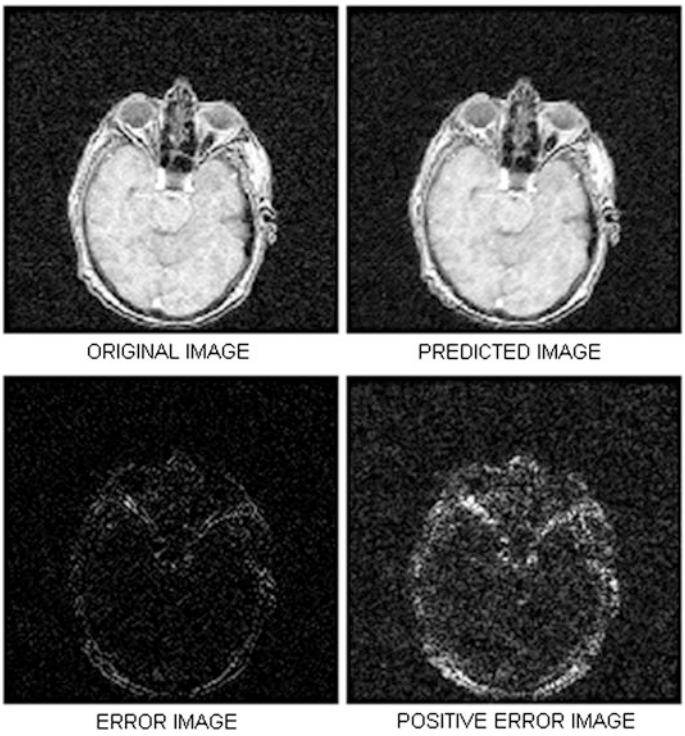


Fig. 12 Result of gap predicted image with positive error $\begin{matrix} a, b \\ c, d \end{matrix}$

- The presented GAPPE algorithm is tested for its quality of reconstruction; it is found that complete recovery is achieved. The two images original image under test and reconstructed image after decoding are identical. Hence there is no need to check the quality of image at reconstructed end with any further objective methods.
- The JPEG-LS works on the principle of advance error modeling along with context formation process which searches for repetition of context in previously encoded data. If such repetition is found, then present predicted pixel error is modelled according to past pixel error information in same block context. {Example for 8 pixel block if initial prediction error is −5, and if such 8 pixel block context is again repeated after some rows or columns, then the new error will be [(−5/2) = 2.5]} Due to this process even though the error values are going to decrease but at the time of entropy coding these new error values act as separate entity. In GAPPE no such context formation is done and hence error values are not corrected. This is useful at the time of entropy coding because of which we are getting compression.

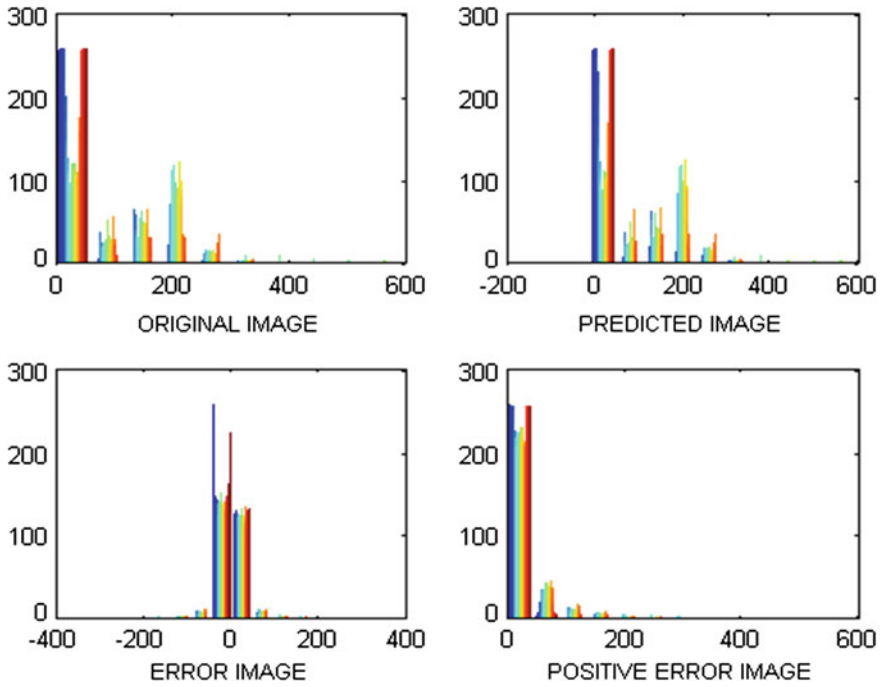


Fig. 13 Probability distribution of gap predicted image with positive error at different stages $\begin{matrix} a, b \\ c, d \end{matrix}$

Table 8 Performance of GAPPE predictor as compared with other compressors on sample images

Sr. no	Input image		Compression ratio of encoded data				
	Name	Size in B	MED	CALIC	JPEG LS	JPEG 2000-LS	GAPPE
1	1.dcm	131984	0.26706	0.26828	0.291422	0.242203	0.198676
2	2.dcm	131984	0.28298	0.26534	0.260782	0.185606	0.180923
3	3.dcm	131984	0.2012	0.16294	0.195865	0.188837	0.15023
4	4.dcm	140774	0.54251	0.46546	0.366509	0.344863	0.338393
5	5.dcm	140774	0.53333	0.45659	0.361324	0.336538	0.331006
6	6.dcm	134816	0.09256	0.13796	0.099729	0.083951	0.044705
7	7.dcm	140758	0.52802	0.45363	0.358758	0.334593	0.328656
8	8.dcm	134816	0.12034	0.1844	0.122218	0.111322	0.066157
9	9.dcm	134816	0.13131	0.1896	0.13149	0.119711	0.074902

7 Block Coding

Image can be segmented into blocks for efficient lossless compression. Redundancy can be found in blocks of pixels with maximum correlation. Venugopal et al., have proposes a block based lossless image compression algorithm using Hadamard transform and Huffman encoding [44]. The complexity can be minimised if use of transform is avoided.

7.1 Block Truncation Coding

Block Truncation Coding (BTC) method is a lossy compression technique, efficient for greyscale images. It divides the original image into the blocks and then applies the quantizer to reduce the number of gray levels in the block, maintaining the same mean and standard deviation. Using sub-blocks of 4 × 4 pixels gives a compression ratio of 25 % assuming 8-bit integer values are used during transmission or storage. Larger blocks allow greater compression; however quality reduces with the increase in block size, due to the nature of the algorithm [45].

So we have developed a different technique of block truncation coding on the image to achieve the lossless compression. The following steps describe the technique:

Algorithm:

a. **Encoder side**

1. The image is divided in blocks of 8 × 8 pixels.
2. These 8 × 8 pixel blocks are considered for texture blocking.
3. The pixel pattern of Nth block is compared with pixel pattern of its previous block i.e. (N − 1)th block.
4. A block subtraction is used, and only error between the subtraction is preserved in the place of Nth block.

At decoder side the reverse process is carried out as that of encoding. First the data from received packet is separated. Next a cell matrix is formed with each cell of 8 × 8 pixel size. Now Nth pixel block is re-constructed by combination of (N − 1)th block and the error data available at Nth place. The cell matrix is then converted to normal matrix form. This matrix represents the decoded image.

Observations

- The performance of developed lossless block code is shown in Table 9. It is verified that the block codes are fast in execution but the compression ratio is approximately 38 % of original for tested images of different sizes and modalities.
- It is observed that the developed method is completely lossless. The simplest method to judge the loss-less-ness is pixel wise comparison, and it is done for

Table 9 Performance of lossless block code

Sr. no	Original image				Block code	
	Name	M	N	Size in B	CR	Time
1	1.dcm	256	256	131984	0.463094	0.609375
2	2.dcm	256	256	131984	0.372788	0.703125
3	3.dcm	256	256	131984	0.256895	0.046875
4	4.dcm	256	256	140774	0.55702	0.06250
5	5.dcm	256	256	140774	0.54807	0.046875
6	6.dcm	256	256	134816	0.083187	0.046875
7	7.dcm	256	256	140758	0.548033	0.046875
8	8.dcm	256	256	134816	0.120846	0.046875
9	9.dcm	256	256	134816	0.133374	0.046875
10	10.dcm	512	512	534210	0.422463	0.156250
11	11.dcm	256	256	141024	0.578469	0.046875
12	12.dcm	512	512	534170	0.410536	0.12500
13	13.dcm	256	256	131984	0.369893	0.03125
14	14.dcm	256	256	131982	0.381166	0.03125
15	15.dcm	256	256	131984	0.56128	0.046875

the purpose of quality checking. Since reconstructed image is 100 % identical to original input image, there is no need to check for any other quality measures.

To further increase the amount of compression, predictive schemes can be used in conjunction with block based schemes.

7.2 Predictor with Block Coding

MED predictor is used as a Median Edge Predictor in the first stage of encoding. The previously encoded pixels are considered to predict the next pixel in a raster scan order. A large part of spatial redundancy is eliminated by the use of such a predictor. However, attempts are being made to skilfully combine predictive coding with block based techniques. Experimentation has shown and proved that fixed-size block based segmentation is not as effective as variable block based segmentation for lossless image compression.

After segmentation of original image into 8 × 8 sized blocks and. if pixel values are predicted inside each block to give predicted blocks, then for each block one needs to keep few reserved pixel for initial prediction. Due to which there is limitation on block—followed by internal prediction. On other hand, the error image from the predictor is taken as input for the block encoder. This image is segmented into 8 × 8 blocks. The blocks are then successively subtracted from each other, to further lower the pixel values. Following this, arithmetic coding is applied. The performance with respect to compression ratio and time for this method is shown in Table 10.

Table 10 Performance of subtracted blocks after prediction

Sr. no	Name of input image	Image dimensions M × N	Size on disk KB	Output file size KB	CR	Time seconds
1	1	256 × 256	128	28.69956	0.224215	21.6417
2	2	256 × 256	128	28.19964	0.22031	47.4775
3	3	256 × 256	128	21.79825	0.170299	34.3463
4	4	256 × 256	137	51.8949	0.378795	61.0751

7.3 *Blocking Coding with Variable Block Size*

Now instead of applying arithmetic coding, the difference between the maximum and minimum value is checked for in each block. If the difference value is large, it indicates large variation and thus less redundancy. In such a case, the blocks are subdivided using the same criterion till a 2 × 2 block is reached as shown in Fig. 14. In the other case, where redundancy is large, the entire 8 × 8 block is passed as it is. The performance with respect to compression ratio and time for variable block size method is shown in Table 11.

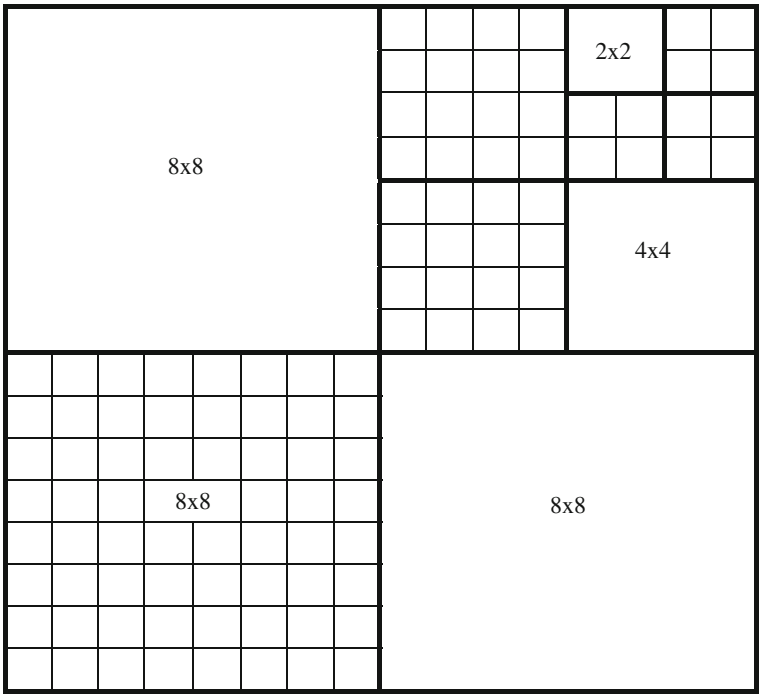


Fig. 14 Variable block size in block coding

Table 11 Performance of variable size block formation

Sr. no	Name of input image	Image dimensions M × N	Size on disk KB	Output file size KB	CR	Time seconds
1	1	256 × 256	128	28.49598	0.222625	32.17885
2	2	256 × 256	128	27.99727	0.218729	35.75553
3	3	256 × 256	128	21.59978	0.168748	38.50462
4	4	256 × 256	137	51.7983	0.37809	49.09732

7.4 Multidimensional Scanning of Blocks

The original image is subjected to an MED predictor, resulting in a predicted image. This image has pixel values differing in small values from the original image pixel values.

The next step involves transforming the 2-D matrix of the predicted image into a block form, of fixed 8 × 8 sized blocks. Since the inputs to our system are of dimensions 256 × 256, the resulting cell would be a 32 × 32 cell, with each entry being an 8 × 8 block.

To further minimize the values in the blocks, we subtract consecutive blocks, and create a new 32 × 32 cell where each 8 × 8 block of predicted image is now replaced by the 8 × 8 block of subtracted value. In doing so, the values are reduced to low values, especially in regions having a more uniform background. Blocks may repeat themselves in the 32 × 32 cell format. Reducing or eliminating this repetition results in a compression. Thus, each block in the cell, is compared with its 3 immediate neighbours: horizontally located neighbour to its right, vertically located neighbour below and its diagonally located neighbour. If the same block is found in any of the neighbouring positions, the corresponding position in a new matrix is marked with a '1', while each new block is represented by a '0' in this new matrix consisting of 0 and 1 s.

To simplify the decoding process, every '0' recorded in the matrix is matched with the corresponding blocks at their respective locations in the cell. This helps to maintain a vector of every 'new' block that is encountered in the encoding process. The 8 × 8 block is broken into its pixel form, resulting in 64 pixels for every 8 × 8 block. The end processing done on the vector is Arithmetic Encoding. The performance this method is shown in Table 12.

Table 12 Performance of multidimensional scanning in block coding

Sr. no	Name of input image	Image dimensions M × N	Size on disk KB	Output file size KB	CR	Time seconds
1	1	256 × 256	128	28.49598	0.222625	32.17885
2	2	256 × 256	128	27.99727	0.218729	35.75553
3	3	256 × 256	128	21.59978	0.168748	38.50462

At the decoder, the reverse Arithmetic Decoding process is carried out to recover the vector containing pixels that form the 'new' blocks. The 0 and 1 s matrix is checked for every 0 that is encountered, and at that pixel position, the decoder places an 8 × 8 block formed from the 64 pixels in the vector. This process helps to recover the original blocks that describe the predicted image. The following step is a simple decoding of the MED predictor to recover the original image without loss.

A multi-direction scanning of blocks is carried out to determine any repetition of blocks. This approach is quite effective but the trade-off between the time and CR is an important point to consider.

Discussion on Block Codes

It is concluded that, the block based coding is useful in the case of plane of pixel with same intensity values or small variation of pixel values. For smaller blocks one needs to store the block size as a separate entity, which increases its size by acting as overhead of actual data. Also due to block representation there may be difficulty in using predictor as predictor needs some values as it is. The computational complexity is more for block codes than pixel based predictor codes, as in block codes one needs to track complete block. It is much simple to use pixel base prediction algorithms as pixel is the smallest unit which cannot be divided further.

8 Symmetry Based Compression

This section deals with employing symmetry as a parameter for compression of biomedical images. The approach presented in this section offers great potential in complete lossless compression of the medical image under consideration, without degrading the diagnostic ability of the image.

8.1 Concept of Symmetry

In geometry, symmetry is exact similarity of position or forms about a given point, line or plane. The extraction of exact symmetry axis is possible only for intrinsically symmetrical objects [46]. In fact, even for symmetrical objects, perfect symmetry is impossible to obtain in digital imaging due to imperfect lighting, digitization or occlusion. Axis of Symmetry is a line that divides the figure into two symmetrical parts. Human body naturally has approximate mirror symmetry. There are some examples of human body organs which possess approximate bilateral symmetry like axial view of brain, pupil, labia, cervical, lumbar, chest, thorax, larynx, lungs, etc. All the above mentioned body parts exhibit more or less bilateral symmetry. An example of the symmetry exhibited in human brain is as shown in the Fig. 15.

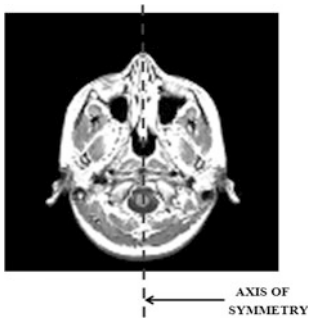


Fig. 15 Symmetric medical image of human brain [46]

The above medical image presents an opportunity to compress it, using symmetry as a parameter. The data present in both halves of this image, considering a vertical axis by crude observation, is almost equal, barring a few pixels. Thus, if data from only a single half is transmitted instead of transmitting both halves, then substantial amount of compression can be achieved.

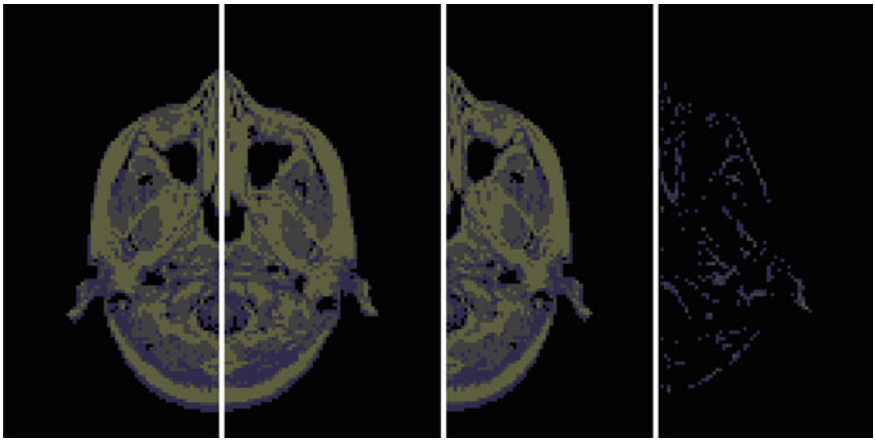
As seen in Fig. 16, it is clear that, if symmetry is used as a parameter to measure redundancy within image then, half of the redundancy can be removed. If the image is perfectly symmetrical, then one can reconstruct one half of image by using information available in other half, in this was 50 % compression will be achieved. But for real times images some asymmetry might be available in it. For complete lossless compression process one needs to preserve the residual part of the subtraction process. The first task is to obtain the axis of symmetry for any kind of image and that too without any error. This axis will act as a differentiator between the two halves of the medical image [45].

Sample Results

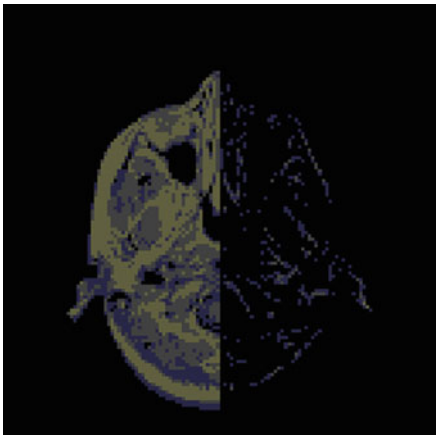
The developed algorithm is tested for its performance on images in database [36, 30, 37]. The developed method is used as one of the preprocessing block along with standard compression algorithms like Huffman coding, Arithmetic coding and JPEG-LS and the results are presented in Tables 13 and 14.

Observations

- Form the Fig. 17 it is clear that using symmetry as a parameter to measure redundancy in an image, one can get lossless compression up to 50 % (if the image is perfectly symmetrical). From Table 13, it is clear that for CT images using symmetry as a pre-processing block, compression ratio of traditional arithmetic conceder is enhanced by 12 % and compression ratio of Huffman encoder is enhanced by 11 %. Also for MRI images an average of 7 % improvement (decrease) is observed. It is clear that even after symmetry processing the Huffman encoder is better than arithmetic encoder.



a) Actual Left Part b) Actual Right Part c) Flip of Left Part d) Right – Flip of left



e) Symmetry Processed Image

Fig. 16 The concept of symmetry processing [46]

- Form Table 14, it is observed that using symmetry as a pre-processing block, compression ratio is enhanced by 47 % as compared to traditional JPEG lossless conceder, and by 81 % as compared to JPEG 2000 LS standard.
- The enhancement in the performance of well know standard algorithm is due to use of symmetry as pre-processing block.
- Using symmetry as pre-processing block, nearly 50 % (depends on degree of similarities in the two half) of the image portion is carrying lower intensity pixels (near to '0').

Table 13 Compression ratio comparison between symmetry based compression and other encoders [46]

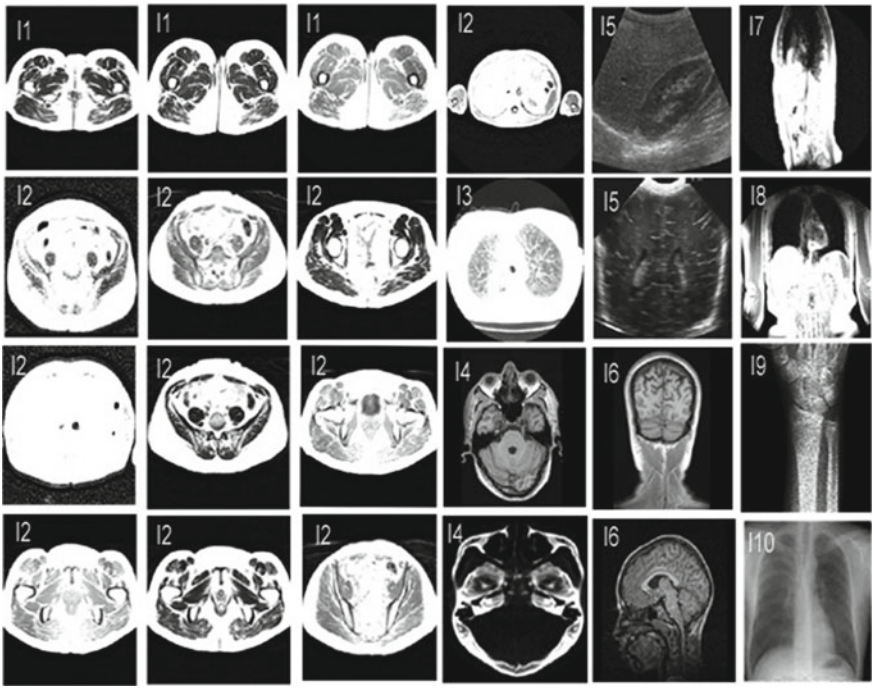
Sr. no	(10 images in each type)	Average compression ratio			
	Name of type	Only Arithmetic encoder	Symmetry + Arithmetic	Only Huffman	Symmetry + Huffman
1	CT-abdomen	0.643574	0.567615	0.485414	0.42882
2	CT-head	0.746698	0.648391	0.562288	0.49331
3	CT-neck	0.751352	0.670353	0.563801	0.509612
4	CT-spine	0.772324	0.676357	0.582679	0.515197
5	MR-abdomen	0.572833	0.53667	0.426529	0.422404
6	MR-head	0.47509	0.4475	0.35847	0.31028
7	MR-pelvic	0.550559	0.522088	0.415903	0.39943
8	X-Lumber	1.06	0.9423	0.7103	0.54493

Table 14 Compression ratio analysis for symmetry processed image [46]

Sr. no	Image type	Size of original image (KB)	Compression ratio			
			Only JPEG-LS	Symmetry + JPEG-LS	Only JPEG −2000 LS	Symmetry + JPEG 2000 LS
1	CT_Abdomen	520	0.321153846	0.228846154	0.25192308	0.104038462
2	CT_Head	520	0.498076923	0.209615385	0.44230769	0.093461538
3	CT_Spine	520	0.457692308	0.185769231	0.38076923	0.081538462
4	MRI_Pelvic	130	0.396923077	0.279230769	0.33538462	0.147692308
5	MRI_Brain	130	0.384615385	0.351538462	0.35307692	0.146153846
6	CT_Chest	514	0.328793774	0.148832685	0.27042802	0.064785992
7	MRI_Dental	514	0.579766537	0.307392996	0.51750973	0.145914397
8	MRI_Chest	563	0.428063943	0.056305506	0.330373	0.026642984
	Average		0.424385724	0.220941398	0.360221535	0.101278499

9 Volumetric Image Compression

Over the past decade, Volumetric or three dimensional (3D) medical imaging as shown in Fig. 18, has become a main pillar of clinical research and practice. Generating such kind of volumetric type of image data an integral part of any patient's diagnosis and records. Medical images acquired by tomography scanners for instance are often given as stacks of cross sectional image slices. Such images are called volumetric because they depict objects, in their entire 3D extent rather than just a projection onto 2D image plane. Since huge amount of volumetric data is being continuously produced in many places around the world techniques for its



L1: Brain CT, L2: Brain MRI, L3: Other Organ CT, L4: Brain L5: Ultra Sound, L6: Brain, L7: Other Organ, L8: Other Organ, L9: Hand X-Ray, L10: Chest X-Ray

Fig. 17 Few representative images in database

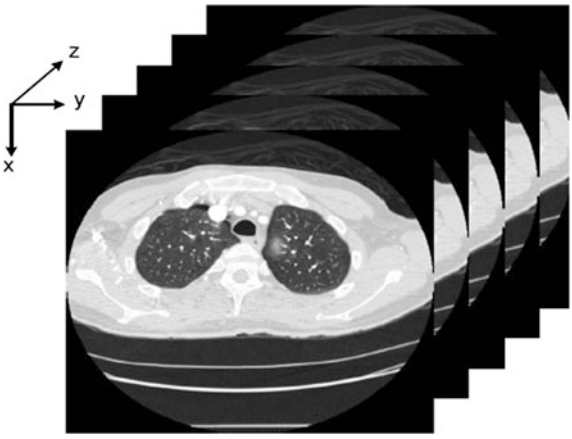


Fig. 18 Volumetric images (3D image)

analysis are more important. It is crucial to improve the techniques that would enable efficient storage and quick access to 3D medical images for future study and follow-up.

The proposed work is focused compression of volumetric data, So as to achieve reduction in storage space requirement of DICOM images in HIS.

Second objective of this section is to achieve compression of the sequence of medical images depending upon the similarity present in the image sequence frames. Stack of the medical images is nothing but the similar type of images with slight differences in their appearance. Idea of image sequence can be understood by the basics of video sequence. The major difference between video sequence and image sequence is as follows,

- Video sequences consist of the sequence of same scene. But these sequences of frames just differ in their position. Video sequence exhibits motion into the frame. So we need to compensate the motion vector and need to find the motion vector from reference frame to the next frame. But this is not the case with medical image sequence; medical image sequence does not consist of any motion in the successive images. So motion compensation is not useful in case of image sequence.
- Speed of video sequence processing is very large, so the method applied to video processing is not useful in case of image sequence. If we apply the similar method as that of video sequence. We might lose valuable information from the image sequence. There is need for different method for image sequence compression. To solve this problem, we have presented lossless method for compressing volumetric data.

Working: The symmetry processed image is the input to the algorithm. First Inter frame and then Intra frame differential coding is applied as shown in Fig. 19. Finally Huffman encoding is used for further compression. The method is completely lossless in nature.

Figure 19 represent the results of various test image stacks of different modality. In each set 5 images are taken which are in stack. The developed algorithm is applied to various image sets. The performance of algorithm is tested with respect to JPEG LS. The graph is plot with file size on y-axis verses different data sets on x-axis (Fig. 20).

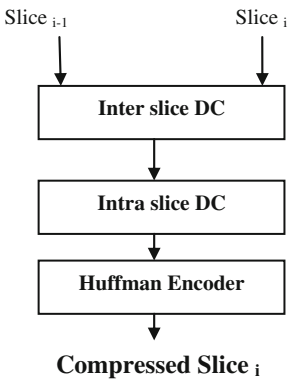


Fig. 19 Block diagram of developed approach for sequence of images

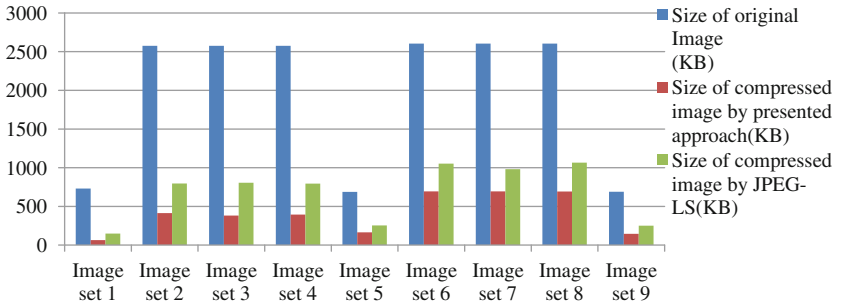


Fig. 20 Comparative analysis in graphical form of presented approach and JPEG-LS

Table 15 Compression performance of symmetry processed image on image sequence

Sr. no	Image set	Size, no of images	Total Orig. Size	CR of developed method	CR of JPEG-LS
1	MR_Overlay1	256 × 256, 5	730 KB	0.088	0.202
2	MRI_Prostate1	512 × 512, 5	2575 KB	0.160	0.309
3	MRI_Brain1	512 × 512, 5	2605 KB	0.266	0.404
4	CT_Spine	512 × 512, 59	29.9 MB	0.397	0.428
5	CT_thorax	512 × 512, 76	38.1 MB	0.296	0.406

The algorithm is tested and verified on more than 500 MRI/CT/Ultra Sound/X-Ray, DICOM images (available from private domain database as well as from public domain images) [36, 30, 37], few of them representative images from database are shown in Fig. 17.

Observations

- Form the Table 15, it is clear that, One can effectively use the intra frame redundancy present in sequence of images. The symmetry based compression method for image stack is superior to JPEG-lossless than 30.99 %.

Discussion on Symmetry Based Compression

The results show that the symmetry based compression approach gives enhanced compression ratios as compared to the standalone compression method. This symmetry based compression technique is completely lossless and hence the diagnostic ability of the image is fully retained. Thus, higher compression ratios are obtained but at a cost of slightly higher time of (near about 20 %) processing.

The symmetry hence proves to be an important parameter which can be exploited redundancy in medical image, for achieving greater compression while at the same time not compromising with the data quality. The symmetry based technique can be as pre-processing block for any compression algorithm to enhance its compression performance.

10 Conclusion

Big data analytics is going to play a vital role in healthcare in the future. A major goal of telemedicine is to eliminate unnecessary travel of patients and their escorts. Telemedicine is characterized by limitations of bandwidth in transmission media for Information and Communication Technology (ICT), especially for rural area. Image compression is beneficial for telemedicine in reducing the storage and transmission bandwidth requirements of medical images for telemedicine. Here in this chapter we have discussed in general about Big Data Analytics in Telemedicine, specifically about A Role of Medical Image Compression.

In comparison with transform based compression, the prediction based compression methods are simple and fast. For fast lossless compression prediction based compression is to be preferred. The prediction algorithms give much deviation from original value resulting in more error at the edge. To overcome this drawback, the edge based prediction method is presented, which gives better result in terms of compression ratio as compared with other prediction methods. The overall approach is compared with available standard i.e. JPEG, and it is found that the presented approach is 21.68 % enhanced than JPEG with respect to compression ratio. The developed GAPPE approach is found to be 33.68 % enhanced with respect to compression ratio and it takes almost 17.80 % less time as compared than CALIC.

The new block based differential prediction approach is presented and tested on various images. From the experiment, on varying the size of block it is found that 8×8 block size of pixel is most suitable for MRI, CT, US, X-Ray modality of images. It is concluded that pixel based prediction algorithms are much simple to use, as pixel is the smallest unit which cannot be divided further, compared to block based coding on medical images.

It is observed that for an image of size $512 \times 512 \times 8$ bit, the storage size required on disk is 256 KB with transmission time of ~35 s on 64 kbps, ISDN line, but due to compression up to 25 % of original size, the transmission time gets reduced to ~9 s. In tele-health services for uncompressed data a bandwidth of 384 kbps is recommended but with compression, even 128 kbps bandwidth would be sufficient for telemedicine. Thus, medical image compression is largely necessary in big data-Telemedicine applications especially in rural areas, where bandwidth constraint is a major issue.

References

1. Lavanian D (2009) Technology enabling rural healthcare. Lecturer presentation Apollo Telemedicine Networking Foundation, India
2. Janet J, Mohandass D, Meenalosini S (2011) Lossless compression techniques for Medical images in Telemedicine. Advances in Telemedicine: Technologies, Enabling Factors and Scenarios, INTECH Open Access Publisher, pp 111–30.

3. Sajeesh K (2008) Introduction to telesurgery. Springer Berlin Heidelberg, pp 1–8
4. Ganapathy K (2016) Telemedicine in India-the apollo experience: http://www.thamburaj.com/telemedicine.html accessed on 29 July 2016
5. Delgorge C, Rosenberger C, Vieyres P, Poisson G (2002) JPEG 2000, an adapted compression method for ultrasound images? A comparative study. In: International conference systemics, cybernetics and informatics, vol 9, pp 536–539
6. Bollineni R (2011) Case study on apollo telemedicine networking foundation, Access Health International, Indian School of Business, Hyderabad
7. Bedi BS (2003) Telemedicine in India: Initiatives and perspective, Journal of eHealth 2003: Addressing the digital divide-17th Oct
8. Wang Z, Gu H (2009) A review of telemedicine in China. Int J Telemed Telecare 15(1):23–27
9. Kombaiya K, Palanisamy V (2009) Wavelet based image compression using ROI SPIHT Coding. Int J Comput Intell Res 5(1):67–74
10. Giakoumaki A, Perakis K, Tagaris A, Koutsouris D (2006) Digital watermarking in telemedicine applications—towards enhanced data security and accessibility. In: Proceedings of the 28th IEEE EMBS international conference, New York City, USA, pp 6328–6331
11. Baeza I, Verdoy A (2009) ROI-based procedures for progressive transmission of digital images: a comparison. Elsevier J Math Comput Model 50(6):849–859
12. Tracy J (2004) A guide to getting started in telemedicine. University of Missouri—School of Medicine Publisher, Missouri, pp 10–50
13. http://www.cdacmohali.in. Accessed Jan 2007
14. Miaou S-G, Ke F-S, Chen S-C (2009) A lossless compression method for medical image sequences using JPEG-LS and interframe coding. IEEE Trans Inf Technol Biomed 13 (5):818–821
15. Digital Imaging and Communications in Medicine (DICOM), Part 5: Data Structures and Encoding. http://medical.nema.org/standard.html. Accessed April 2012
16. Bagchi S (2006) Telemedicine in rural India. Online J Public Libr Sci Med 3(3):e82 (Pub Med Publication, pp 1–4)
17. DICOM Traffic Performance and WAAS Application Deployment Guide. http://www.cisco.com. Accessed Dec 2011
18. Mukherjee J (2003) Telemedicine: lecture note. Department of Computer Science and Engineering, I.I.T., Kharagpur, India, pp 1–6
19. Technical Working Group on Telemedicine Standardization (2003) Recommended guidelines & standards for practice of telemedicine in India. Published by Ministry of Communications and Information Technology Government of India
20. Wu D, Tan EC (2002) Comparison of lossless image compression algorithm. In: Proceedings of the IEEE region 10 conference TENCON 99. 1999. vol 1, pp 718–721 (Curr. Ver 2002)
21. Sayood K (2010) Introduction to data compression, 3rd edn. Elsevier Publication, San Francisco, CA, pp 473–512
22. Elizabeth SK, Krupinski A (2008) Teleradiology, 1st edn. Springer London Publication, London, pp 10–21
23. Ali TJ, Akhtar P (2008) "Significance of region of interest applied on MRI and CT images in teleradiology-telemedicine. In: Gao X (ed) Medical imaging and informatics, vol 4987. Springer, Berlin, pp 151–159
24. Gornale SS, Humbe VT, Jambhorkar SS, Yannawar P, Manza R, Kale KV (2007) Multi-resolution system for MRI (Magnetic Resonance Imaging) image compression: a heterogeneous wavelet filters bank approach. In: IEEE conference preceding on computer graphics, imaging and visualization (CGIV 2007) Bangkok, pp 68–73
25. Gonzalez RC, Woods RE (2010) Digital image processing. Prentice Hall Publication, New Jersey
26. Technical Report on the second global survey on eHealth (2010) Telemedicine Opportunities and Developments. Global Observatory for eHealth series, World Health Organization, vol 2, 2010

27. Sanchez V, Abugharbieh R, Nasiopoulos P (2009) Symmetry-based scalable lossless compression of 3D medical image data. IEEE Trans Med Imaging 28(7):1062–1072
28. Knezovic J, Kovac M, Mlinaric H (2006) A new adaptive blending predictor for lossless image compression. In: IEEE ITI 4th international conference on information and communications technology, ICICT'06, Cairo, Egypt, pp 1–12
29. Van Boom D (2011) Method and device for encoding and decoding of data in unique number values. United States Patent Application Publication, Pub no. US2011/0122113 A1, May 26, 2011
30. Karimi N, Samavi S, Reza Soroushmehr SM, Shirani S, Najarian K (2016) Toward practical guideline for design of image compression algorithms for biomedical applications. Elsevier J Expert Syst Appl 56:360–367
31. The Whole Brain—Atlas. http://www.med.harvard.edu/AANLIB/home.htm. Accessed Dec 2011
32. Weinlich A, Amon P, Hutter A (2016) Probability distribution estimation for autoregressive pixel-predictive image coding. IEEE Trans Image Process 25(3):1382–1395
33. The National Library of Medicine (NLM). http://www.nlm.nih.gov. Accessed July 2011
34. Penrose AJ (2001) Extending lossless image compression. Technical Report No. 526, University of Cambridge
35. Amar A, leshem A, Gastpar M (2010) Recursive implementation of the distributed Karhunen-Loève transform. IEEE Trans Signal Process 58(10):5320–5330
36. Wallace GK (1992) The JPEG still picture compression standard. IEEE Trans Consum Electron 38(1):xviii–xxxiv
37. Memon N, Sayood K (1995) Lossless image compression: a comparative study. Proc SPIE Still Image Compress 2418:8–20
38. Sepehrband F, Mortazavi M, Ghorshi S, Choupan J (2011) Simple lossless and near-lossless medical image compression based on enhanced DPCM transformation, IEEE Intentional Conference PacRim, Canada, Aug. 2011, pp 66–72
39. Computer vision group. http://decsai.ugr.es/cvg/index2.php. Accessed Aug 2011
40. American College of Radiology (2002) ACR standard for teleradiology, pp 13–21. http://imaging.stryker.com/images/ACR_Standards-Teleradiology.pdf. Accessed June 2007
41. Avramovic A (2011) Lossless compression of medical images based on gradient edge detection. In: Proceedings of IEEE international conference, 19th telecommunications forum TELFOR 2011, Belgrade, pp 1199–1202
42. Peng Z, Huang Y-F, Costello DJ (2000) Turbo codes for image transmission-a joint channel and source decoding approach. IEEE J Sel Areas Commun 18(6):868–879
43. Hu H (2004) A study of CALIC. MS Thesis, University of Maryland, Baltimore
44. Moursi SG, El-Sakka MR (2007) Improving CALIC compression performance on binary images. In: Proceeding of EURASIP international conference on picture coding symposium 2007, Portugal, pp 1–4
45. Rane SD, Sapiro G (2001) Evaluation of JPEG-LS, the new lossless and controlled-lossy still image compression standard, for compression of high-resolution elevation data. IEEE Trans Geo Sci Remote Sens 39(10):2298–2306
46. Weinberger MJ, Sapiro G, Seroussi G (2000) The LOCO-I lossless image compression algorithm: principle and standardization into JPEG-LS. IEEE Trans Image Process 9(8):1309–1324

A Bundle-Like Algorithm for Big Data Network Design with Risk-Averse Signal Control Optimization

Suh-Wen Chiou

Abstract A big data network design with risk-averse signal control optimization (RISCO) is considered to regulate the risk associated with hazmat transportation and minimize total travel delay. A bi-level network design model is presented for RISCO subject to equilibrium flow. A weighted sum risk equilibrium model is proposed to determine generalized travel cost at lower level problem. Since the bi-objective signal control optimization is generally non-convex and non-smooth, a bundle-like efficient algorithm is presented to solve the equilibrium-based model effectively. A big data bounding strategy is developed to stabilize solutions of risk-averse signal control optimization with modest computational efforts. In order to investigate computational advantage of proposed algorithm for big data network design with signal optimization, numerical comparisons using real data example and general networks are made with current best well-known algorithms. The results strongly indicate that the proposed algorithm becomes increasingly computationally comparative to best known alternatives as the size of network grows.

Keywords Bundle methods · Big data network design · Signal optimization · User equilibrium · Bi-level programming · Numerical computation

1 Introduction

For most urban road networks, the reliability of a signal-controlled road network heavily depends on its vulnerability to a dangerous mix of probabilistic threats such as lane closure and road capacity loss. Particularly, transportation of hazardous material (hazmat) is of primary concern to decision makers due to serious safety, human health, and environmental risks associated with the release of hazmat. Because of the danger associated with the accidental release of hazmat, the people

S.-W. Chiou (✉)
Department of Information Management,
National Dong Hwa University, Hualien County, Taiwan
e-mail: chiou@mail.ndhu.edu.tw

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_8

living and working around the roads heavily used for hazmat thus incur most of the risk during transportation. For a signal-controlled road network, one direct approach usually adopted by government to simultaneously regulate the risks associated with hazmat shipment and reduce travel delay is the prohibition on the use of certain roads by hazmat traffic [1–5]. A more amiable and flexible policy for regulators on effective road control can be achieved through appropriate signal design. In effect, the signal-controlled road network available for the carriers' operations is determined by the regulator. A multi-objective program has long been regarded as one of the most popular approaches to tackle a general network design problem where the interest of stakeholders is in conflict [6–10]. The multi-objective program often assumes that the multiple objectives which are established by the same decision makers and located at the same level. As the objectives have to be optimized simultaneously, a tradeoff needs to be determined for compromising the multiple objectives. For example, [6] considered a multi-objective program in which two conflicting objectives between system cost and greenhouse gas emissions are integrally considered. Zhao and Verter [7] presented a bi-objective model for the location and routing problem to simultaneously minimize environment risk and cost. A weighted goal programming was employed to solve the location-routing problem with a case study. Since multi-objective programs can hardly solve the problems with multiple decision makers at different levels, on the other hand, the bi-level decision-making following a leader and follower relationship to sequentially optimize the objectives has recently been noticed [3–5, 11]. In this regard, [5] proposed a family of valid cuts and a typical cutting plane approach for the combinatorial formulation of the bi-level hazmat transport network design. Numerical tests were performed using real and random data and comparisons were widely made with current popular methods. The results obviously show the proposed cutting plane algorithm is faster than all alternatives in the literature for combinatorial formulation of the bi-level hazmat network design problem. More recently, [8] proposed a multi-objective bi-level location planning problem for stone industrial parks with a hierarchical structure under a random environment. The proposed model captured cost uncertainties and multiple decision makers with conflicting interests and solved by adaptive chaotic particle swarm optimization.

For a signal-controlled road network most regulators do not have the authority to impose routes on hazmat carriers. Therefore, transportation risk of hazmat is an outcome of the carriers' route choice over the available signal-controlled network. While the authority is primarily concerned about mitigation of public risk, the carriers are more interested in minimizing transport costs. Considering the rationale response of carriers, [12] were the first ones to propose a bi-level network design model for hazmat transportation with single objective of risk minimization. A bi-level problem of minimizing population exposure risk is addressed by banning hazmat traffic from traveling on certain segments of road network. Numerical results have shown significant reduction in population exposure can be achieved through effective government interventions while the carriers incurred more increased travel costs compared to the use of minimum cost routes. Furthermore, [2] improved the bi-objective bi-level hazmat network design problem by a proposed

heuristic. Numerical computations were performed using real data road network and good computational performance of proposed heuristic has been reported. Veter and Kara [13] also presented a path-based formulation for a bi-level hazmat network design to find compromise solutions between the regulator and the carriers. Despite traditional risk model taking the form of expected consequence of incidents is regarded as the most-widely used one in literature of modeling transport risk of hazmat [13–15], the development of risk-averse models for low probability but high risk hazmat routing has attracted the attention of researchers only fairly recently [16–18]. In this regard, [16] introduced three catastrophe-avoidance models for hazmat routing. The first model is the maximum risk model aiming to avoid catastrophes by minimizing the maximum risk along a route. The second model incorporates the variance of consequence along a route into route choice, and the third one minimizes the expected consequence using a convex utility function. It is shown that all three models can be solved as typical shortest path problems, and the first is recommended as an appropriate one in terms of computational tractability. By assuming that route usage probabilities and link accident probabilities are two person non-cooperative and zero-sum players in the mixed strategy Nash game, [18] proposed an alternative to reduce the maximum risk along a route. A mixed route strategy by sharing shipments among routes is considered to effectively reduce the maximum population exposure under uncertain probability of incidents. As a result, it generates the potential link usage probabilities and most vulnerable links with potentially high accident probabilities. However, numerical computations for various risk models proposed in [16–18] have not well been investigated and no comparison was numerically made with alternatives. Although the risk models in [16] and [18], for example, provide an interesting framework for tackling hazmat network design, apparently the advantage of risk-averse models has not yet been fully exploited from the numerical point of view, in particular for a general road network with signal settings.

A bi-level road network design problem can be regarded as one of the most challenging problems in the field of operations research and computation science facing decision makers at various levels of society. In order to effectively determine optimal allocation of limited resource, decision makers at upper level need to consider entire benefit of whole society whilst taking account of road users' response at lower level. The bi-level continuous network design can be regarded as a special case of a mathematical program with equilibrium constraints (MPEC) when the lower level problem can be expressed as an equilibrium constraint. As widely commented in the literature [19–21], the MPEC is difficult to solve with a global optimum due to its non-convexity and therefore becomes increasingly computationally intractable as the instance of problem grows. Recently, [22] presented a global optimum solution for a linearized network design problem with single objective where the cost function is approximated to a sequence of piecewise linearized functions. A mixed-integer linear program was proposed to effectively solve the approximation of bi-level continuous network design and preliminary numerical tests were conducted using hypothetical small-scale road network. [23] proposed a similar link-based global optimization method (LMILP) for a mixed

transportation network design problem through a mixed-integer linear programming approach. For a transportation network design with mixed variables (continuous and discrete) at upper level, a traditional Wardrop equilibrium problem is considered at the lower level and formulated as a general variational inequality. A bi-level network design problem with single objective can be transferred as a single-level mixed-integer linear program via a sequence of piecewise-linear programs. A cutting constraint method was proposed to solve the mixed-integer linear program with success and preliminary computational comparisons to alternatives in terms of solution effectiveness were also made using a small-scale road network. However, computational tractability of the proposed algorithm solving general road network of large scale has not been fully investigated. Li et al. [24] also presented a global optimization method for general continuous network problems. Based on the concept of gap function, the bi-level continuous network design problem can be transferred to a single-level program. A novel approach employing multiple cutting planes with penalty (PMC) was proposed to globally solve the continuous network design problem with single objective. Numerical calculations were illustrated using a small-scale road network. Although experimental comparisons were numerically made with limited options, the computational tractability of proposed algorithm for general road networks has not yet been fully explored. The merit of proposed algorithm in computational efficiency accordingly has not yet been completely revealed when applying to large-scale road networks. More recently, [25] presented a relaxation heuristic for a bi-level continuous network design with single objective to reduce traffic congestion through tradable credit scheme and equity constraints. Numerical computations were performed using a real-data network. Despite the proposed heuristic worked far better than most alternatives [22–24], providing an interesting framework for dealing with continuous road network design, apparently the advantage of proposed heuristic has not yet been fully explored from the numerical prospective, in particular for a general road network of large scale size.

As mentioned in the literature [19–21], a bi-level programming problem is very difficult to solve particular for large scale road network and unfortunately there are only local optimal solutions that can be found due to the non-linearity of the constraints at the lower level problem. In this paper, a big data risk-averse signal control optimization (RISCO) is considered to simultaneously regulate the risk associated with hazmat traffic and minimize total travel cost. An equilibrium-based network design model (EQ-based) is introduced to tackle transport risk and generalized travel cost. A proximal bundle-like method (PBM) is proposed to solve the EQ-based model with global convergence. At the lower level a weighted sum risk equilibrium model (WSM) is presented for hazmat shipment with low probability but high consequence. The maximum risk exposure of population adjacent to selected routes of hazmat and that over entire road network can be accordingly minimized. For a hazmat network design, two benchmarks: the cost minimum routes (CM) in a risk neutral model [26] and the maximum risk model (MM) [16, 18], are numerically compared with WSM. At the upper level, a bi-objective performance measure of signal-controlled network for hazmat transport can be evaluated using a well-known traffic model TRANSYT [27]. The multiple criteria

performance measure for traffic movements has been approximately derived as mathematical expressions [28]. Due to the non-convexity of RISCO, in this paper, a bounding solution aiming to reduce relative gaps between iterations for non-smooth RISCO is developed. The minimization of non-smooth functions that are given by bundle type methods has been successfully approached [29–33].

For non-smooth optimization problems, the bundle-like algorithms are currently considered ones of the most efficient optimization methods as compared to classical cutting plane (CP) method [30–32, 34–38]. The conventional approach to designing algorithms for non-smooth optimization is to stabilize steepest descent by exploiting gradient or sub-gradient information evaluated at multiple points, which constitutes the essential idea of bundle methods [34, 37, 38]. From recent numerical experiments in non-smooth optimization methods and solvers [29], the PBM was highly recommended as one of the most efficient solvers for the piecewise quadratic and piecewise linear problems due to the least number required of evaluations for problems of any size. Especially, the PBM is suggested as a good choice for a solver when the objection function value and the sub-gradient are expensive to compute. In this paper, we provide a fully implementable variant of bundle method that opens the way to bundling recent information of the smooth mapping when available. A hybrid of the bundle methods [34, 35] is proposed to fill the gap, which exists in the field of demanding bi-objective risk-averse signal control optimization (RISCO) with large number of variables. The new method exploits the ideas of the bundle methods, namely the utilization of sub-differential, aggregation of collective sub-gradients, and the sub-gradient locality measures. By employing prox-center updates, the bundle gradients projected onto null space of active constraints are used to search for a better local solution for the EQ-based model. By successive executing the PBM, consecutive bundle gradients can make the EQ-based model more computationally efficient. Therefore, the improvements in CPU time reduction can be apparently observed in this paper as compared to those did the CP via a variety of extensive numerical experiments in continuous road network design.

The contributions made from this paper can be briefly stated as follows. Firstly, a big data risk-averse signal control optimization (RISCO) is proposed to simultaneously mitigate maximum risk exposure and reduce travel delay. An equilibrium-based bi-level network design model (EQ-based) is presented to solve the RISCO. A weighted sum risk equilibrium model (WSM) is established to determine the maximum risk exposure at the lower level where Wardrop's first principle is respected. Due to kinky structure of RISCO, secondly, a prox-center bundle like method (PBM) is proposed to stabilize solutions of RISCO with global convergence. Numerical computations are performed twofold using a variety of example road networks. In the first place, we compare proposed EQ-based model to well-known benchmarks such as CM [26], MM [16] and MM2 [18]. Computations are performed at two sets of initial data using a real data benchmark example network. Numerical comparisons are also made with alternatives such as CP [31, 32, 39]. By contrast with recent developments in solving continuous road network design problem (CNDP) [23–25], in the second place, we compare PBM with best well-known solvers for CNDP such as LMILP [23] and PMC [24] using general

road network of large-scale size. The trade-offs between risk exposure and generalized travel costs of comparative algorithms against CPU time are also empirically investigated. As it shows, PBM consistently exhibits considerable advantage on computation efficiency in all cases whilst guaranteeing minimum risk exposure spreading over road network. The results strongly indicate that PBM becomes increasingly comparative when improving system performance whilst incurred relatively less CPU time as size of network grows. The rest of the paper is organized as follows. Section 2 introduces a big data network design with risk-averse signal control optimization (RISCO). An equilibrium-based network design model (EQ-based) is presented. In Sect. 3, a prox-center bundle like method (PBM) to effectively solve the EQ-based model is proposed. In Sect. 4, numerical computations of proposed scheme are performed using a variety of example networks. Extensive numerical comparisons are empirically made with current best known algorithms solving continuous network design models. Conclusions for this paper and extensions of the proposed approach to topics of interest are briefly summarized in Sect. 5.

2 A Big Data Network Design Model

An equilibrium-based bi-level network design model (EQ-based) is proposed in this section. At the lower level a weighted sum risk equilibrium model (WSM) is established to determine the maximum risk exposure for equilibrium flow. The WSM is presented in which maximum risk on links along selected routes can be identified when the probability of risk exposure is not known a priori. At the upper level the performance measure of signal-controlled network can be evaluated using a well-known traffic model TRANSYT. The calculations of indicator of traffic condition are mathematically derived in [28]. The performance index in a signal-controlled road network thus can be expressed as a weighted sum of risk exposure and a linear combination of the rate of delay and the number of stops for each link. Notation used throughout this paper is stated first.

2.1 Notation

Let $G(N, L)$ denote a directed road network, where N represents a set of fixed time signal controlled junctions and L represents a set of links denoted by (i, j), $\forall (i, j) \in L$. Each traffic stream approaching any junction is represented by its own link.

W	a set of origin-destination (OD) pairs
R_w	a set of routes between OD pair w, $\forall w \in W$
$T = [T_w]$	the matrix of travel demands for origin-destination pair w, $\forall w \in W$

ζ	the reciprocal of the common cycle time
$\zeta_{\min}, \zeta_{\max}$	the minimum and maximum reciprocal of the common cycle time
$\theta = [\theta_{am}]$	the vector of starts of green for various links as proportions of cycle time where θ_{am} is start of next green for signal group a at junction m
$\phi = [\phi_{am}]$	the vector of durations of green for various links as proportions of cycle time where ϕ_{am} is the duration of green for signal group a at junction m
τ_{abm}	the clearance time between the end of green for signal group a and the start of green for incompatible signal group b at junction m
$\Psi = (\zeta, \theta, \phi)$	the set of signal setting variables, respectively for the reciprocal of common cycle time, start and duration of greens
$\lambda_{(i,j)}$	duration of effective green for link (i,j), $\forall (i,j) \in L$
$\lambda_{\min}$	the minimum green
$\Omega_m(a,b)$	collection of numbers 0 and 1 for each pair of incompatible signal groups at junction m; where $\Omega_m(a,b) = 0$ if the start of green for signal group a proceeds that of b and $\Omega_m(a,b) = 1$, otherwise
$D_{(i,j)}$	the rate of delay on link (i,j), $\forall (i,j) \in L$
$S_{(i,j)}$	the number of stops per unit time on link (i,j), $\forall (i,j) \in L$
W_D	weighting factor for rate of delay
W_S	weighting factor for number of stops
M_D	monetary factor associated with $D_{(i,j)}$
M_S	monetary factor associated with $S_{(i,j)}$
$\rho_{(i,j)}$	maximum degree of saturation for link (i,j), $\forall (i,j) \in L$
$s_{(i,j)}$	saturation flow on link (i,j), $\forall (i,j) \in L$
$q_{(i,j)}$	incidental probability of accidental release of hazmat on link (i,j), $\forall (i,j) \in L$
$r_{(i,j)}$	incidental consequence of accidental release of hazmat on link (i,j), $\forall (i,j) \in L$
$f_{(i,j)}$	hazmat traffic flow on link (i,j), $\forall (i,j) \in L$
h_k	hazmat traffic flow on route k between OD trips, $\forall k \in R_w, w \in W$
Λ	a link-route incidence matrix with entry $\Lambda^k_{(i,j)} = 1$ if route k uses link (i,j), and $\Lambda^k_{(i,j)} = 0$ otherwise, $\forall (i,j) \in L$, $\forall k \in R_w, w \in W$
Γ	an OD-route incidence matrix with entry $\Gamma^w_k = 1$ if path k connects OD trip w, $\forall w \in W$, and $\Gamma^w_k = 0$ otherwise, $\forall k \in R_w, w \in W$
$c_{(i,j)}$	the travel time on link (i,j), $\forall (i,j) \in L$
$c^0_{(i,j)}$	the un-delayed travel time on link (i,j), $\forall (i,j) \in L$
$d_{(i,j)}$	the average delay on link (i,j), $\forall (i,j) \in L$
C_k	the travel time on route k, i.e. $C_k = \sum_{(i,j) \in L} \Lambda^k_{(i,j)} c_{(i,j)}$, $\forall k \in R_w, w \in W$
σ	a converting factor from risk to monetary factor
σ_c	a converting factor from expected risk to travel cost

2.2 The Lower Level Problem

At the lower level problem, a weighted sum risk model combining signal delay at downstream junction is proposed in this section.

2.2.1 The Cost Minimum (CM) Model

For a signal-controlled road network, the travel time on link can be calculated as a sum of cruise travel time $c^0_{(i,j)}$ on the link (i,j) and the average delay $d_{(i,j)}$ incurred at the downstream junction, i.e.

$$c_{(i,j)}(f,\Psi) = c^0_{(i,j)} + d_{(i,j)}(f,\Psi) \tag{1}$$

The cost minimum routes (CM) among pairs of OD w can be found through a system optimum formulation in the following way:

$$\underset{f}{Min} \sum_{(i,j)\in L} f_{(i,j)} c_{(i,j)}(f,\Psi) \tag{2}$$

$$\text{subject to } \sum_{k\in R_w} h_k = T_w, \forall w \in W$$

$$f_{(i,j)} = \sum_{w\in W} \sum_{k\in R_w} \Lambda^k_{(i,j)} h_k, \forall (i,j) \in L$$

$$h_k \geq 0, \forall k \in R_w, w \in W$$

Let Ω denote the feasible set for feasible traffic flow f in a following vector form, i.e.

$$\Omega = \{f : f = \Lambda h, \Gamma h = T,\ h \geq 0\}$$

2.2.2 A Maximum Risk Model (MM)

According to [26], the expected consequence due to hazmat transport on link can be generalized as a following form: let $c^H_{(i,j)}$ denotes a public risk with probability $q_{(i,j)}$ of accidental release for hazmat to population exposure on link *(i, j)*, we have

$$c^H_{(i,j)} = r_{(i,j)} q_{(i,j)} \tag{3}$$

According to (1), the generalized cost $c^G_{(i,j)}$ on a link (i,j) can be expressed as follows.

$$c^G_{(i,j)}(f,\Psi) = c_{(i,j)}(f,\Psi) + \sigma_c c^H_{(i,j)} \tag{4}$$

Let $C^G_k(f,\Psi)$ denote the generalized route cost on route k, we have

$$C^G_k(f,\Psi) = \sum_{(i,j)\in L} c^G_{(i,j)}(f,\Psi)\Lambda^k_{(i,j)} \tag{5}$$

Thus total travel cost in a signal-controlled road network can be expressed in the following manner:

$$TC^G(f,\Psi,q) = \sum_{(i,j)\in L} f_{(i,j)} c^G_{(i,j)}(f,\Psi) = \sum_{(i,j)\in L} f_{(i,j)} c_{(i,j)}(f,\Psi) + \sigma_c \sum_{(i,j)\in L} f_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{6}$$

Since the distribution of occurrence of probability q in (6) is not always available, the expected consequence of incident in a signal-controlled network can be approximated by a maximum risk model. According to [16], a maximum risk link along a chosen route k can be identified as a following form: for every route k,

$$c^M_k = \underset{(i,j)\in k}{Max}\left\{c^H_{(i,j)}\right\} \tag{7}$$

Thus a least maximum risk model (MM) can be expressed as follows.

$$\underset{k\in R_w, w\in W}{Min} c^M_k = \underset{k\in R_w, w\in W}{Min}\ \underset{(i,j)\in k}{Max}\left\{c^H_{(i,j)}\right\} \tag{8}$$

Find a route k' with a least risk in MM (8) such that for every route k, we have

$$k' = \arg\underset{k}{Min}\, c^M_k \tag{9}$$

2.2.3 A Maximum Risk Model with Mixed Routes (MM2)

Following [16, 18] considered mixed routes and proposed an alternative to reduce the maximum risk along a route in the following manner: let p_k denote path use probability for a trip w such that

$$1 = \sum_{k\in R_w} p_k \tag{10}$$

Let $p_{(i,j)}$ denote a link use probability such that for every link (i,j), by definition, we have

$$p_{(i,j)} = \sum_{k \in R_w} \Lambda^k_{(i,j)} p_k \tag{11}$$

The expected consequence in (3) with link use probability $p_{(i,j)}$ can be re-expressed in a following form:

$$c^H_{(i,j)} = p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{12}$$

subject to (10) and (11). A maximum risk model $c^M(p)$ with a mixed-route selection probability p_k can be described as a following form: for any link use probability $p_{(i,j)}$ satisfying (10) and (11), we have

$$c^M(p) = \underset{q_{(i,j)}}{Max} \sum_{(i,j) \in L} c^H_{(i,j)} = \underset{q_{(i,j)}}{Max} \sum_{(i,j) \in L} p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{13}$$

subject to $1 = \sum_{(i,j) \in L} q_{(i,j)}$

Therefore a least maximum risk model with link use probabilities (MM2) over entire road network can be re-expressed as follows.

$$\underset{p_{(i,j)}}{Min}\, c^M(p) = \underset{p_{(i,j)}}{Min}\, \underset{q_{(i,j)}}{Max} \sum_{(i,j) \in L} p_{(i,j)} r_{(i,j)} q_{(i,j)} \tag{14}$$

subject to

$$\sum_{(i,j) \in L} q_{(i,j)} = 1$$

and

$$1 = \sum_{k \in R_w} p_k \tag{15}$$

together with

$$p_{(i,j)} = \sum_{k \in R_w} \Lambda^k_{(i,j)} p_k \tag{15}$$

2.2.4 A Weighted Sum Risk Equilibrium Model (WSM)

According to Wardrop's first principle, a weighted sum risk equilibrium model (WSM) taking account of signal delay can be proposed as follows.

$$TC^{Q}(f,\Psi)=\sum_{(i,j)\in L}\int_{0}^{f_{(i,j)}}c_{(i,j)}(x,\Psi)dx+\sigma_c\underset{q_{(i,j)}}{Max}\sum_{(i,j)\in L}f_{(i,j)}r_{(i,j)}q_{(i,j)} \tag{16}$$

subject to $\sum_{(i,j)\in L}q_{(i,j)}=1$. The occurrence of probability $q_{(i,j)}$ in (16) maximizing total risk over entire road network can be determined as follows.

$$q^{M}=\arg\underset{q}{Max}\sum_{(i,j)\in L}\int_{0}^{f_{(i,j)}}c_{(i,j)}(x,\Psi)dx+\sigma_c\sum_{(i,j)\in L}f_{(i,j)}r_{(i,j)}q_{(i,j)} \tag{17}$$

subject to $\sum_{(i,j)\in L}q_{(i,j)}=1$

Thus we have

$$TC^{Q}(f,\Psi)=\sum_{(i,j)\in L}\int_{0}^{f_{(i,j)}}c_{(i,j)}(x,\Psi)dx+\sigma_c\sum_{(i,j)\in L}f_{(i,j)}r_{(i,j)}q^{M}_{(i,j)} \tag{18}$$

Therefore the WSM (16) can be expressed in the following way:

$$\underset{f}{Min}\,\underset{q}{Max}\sum_{(i,j)\in L}\int_{0}^{f_{(i,j)}}c_{(i,j)}(x,\Psi)dx+\sigma_c\sum_{(i,j)\in L}f_{(i,j)}r_{(i,j)}q_{(i,j)} \tag{19}$$

subject to

$$\begin{aligned}&\sum_{k\in R_w}h_k=T_w,\forall w\in W\\&f_{(i,j)}=\sum_{w\in W}\sum_{k\in R_w}\Lambda^{k}_{(i,j)}h_k,\forall(i,j)\in L\\&h_k\geq 0,\forall k\in R_w,w\in W\\&\sum_{(i,j)\in L}q_{(i,j)}=1\end{aligned}$$

According to (18), it implies

$$\underset{f}{Min}\sum_{(i,j)\in L}\int_{0}^{f_{(i,j)}}c_{(i,j)}(x,\Psi)dx+\sigma_c\sum_{(i,j)\in L}f_{(i,j)}r_{(i,j)}q^{M}_{(i,j)} \tag{20}$$

subject to

$$\sum_{k \in R_w} h_k = T_w, \forall w \in W$$
$$f_{(i,j)} = \sum_{w \in W} \sum_{k \in R_w} \Lambda^k_{(i,j)} h_k, \forall (i,j) \in L$$
$$h_k \geq 0, \forall k \in R_w, w \in W$$

The optimization problem (20) can be equivalently expressed as a following variational inequality: let

$$\tilde{c}(f', \Psi) = c(f', \Psi) + \sigma_c r q^M \tag{21}$$

for every traffic flow $f' \in \Omega$, it is to find a traffic flow with an obnoxious incident probability q^M such that

$$\tilde{c}(f, \Psi)(f' - f) \geq 0 \tag{22}$$

Let Σ_Q denote a solution set for (22).

2.3 The Upper Level Problem

At the upper level, a bi-objective signal control optimization with minimization of total travel delay and mitigation of total risk can be represented as follows. Let P be a performance index for signal settings and traffic flow, we thus have

$$\underset{\Psi \in \Pi, f \in \Sigma_Q}{Min} P = P_0(\Psi, f) \tag{23}$$

In (23), the set Π defines the constraints of signal settings. For instance, the cycle time constraint can be expressed as

$$\zeta_{\min} \leq \zeta \leq \zeta_{\max} \tag{24}$$

For each signal controlled junction m, the phase a green time for all signal groups at junction m is expressed as

$$\lambda_{\min} \zeta \leq \phi_{am} \leq 1 \tag{25}$$

The link capacity for all links leading to junction m is expressed as

$$f_{(i,j)} \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)} \tag{26}$$

The clearance time τ_{abm} for incompatible signal groups a and b at junction m is expressed as

$$\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a,b) \tag{27}$$

Therefore a big data risk-averse signal control optimization (RISCO) can be expressed as a following equilibrium-based network design model (EQ-based): for every flow f, which is solved by the WSM (20), we have

$$\underset{\Psi, f \in \Sigma_Q}{Min}\ P = \sum_{(i,j)\in L} D_{(i,j)}(\Psi, f) W_D M_D + S_{(i,j)}(\Psi, f) W_S M_S + \sigma \sum_{(i,j)\in L} f_{(i,j)} r_{(i,j)} \tag{28}$$

subject to

$$\begin{aligned}
&\zeta_{\min} \leq \zeta \leq \zeta_{\max} \\
&\lambda_{\min}\zeta \leq \phi_{am} \leq 1 \\
&f_{(i,j)} \leq \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}, \forall (i,j) \in L \\
&\theta_{am} + \phi_{am} + \tau_{abm}\zeta \leq \theta_{am} + \Omega_m(a,b)
\end{aligned}$$

3 A Bundle-like Method

Bundle methods are the most promising methods for non-smooth optimization [31, 34, 36, 37, 40]. The information around current iterate together with the past are collected and accumulated in bundles in order to establish a polyhedral approximation of the objective function at the actual iterate. Because of implicit form of equilibrium constraints, a general bi-level problem is usually not a convex program as widely mentioned in the literature [21–24, 41]. Due to the kinky structure of the performance function in RISCO, the EQ-based model (28) for single iterate could be possibly not precise for approximating the performance function for RISCO. Thus more information collected around current signal settings will be mobilized to acquire a more reliable model for RISCO. While non-smooth optimization problem can be effectively solved by the bundle methods as noticed in the literature [29, 31, 32, 36], applications of bundle methods to bi-level network design have not yet received much attention in the field of computer sciences and operations research.

Most of the existing algorithms for non-smooth optimization fall in the class of the sub-gradient and space dilatation algorithms [40], or of the bundle methods [34, 36]. In particular, the bundle methods are based on the cutting plane (CP) method, where the convexity of the objective function is the fundamental assumption. In fact the extension of the CP method to the non-convex case is not straightforward. A basic observation is that in general the first order information does not provide any longer a lower approximation to the objective function, independently on the non-smoothness assumption. Therefore, the optimization of the cutting plane

approximation does not necessarily give an optimistic estimate of the obtainable reduction in the objective function. Furthermore the CP based model might even fail to interpolate the objective function at the points where its value is known. From most recent numerical developments in bundle methods [29, 30], it obviously indicated that the CP based bundle methods become much less computationally robust and relatively unreliable in most general cases as compared to alternatives for non-convex optimization problem.

3.1 A Cutting Plane (CP) Model

The cutting plane (CP) algorithm of [32, 39, 42] for (28) in the neighborhood of $\Psi^{(k)}$ can be built by employing a bundle of past sub-gradients $\left\{P(\Psi^{(i)}), z^{(i)}, \Psi^{(i)}; z^{(i)} \in \partial_\Psi P(\Psi^{(i)}), 1 \le i \le k\right\}$. The generalized gradient of EQ-based model (28) can be expressed whenever a sub-gradient $z \in \partial P(\Psi)$ exists.

$$\partial_\Psi P(\Psi^*) = co\left\{\lim_{k\to\infty} \nabla_\Psi P(\Psi^{(k)}) : \Psi^{(k)} \to \Psi^*, \nabla_\Psi P(\Psi^{(k)}) exists\right\} \tag{29}$$

And the generalized gradients can be detailed as follows.

$$\begin{aligned}\nabla_\Psi P = &(\nabla_\Psi D(\Psi, f) + \nabla_f D(\Psi, f)\nabla_\Psi f(\Psi))W_D M_D \\ &+ (\nabla_\Psi S(\Psi, f) + \nabla_f S(\Psi, f)\nabla_\Psi f(\Psi))W_S M_S + \sigma\nabla_\Psi f(\Psi) r\end{aligned} \tag{30}$$

In (30), the directional derivatives for traffic conditions can be referred to [28]. The directional derivatives of responding flow in (22) can be found by solving a following linearized variational inequality according to recent developed results in [43, 44]. Introduce

$$\Omega(\Delta) = \{\nabla f : \nabla f = \Lambda(\Delta h), \Gamma(\Delta h) = 0, \exists \Delta h \in K_0\} \tag{31}$$

and

$$K_0 = \left\{ \Delta h : \begin{array}{llll} (i)\Delta h_k \quad free, & if & h_k^* > 0, & \\ (ii)\Delta h_k \ge 0, & if & h_k^* = 0, \tilde{C}_k = \pi_w, & \forall k \in R_w, \forall w \in W \\ (iii)\Delta h_k = 0, & if & h_k^* = 0, \tilde{C}_k > \pi_w & \end{array} \right\} \tag{32}$$

In (32) π_w denote minimum travel cost for trip w. For every flow perturbation in (31), i.e.$f' \in \Omega(\Delta)$, a directional derivative ∇f along a direction Δ in signal settings Ψ^* can be determined such that

$$(\nabla_\Psi \tilde{c}(f, \Psi^*)\Delta + \nabla_f \tilde{c}(f, \Psi^*)\nabla f)(f' - \nabla f) \ge 0 \tag{33}$$

The gradients $\nabla_{\Psi}\tilde{c}(f,\Psi^{*})$ and $\nabla_{f}\tilde{c}(f,\Psi^{*})$ in (33) are evaluated at Ψ^{*} when perturbations in Δ are specified. Therefore for a direction $\Delta^{(k)}$ in the neighborhood of $\Psi^{(k)}$, the directional derivative $DP(\Psi^{(k)};\Delta^{(k)})$ for any sub-gradient $z\in\partial P(\Psi^{(k)})$ can be calculated as follows.

$$DP(\Psi^{(k)};\Delta^{(k)})=z\Delta^{(k)} \tag{34}$$

Let $P^{(k)}$ denote a linear approximation of $P(\Psi)$ close to $\Psi^{(k)}$ at iteration i, $1\le i\le k$ we have

$$P^{(k)}\approx \underset{1\le i\le k}{Max}\left\{z^{(i)}(\Psi-\Psi^{(i)})+P(\Psi^{(i)})\right\} \tag{35}$$

Let

$$\varepsilon_{i,k}=P(\Psi^{(k)})-(P(\Psi^{(i)})+z^{(i)}(\Psi^{(k)}-\Psi^{(i)})) \tag{36}$$

denote an error bound for a linear approximation of $P(\Psi^{(k)})$. A cutting plane model $\hat{P}^{(k)}$ for a linear approximation of $P^{(k)}$ in (35) can be established in terms of bundle gradients, i.e.

$$\hat{P}^{(k)}=\underset{1\le i\le k}{Max}\left\{z^{(i)}(\Psi-\Psi^{(k)})-\varepsilon_{i,k}\right\}+P(\Psi^{(k)}) \tag{37}$$

Therefore the EQ-based model (28) can be approximated as a following cutting plane (CP) model $\hat{P}^{(k)}$.

$$\underset{\Psi}{Min}\,\hat{P}^{(k)}=P(\Psi^{(k)})+\underset{1\le i\le k}{Max}\left\{z^{(i)}(\Psi-\Psi^{(k)})-\varepsilon_{i,k}\right\} \tag{38}$$

subject to

$$\begin{aligned}
&\zeta_{\min}\le\zeta\le\zeta_{\max}\\
&\lambda_{\min}\zeta\le\phi_{am}\le 1\\
&f_{(i,j)}(\Psi)\le\rho_{(i,j)}s_{(i,j)}\lambda_{(i,j)},\forall(i,j)\in L\\
&\theta_{am}+\phi_{am}+\tau_{abm}\zeta\le\theta_{am}+\Omega_{m}(a,b)
\end{aligned}$$

Due to the instability of CP, in (38) the minimization of $\hat{P}^{(k)}$ that may give us very slow convergence as commented in the literature [29, 34–36]. Bundle methods are refinements of the classical CP method. A comprehensive treatment of the bundle methods can be found in [31, 34]. The idea is to maintain a stability center or a, that is, to distinguish one of the iterates generated so far. The is updated every time a significantly better solution was found. Roaming away from current is

penalized. The bundle methods thus reduce the influence of the inaccuracy of the CP approximation and therefore reducing instability of the CP method. Following recent numerical developments in non-smooth optimization [29, 31, 32, 38, 40, 45–47], the proximal bundle method (PBM) was recommended to provide a most efficient way to reliably solve the minimization of $\hat{P}^{(k)}$.

3.2 A Proximal Bundle Method (PBM)

According to [38, 45], the proximal bundle method for (38) can be expressed below by adding a quadratic term via a predetermined parameter t_k.

$$\underset{\Psi}{Min}\ \hat{P}^{(k)} + \frac{1}{2} t_k (\Psi - \Psi^{(k)})^2 \tag{39}$$

subject to

$$\begin{aligned}
&\zeta_{\min} \le \zeta \le \zeta_{\max} \\
&\lambda_{\min}\zeta \le \phi_{am} \le 1 \\
&f_{(i,j)}(\Psi) \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}, \forall (i,j) \in L \\
&\theta_{am} + \phi_{am} + \tau_{abm}\zeta \le \theta_{am} + \Omega_m(a,b)
\end{aligned}$$

i.e.

$$\underset{\Psi}{Min}\ \underset{1 \le i \le k}{Max}\ P\left(\Psi^{(k)}\right) + z^{(i)}\left(\Psi - \Psi^{(k)}\right) - \varepsilon_{i,k} + \frac{1}{2} t_k \left(\Psi - \Psi^{(k)}\right)^2 \tag{40}$$

subject to

$$\begin{aligned}
&\zeta_{\min} \le \zeta \le \zeta_{\max} \\
&\lambda_{\min}\zeta \le \phi_{am} \le 1 \\
&f_{(i,j)}(\Psi) \le \rho_{(i,j)} s_{(i,j)} \lambda_{(i,j)}, \forall (i,j) \in L \\
&\theta_{am} + \phi_{am} + \tau_{abm}\zeta \le \theta_{am} + \Omega_m(a,b)
\end{aligned}$$

3.3 A Bounding Strategy

The EQ-based model (28) can be effectively solved by a PBM algorithm. A bounding strategy is developed to solve the EQ-based model (28) in the following steps.

Step 1. Start with initial signal settings $\Psi^{(k)}$ and set index $k = 1$. Set the stopping threshold $\varepsilon, \varepsilon \geq 0$.

Step 2. Solve a weighted sum model at lower level with signal setting $\Psi^{(k)}$.

- Step 2.1. Find a most occurrence of obnoxious incidents with probability via (17).
- Step 2.2. Find the WSM based equilibrium flow f via (20).
- Step 2.3. Characterize the generalized cost via (21).

Step 3. Find directional derivatives $\nabla f\left(\Psi^{(k)}\right)$ via (33).

Step 4. Determine generalized gradient $\nabla_{\Psi} P\left(\Psi^{(k)}\right)$ via (30).

Step 5. Solve EQ-based model (28) via PBM algorithm.

- Step 5.1. Solve a cutting plane model $\hat{P}^{(k)}$ via (38).
- Step 5.2. Find a PBM updated $\Psi^{(k+1)}$ via (40).
- Step 5.3. Calculate a tentative direction $\Delta^{(k)} = \Psi^{(k+1)} - \Psi^{(k)}$.
- Step 5.4. Compute directional derivative $DP(\Psi^{(k)}; \Delta^{(k)})$ via (34). If $DP(\Psi^{(k)}; \Delta^{(k)}) > -\varepsilon$, stop and $\Psi^{(k)}$ is the solution for EQ-based model (28); otherwise move iteration k to $k + 1$ and go back to Step 2.

4 Numerical Computations

Numerical computations were made with solvers threefold using a real data benchmark problem of aggregated Sioux Falls city network [48] and grid-size general road networks as shown in Figs. 1, 2 and 3. In the first place, we compare with three benchmarks commonly employed for hazmat network design problem: CM [26], MM [16] and MM2 [18]. A real data example road network [48] was employed where 9 signal-controlled junctions were considered at two initial data sets, as shown in Fig. 1. In the second place, we compare CPU time elapsed using the same road network with 16 signal-controlled junctions at 10 sets of initial data, as shown in Fig. 2. Numerical computations were performed with a variety of weights between risk and cost. In the third place, we compare with continuous network design algorithms like LMILP [23] and PMC [24] using grid-size general road networks, as shown in Fig. 3. The performance index (PI) used in EQ-based model (28) was calculated on the basis of [28]. The travel time functions for

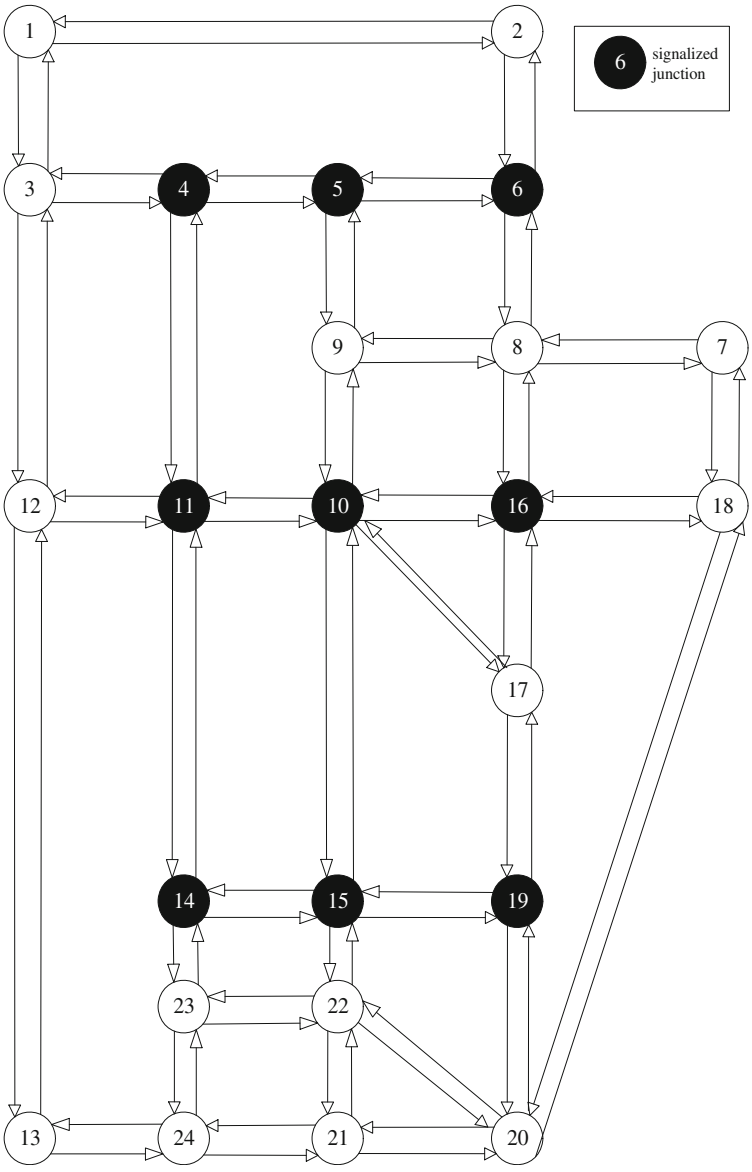


Fig. 1 Sioux Falls real-data test network [48]

non-signal controlled links were adopted from [48]. The average travel speed for hazmat traffic on link is supposed to be 15 m/s. Thus the average length on a link can be calculated as the product of average speed and cruise travel time along the link. Assuming incidental consequence of accidental release of hazmat is measured by average population exposure along a link, the population exposure is set 10 units

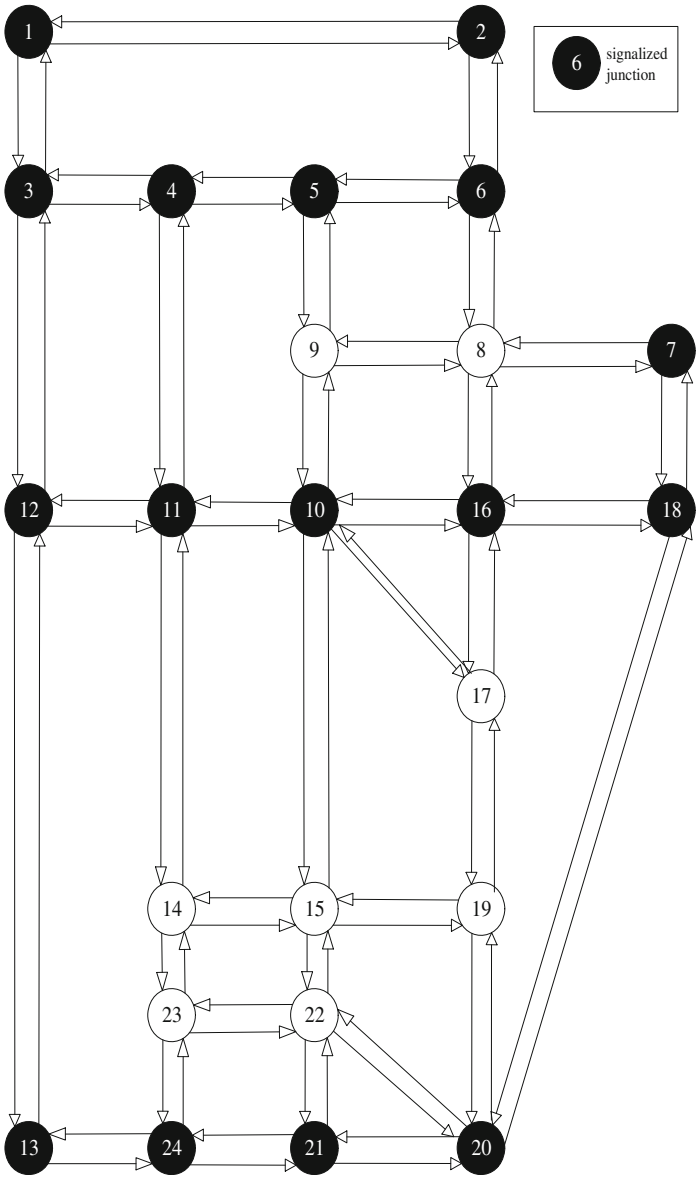


Fig. 2 Sioux Falls real-data test network [48]

per 300 m. Therefore the incidental consequence of accidental release of hazmat along a link can be computed accordingly. Using typical values found in practice, the minimum green time for each signal-controlled group is 7 s, and the clearance times are 5 s between incompatible signal groups. The maximum cycle time is set at 180 s. Implementations were performed on DELL T7610, Intel Xeon 2.5 GHz

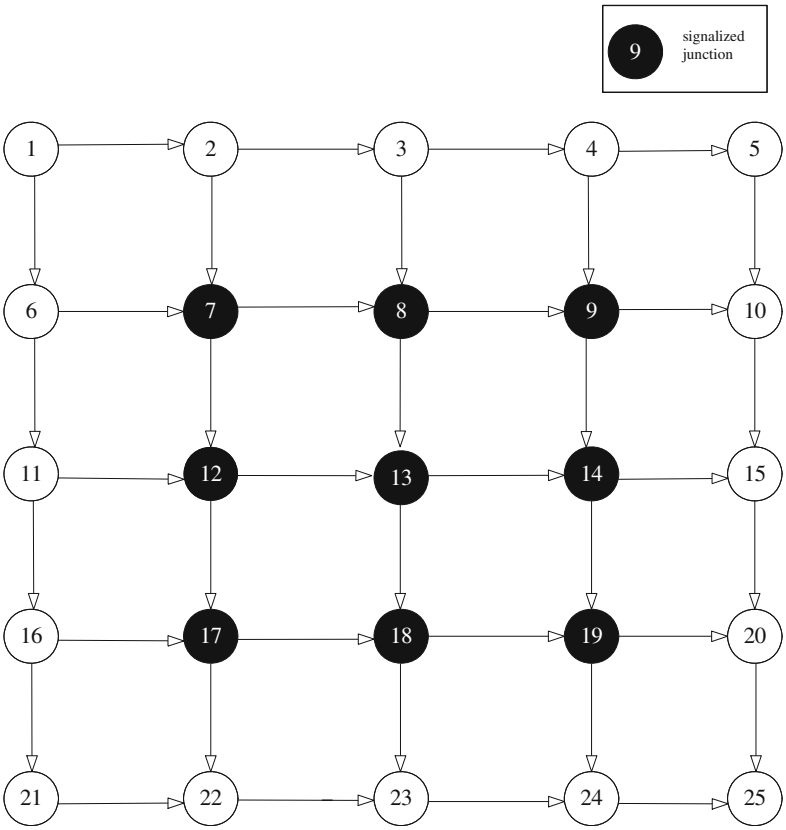


Fig. 3 25-node grid network

processor with 32 GB RAM under Windows 7 OS using C++ compiler. The stopping criterion is set when the relative difference in the objective function value is less than 0.15 %.

4.1 The First Example Road Network

A real-data example road network was used for numerical experiments as shown in Fig. 1 [48]. Two initial data sets are given in Table 1. For EQ-based model (28), the weight between generalized travel cost and risk exposure is supposed to be fixed and the value of weight equals one throughout the first example. As shown in Table 1, the initial performance index in terms of PI, risk exposure and travel cost were respectively expressed the last three rows for two distinct initial data sets. Computation results for three benchmark solvers CM, MM and MM2 and EQ-based model were respectively reported in Tables 2, 3, 4, and 5. Computation

results in Tables 2 3, 4, and 5 also contain most commonly used non-smooth minimization algorithms: CP. In Table 2, CPU time elapsed in seconds required for each solver was revealed the last row. As shown in Table 2, EQ-based model required significantly less CPU time elapsed as compared to those did CM, MM and MM2. The resulting performance values in terms of PI, cost and risk were reported with each solver. Performance improvement rates over initials in Table 2 with each solver in terms of PI reduction, risk reduction and cost reduction were expressed as PI %, risk % and cost % the last 2–4 rows. As expected, both MM and MM2 provided the best risk exposure reduction 55.8 and 55.2 % followed by EQ-based model 52.9 % on the average with non-smooth optimization variants. CM provided the least 36.5 %. Due to the effect of trade-off between risk exposure and travel cost, on the other hand, negativity of cost reduction rate over initial reveals relative increment of travel cost for alternatives. As expected, CM provided a least cost increment rate 6.4 % followed by EQ-based model. PBM provided 9.9 % cost increment over initial. EQ-based model gave in Table 2 the most 36.9 % on the average for PI, followed by MM2 34.3 %. CM provided a least PI improvement 26.3 % in all alternatives followed by MM 32.5 %. Comparison results obtained of PBM with solvers were also made and briefly reported in Table 3. When compared to those did CM, the first four rows in Table 3 reported relative reduction rates with solvers in terms of PI%, risk%, cost%, and CPU time%. For example, PBM gave a most PI reduction rate 18 % whilst MM gave a least 8.4 %. On the risk reduction, as expected, MM exhibited a superior advantage 30.3 %, followed by MM2 and PBM 29.4 and 29.2 %, respectively. As expected, none of solvers provided a better solution for cost reduction than did CM. Although most solvers produced positive cost reduction as compared to that did CM, PBM gave a least 3.3 % whilst MM provided a largest 33 %, followed by MM2 24.3 %. Moreover, PBM exhibited a superiority on reduction of CPU time elapsed nearly 90 %, whilst MM produced a positive 10.5 %. Similarly, when compared to MM in Table 3 the middle four rows reported the relative reductions of PI, risk, cost and CPU time elapsed with solvers in terms of PI%, risk%, cost%, and CPU time%. PBM exhibited a superior advantage on reduction of PI 10.5 %. By contrast, CM produced a positive 9.2 %, which apparently indicates that MM is superior to CM. Moreover, as expected, none of solvers provided a comparable solution as did as MM when reducing travel cost. Although most solvers gave positive values of risk reduction as compared to that did MM, PBM gave a slightly less 1.6 % following a comparative MM2 1.4 %. Again, PBM exhibited a superior advantage on CPU time whilst CM produced a least 9.5 %. Finally, when compared to MM2 in Table 3 the last four rows reported that PBM gave a superior PI improvement 8 % over all other solvers. However, both CM and MM induced positive 12.2 and 2.8 %, respectively. As expected, none of solvers provided a comparable solution for risk reduction but MM. Although most solvers gave rise to positive risk reductions, PBM produced a least 0.2 % while CM gave a largest 41.6 %. Again, PBM exhibited a most significant CPU time reduction 90 %. By contrast, both MM and CM gave positive 18 and 6.8 %, respectively.

Table 1 Initial data for Sioux Falls city network

Signal settings	1st set	2nd set
$\phi_{1,4}/\zeta$	25	45
$\phi_{2,4}/\zeta$	25	45
$\phi_{1,5}/\zeta$	25	45
$\phi_{2,5}/\zeta$	25	45
$\phi_{1,6}/\zeta$	25	45
$\phi_{2,6}/\zeta$	25	45
$\phi_{1,10}/\zeta$	25	45
$\phi_{2,10}/\zeta$	25	45
$\phi_{1,11}/\zeta$	25	45
$\phi_{2,11}/\zeta$	25	45
$\phi_{1,14}/\zeta$	25	45
$\phi_{2,14}/\zeta$	25	45
$\phi_{1,15}/\zeta$	25	45
$\phi_{2,15}/\zeta$	25	45
$\phi_{1,16}/\zeta$	25	45
$\phi_{2,16}/\zeta$	25	45
$\phi_{1,19}/\zeta$	25	45
$\phi_{2,19}/\zeta$	25	45
$1/\zeta$	60	100
PI (in$)	5500	5476
Risk (in $)	4184	4151
Cost (in $)	1316	1325

where $1/\zeta$ and ϕ_{jm}/ζ respectively denote the common cycle time and green durations measured in seconds

Similar results for the second data set were also reported in Tables 4 and 5. As shown in Table 4, EQ-based model required much less CPU time as compared to those did CM, MM and MM2. As expected, again, MM and MM2 provided the most reduction of risk exposure 55.5 % and 55 %, respectively, followed by EQ-based model 52.6 % on the average. CM provided the least 37.4 %. Due to the effect of trade-off between risk exposure and travel cost, as widely noticed in the literature [4, 49–53], CM provided a least cost increment 4.9 % followed by EQ-based model 15.2 %. PBM produced 9.4 % cost increment over initial. Moreover, EQ-based model gave the best PI improvement 36.2 % on the average, followed by MM2 and MM 34.1 and 32.5 %, respectively. CM provided a least 27.2 %. As compared to those obtained for the 1st data set, EQ-based model generated the least relative difference in all performance improvement between two distinct data sets while CM provided the largest. Similarly, Table 5 reported corresponding results in terms of PI%, risk%, cost% and CPU time% with solvers at the 2nd data set for the purpose of comparing relative improvement over CM, MM and MM2. PBM in the first four rows gave a most PI improvement 16.4 % as compared to those did CM whilst MM gave a least 7.3 %. MM exhibited a superior reduction

Table 2 Computation results for 1st data set at 1st example network

	CM	MM	MM2	EQ-based	
				CP	PBM
$\phi_{1,4}/\xi$	56	58	40	55	58
$\phi_{2,4}/\xi$	54	64	40	55	54
$\phi_{1,5}/\xi$	55	60	42	56	55
$\phi_{2,5}/\xi$	55	62	38	54	57
$\phi_{1,6}/\xi$	57	57	37	55	56
$\phi_{2,6}/\xi$	53	65	43	55	56
$\phi_{1,10}/\xi$	54	65	41	60	57
$\phi_{2,10}/\xi$	56	57	39	50	55
$\phi_{1,11}/\xi$	55	61	35	58	54
$\phi_{2,11}/\xi$	55	61	45	52	58
$\phi_{1,14}/\xi$	60	58	40	56	62
$\phi_{2,14}/\xi$	50	64	40	54	50
$\phi_{1,15}/\xi$	50	59	40	58	50
$\phi_{2,15}/\xi$	60	63	40	52	62
$\phi_{1,16}/\xi$	55	58	38	55	55
$\phi_{2,16}/\xi$	55	64	42	55	57
$\phi_{1,19}/\xi$	58	59	44	54	59
$\phi_{2,19}/\xi$	52	63	36	56	53
$1/\xi$	120	132	90	120	122
PI (in $)	4055	3715	3615	3618	3325
Risk (in $)	2655	1850	1875	2044	1879
Cost (in $)	1400	1865	1740	1574	1446
PI %	26.3	32.5	34.3	34.2	39.5
Risk %	36.5	55.8	55.2	51.1	55.1
Cost %	−6.4	−41.7	−32.2	−19.6	−9.9
CPU (in s)	267	295	250	51	27

rate 28.9 % on risk exposure, followed by MM2 and PBM 28.1 and 27.6 %, respectively. Although none of solvers can produce a better one than did CM on cost reduction, PBM gave a rather comparable one 4.3 %. Also, PBM exhibited a most CPU time reduction 90.3 %, followed CP 80.9 %. When compared to MM, PBM produced a significant improvement 9.9 % on PI reduction. On the contrary, CM gave a positive 7.8 %, which again indicates that MM works better than does CM when reducing PI. Although none of solvers can provide a comparable risk reduction as MM, PBM produced a rather amiable one following a comparable MM2. PBM also gave a superior CPU time elapsed reduction around 91 % as compared to that did MM. Finally, when compared to those did MM2, again, PBM

Table 3 Comparisons of benchmarks at 1st data set for 1st example network

	CM	MM	MM2	EQ-based	
				CP	PBM
PI %	–	−8.4	−10.9	−10.8	−18.0
Risk %	–	−30.3	−29.4	−23.0	−29.2
Cost %	–	33.2	24.3	12.4	3.3
CPU time %	–	10.5	−6.4	−80.9	−89.9
PI %	9.2	–	−2.7	−2.6	−10.5
Risk %	43.5	–	1.4	10.5	1.6
Cost %	−24.9	–	−6.7	−15.6	−22.5
CPU time %	−9.5	–	−15.3	−82.7	−90.8
PI %	12.2	2.8	–	0.1	−8.0
Risk %	41.6	−1.3	–	9.0	0.2
Cost %	−19.5	7.2	–	−9.5	−16.9
CPU time %	6.8	18.0	–	−79.6	−89.2

gave a best PI improvement 7.7 %. Although none of solvers but MM provided a less risk reduction as compared to MM2, PBM produced a fairly comparable one within marginal value 0.7 %. Moreover, PBM exhibited a superior CPU time reduction 90 %. On the contrary, both MM and CM gave positive CPU time elapsed around 16 and 7 %, respectively.

4.2 The Second Example Road Network

We compare numerical performance with CP using a real-data example road network as shown in Fig. 2. Ten sets of initial data are taken into account. Computation results obtained are summarized and reported in Tables 6, 7 and 8 with solvers considered above. A variety of weights in the interval of [0.05, 1.5] were considered between risk exposure and travel cost in EQ-based model (28). Table 6 summarizes the computation results obtained with CP and PBM in terms of risk exposure, travel cost and PI values together with CPU time, averaged over 10 sets of initial data for various weights between risk and cost in (28). As reported in Table 6, PBM exhibited a highly competitive advantage both in solution robustness and computation time efficiency in all cases. Indeed, an average time used to solve a second example with PBM along various weights was 49 s. As expected, the risk exposure was gradually reduced as more weight between risk exposure and travel cost was taken into account. On the other hand, total travel cost and accordingly the PI values were increased due to the effect of trade-off between cost and risk exposure. To investigate solution robustness and reliability of PBM, we compare performance improvement over that of CP in terms of risk reduction rate, cost

Table 4 Computation results for 2nd data set at 1st example network

	CM	MM	MM2	EQ-based	
				CP	PBM
$\phi_{1,4}/\zeta$	55	55	57	60	41
$\phi_{2,4}/\zeta$	57	65	55	60	41
$\phi_{1,5}/\zeta$	58	58	56	59	44
$\phi_{2,5}/\zeta$	54	62	56	61	38
$\phi_{1,6}/\zeta$	57	55	58	56	37
$\phi_{2,6}/\zeta$	55	65	54	64	45
$\phi_{1,10}/\zeta$	57	65	57	64	43
$\phi_{2,10}/\zeta$	55	55	55	56	39
$\phi_{1,11}/\zeta$	54	60	54	60	36
$\phi_{2,11}/\zeta$	58	60	58	60	46
$\phi_{1,14}/\zeta$	62	59	62	58	42
$\phi_{2,14}/\zeta$	50	61	50	62	40
$\phi_{1,15}/\zeta$	50	57	62	59	40
$\phi_{2,15}/\zeta$	62	63	50	61	42
$\phi_{1,16}/\zeta$	55	56	55	61	39
$\phi_{2,16}/\zeta$	57	64	57	59	43
$\phi_{1,19}/\zeta$	59	60	60	58	46
$\phi_{2,19}/\zeta$	53	60	52	62	36
$1/\zeta$	122	130	122	130	92
PI (in $)	3988	3698	3609	3615	3332
Risk (in $)	2598	1848	1868	2051	1882
Cost (in $)	1390	1850	1741	1564	1450
PI %	27.2	32.5	34.1	34.0	39.2
Risk %	37.4	55.5	55.0	50.6	54.7
Cost %	−4.9	−39.6	−31.4	−18.0	−9.4
CPU (in s)	267	291	250	51	26

reduction rate, CPU time reduction rate, and PI reduction rate with those of PBM. For instance, the risk % for solver x denotes percent relative difference ratio in risk exposure between x and CP. Taking negative value of risk % means that there is a reduction rate in risk exposure for x over CP. Tables 7 and 8 give details obtained for each instance with PBM in the sense of relative performance improvement when compared to those of CP. Results are displayed in four columns, with the risk reduction rate: risk %, the cost reduction rate: cost %, the CPU time reduction rate: CPU %, and the PI reduction rate: PI %, respectively. Numerical comparisons were also detailed in Tables 7 and 8 for weights varying from 0.05 to 1.5 with PBM at 10 sets of initial instances. As shown in Table 7, for the first group, i.e. $\sigma = 0.05$, PBM

Table 5 Comparisons of benchmarks at 2nd data set for 1st example network

	CM	MM	MM2	EQ-based	
				CP	PBM
PI %	–	−7.3	−9.5	−9.4	−16.4
Risk %	–	−28.9	−28.1	−21.1	−27.6
Cost %	–	33.1	25.3	12.5	4.3
CPU time %	–	9.0	−6.4	−80.9	−90.3
PI %	7.8	–	−2.4	−2.2	−9.9
Risk %	40.6	–	1.1	11.0	1.8
Cost %	−24.9	–	−5.9	−15.5	−21.6
CPU time %	−8.2	–	−14.1	−82.5	−91.1
PI %	10.5	2.5	–	0.2	−7.7
Risk %	39.1	−1.1	–	9.8	0.7
Cost %	−20.2	6.3	–	−10.2	−16.7
CPU time %	6.8	16.4	–	−79.6	−89.6

Table 6 Summary results for solvers at 2nd example network

Algorithm	σ	Risk (in $)	Cost (in $)	CPU (in s)	PI (in $)
CP	0.05	2746.0	1178.0	101.9	1315.3
	0.25	2634.3	1348.5	103.3	2007.1
	0.5	2545.3	1383.5	104.4	2656.2
	0.75	2464.8	1403.7	105.2	3252.3
	1.0	2244.6	1430.3	105.7	3674.9
	1.25	2178.0	1503.5	106.9	4226.0
	1.5	2128.4	1557.1	107.5	4749.7
	Average	**2420.2**	**1400.7**	**105.0**	**3125.9**
PBM	0.05	2663.8	1053.6	48.7	1186.8
	0.25	2513.3	1213.9	48.7	1842.2
	0.5	2402.0	1256.6	48.7	2457.6
	0.75	2297.5	1274.5	48.9	2997.6
	1.0	2076.1	1301.8	48.8	3377.9
	1.25	1997.5	1373.2	48.8	3870.1
	1.5	1944.0	1421.7	49.0	4337.7
	Average	**2270.6**	**1270.8**	**48.8**	**2867.1**

was good both in PI and CPU time elapsed reduction. Averaged over 10 sets of instances, the results obtained with PBM were promising, which apparently indicates that PBM has better potential solving EQ-based model as σ increases. In particular, PBM exhibited a significant decrease 52.2 % on the average in CPU time elapsed. PBM also exhibited a promising decrease 9.8 % on the average in PI. Similarly, for the second group, i.e. $\sigma = 0.25$, PBM, again, exhibited a significant decrease rate 52.9 % on the average in CPU time elapsed. Numerical results

Table 7 Improvement rate of bundle-like solvers at 2nd example network

σ	Instance	PBM			
		Risk %	Cost %	CPU %	PI %
0.05	1	−2.9	−10.4	−52.4	−9.7
	2	−3.0	−10.5	−51.0	−9.7
	3	−3.0	−10.7	−53.4	−9.9
	4	−2.9	−10.7	−52.4	−9.9
	5	−3.0	−10.4	−51.5	−9.7
	6	−3.1	−10.1	−52.9	−9.4
	7	−3.1	−10.3	−52.5	−9.5
	8	−2.8	−10.8	−51.0	−9.9
	9	−3.0	−10.6	−51.5	−9.8
	10	−3.0	−11.0	−53.4	−10.2
	Average	**−3.0**	**−10.6**	**−52.2**	**−9.8**
0.25	1	−4.5	−9.9	−52.4	−8.2
	2	−4.6	−10.4	−51.9	−8.5
	3	−4.7	−10.1	−53.4	−8.4
	4	−4.4	−9.9	−52.4	−8.1
	5	−4.6	−10.2	−52.4	−8.4
	6	−4.6	−10.2	−52.9	−8.3
	7	−4.6	−9.7	−53.4	−8.0
	8	−4.6	−9.4	−52.9	−7.8
	9	−4.6	−9.9	−52.4	−8.2
	10	−4.7	−10.0	−54.3	−8.3
	Average	**−4.6**	**−10.0**	**−52.9**	**−8.2**
0.5	1	−5.6	−9.3	−53.3	−7.5
	2	−5.4	−8.9	−51.9	−7.2
	3	−5.7	−8.9	−53.4	−7.3
	4	−5.5	−9.4	−52.9	−7.5
	5	−5.5	−8.9	−52.4	−7.3
	6	−5.7	−9.0	−54.3	−7.4
	7	−5.8	−9.2	−53.4	−7.6
	8	−5.7	−9.3	−52.9	−7.6
	9	−5.7	−9.3	−53.8	−7.5
	10	−5.6	−9.6	−55.1	−7.7
	Average	**−5.6**	**−9.2**	**−53.3**	**−7.5**
0.75	1	−6.7	−9.1	−52.9	−7.7
	2	−6.9	−9.1	−51.9	−7.9
	3	−7.1	−9.1	−54.7	−7.9
	4	−6.7	−9.2	−51.9	−7.8
	5	−6.7	−9.7	−52.9	−8.0
	6	−6.6	−9.2	−54.7	−7.7

(continued)

Table 7 (continued)

σ	Instance	PBM			
		Risk %	Cost %	CPU %	PI %
	7	−6.8	−9.5	−54.3	−7.9
	8	−6.8	−9.0	−53.8	−7.7
	9	−6.8	−9.2	−52.8	−7.8
	10	−6.8	−9.0	−55.1	−7.7
	Average	**−6.8**	**−9.2**	**−53.5**	**−7.8**

obtained for cost and risk reduction rates 10 and 4.6 % with PBM, as shown in Table 7. Similar results can be observed in Table 7 for groups 3–7 when weights range from 0.5 to 1.5. Numerical comparisons with PBM along varying weights were plotted in Figs. 4, 5, 6 and 7, respectively for performance improvement in the sense of risk %, cost %, PI % and CPU %. As seen from Figs. 4, 5, 6 and 7, PBM enjoyed greater performance improvements over those did CP uniformly in all cases.

4.3 The Third Example Road Network

We compare system performance in the sense of risk exposure, total travel cost, CPU time in seconds, and PI values of PBM with continuous network design problem (CNDP) solvers: LMILP [23] and PMC [24] using general road networks of various grid-size. To investigate reliability of PBM when applying in large-size networks, numerical results were performed using varying weights between risk and cost considered in the second example. Table 9 summarizes system performance in the sense of risk exposure, travel cost, PI values and CPU time respectively for various grid-size general road networks with alternative solvers for RISCO. As observed in Table 9, the global optimum provided by LMILP from mixed transportation network design produced the least performance in all cases but CPU time. As expected, the solutions provided by LMILP are approximate for mixed transportation network design problem with linearized objective function. Employing such piecewise linearized functions for RISCO generally simplifies the structure of problem considered in this paper and not apparently appropriate for general network design problem like RISCO. Thus the results obtained in system performance from LMILP can be served as upper bounds for the purpose of computation comparisons. PMC was usually the most efficient solver for CNDP, it was also the one which needed the longest time to compute the CNDP [24]. As observed in Table 9, the global solution provided by PMC makes the most extensive use of CPU time due to PMC suffering from all-or-nothing traffic assignment each time a penalty term in PMC is updated. Indeed, an average time used to solve a third example of various grid-size networks with PMC was 1221 s while with LMILP, CP and PBM, they were 636, 138 and 59 s, respectively.

Table 8 Improvement rate of bundle-like solvers at 2nd example network (cont)

σ	Instance	PBM			
		Risk %	Cost %	CPU %	PI %
1.0	1	−7.5	−8.7	−52.8	−8.0
	2	−7.5	−9.1	−51.9	−8.1
	3	−7.8	−9.1	−54.7	−8.3
	4	−7.5	−9.1	−52.9	−8.1
	5	−7.5	−8.6	−53.8	−7.9
	6	−7.4	−9.0	−54.7	−8.0
	7	−7.6	−8.7	−54.3	−8.0
	8	−7.5	−9.4	−54.2	−8.2
	9	−7.4	−9.1	-53.8	−8.1
	10	−7.4	−9.1	−55.1	−8.1
	Average	**−7.5**	**−9.0**	**−53.8**	**−8.1**
1.25	1	−8.3	−8.5	−53.8	−8.4
	2	−−8.4	−8.8	−52.8	−8.6
	3	−8.6	−8.7	−55.1	−8.6
	4	−8.4	−8.6	−54.2	−8.4
	5	−−8.3	−9.0	−53.8	−8.6
	6	−8.2	−8.8	−55.1	−8.4
	7	−8.0	−8.7	−55.6	−8.3
	8	−7.9	−8.4	−53.3	−8.1
	9	−8.3	−8.5	−54.6	−8.4
	10	−8.4	−8.6	−55.1	−8.5
	Average	**−8.3**	**−8.7**	**−54.3**	**−8.4**
1.5	1	−8.6	−8.4	−55.0	−8.5
	2	−9.0	−8.7	−52.8	−8.9
	3	−8.8	−8.2	−55.1	−8.6
	4	−8.6	−9.0	−54.2	−8.7
	5	−8.9	−9.1	−52.8	−9.0
	6	−8.7	−8.4	−55.1	−8.6
	7	−8.5	−8.5	−55.6	−8.5
	8	−8.5	−9.3	−54.6	−8.7
	9	−8.6	−8.8	−54.6	−8.6
	10	−8.5	−8.6	−54.1	−8.6
	Average	**−8.7**	**−8.7**	**−54.4**	**−8.7**

Moreover, as observed in Table 9, apparently the CPU times elapsed with PMC and LMILP grow superlinearly in the size of grid networks while the growth of the computation times for CP is roughly linear in the size of grid networks. PBM, on the contrary, exhibits the most economic advantage on CPU time elapsed over alternatives. The results obtained with PBM serve the best of all alternatives in all

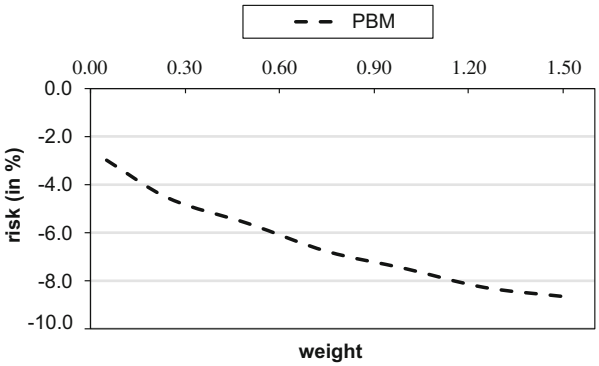


Fig. 4 Risk reduction rate of bundle-like solvers at 2nd example network

Fig. 5 Cost reduction rate of bundle-like solvers at 2nd example network

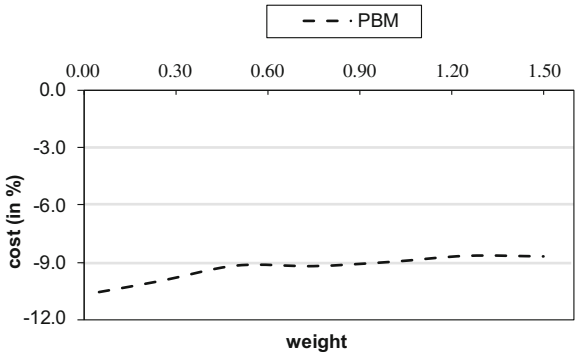
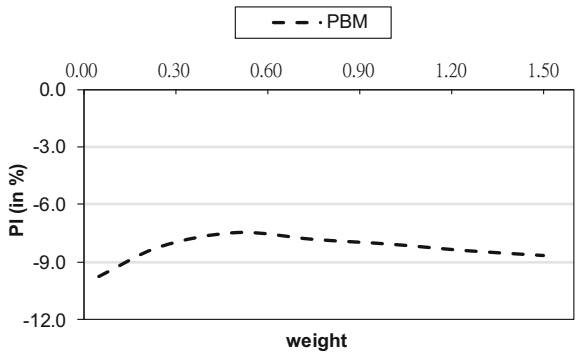


Fig. 6 PI reduction rate of bundle-like solvers at 2nd example network



cases. In particular, as observed in Table 9, PBM achieved significant superiority on CPU time elapsed as the size of general road networks grows. The growth of computation time with PBM was clearly sublinear in the size of grid networks and did not grow as fast as those required with CP when the size of grid networks was increased. Indeed, with large numbers of grid sizes, PBM needed significantly less

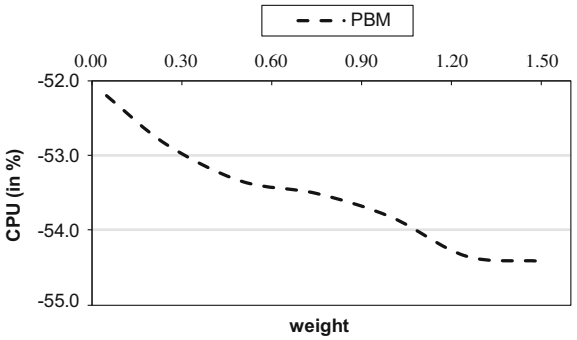


Fig. 7 CPU time reduction rate of bundle-like solvers at 2nd example network

Table 9 Summary results of comparative solvers at 3rd example network

Algorithm	Grid size	Risk (in $)	Cost (in $)	CPU (in s)	PI (in $)
LMILP	5 × 5	838.4	430.6	370.3	1043.3
	7 × 7	963.1	538.7	423.6	1251.2
	9 × 9	1147.7	723.6	517.4	1528.6
	13 × 13	1370.4	1012.1	858.4	1969.8
	15 × 15	1987.4	1226.0	1011.3	2576.1
	Average	**1261.4**	**786.2**	**636.2**	**1673.8**
PMC	5 × 5	815.1	426.0	614.1	1020.5
	7 × 7	942.1	536.4	708.9	1231.1
	9 × 9	1120.3	686.3	982.3	1468.8
	13 × 13	1340.0	970.1	1604.3	1899.1
	15 × 15	1930.9	1176.7	2193.9	2482.1
	Average	**1229.7**	**759.1**	**1220.7**	**1620.3**
CP	5 × 5	816.0	426.3	100.4	1021.5
	7 × 7	943.1	536.3	106.6	1231.7
	9 × 9	1122.0	688.1	120.1	1471.8
	13 × 13	1344.1	971.1	164.4	1903.7
	15 × 15	1934.1	1181.7	195.9	2490.0
	Average	**1231.9**	**760.7**	**137.5**	**1623.7**
PBM	5 × 5	790.0	414.6	38.4	986.8
	7 × 7	907.7	513.3	39.9	1178.1
	9 × 9	1076.7	650.1	44.0	1398.2
	13 × 13	1272.9	912.1	84.6	1791.0
	15 × 15	1803.7	1077.7	89.0	2287.6
	Average	**1170.2**	**713.6**	**59.2**	**1528.3**

CPU time elapsed than did the CP. As for the PI values, PBM provided the best results in all cases of all solvers. PMC and CP exhibited comparable advantage. Moreover, in this example also the risk exposure of PBM was the least of all solvers. The same trend can be seen in the travel time cost. The efficiency of PBM

is mostly due to its efficiency in piecewise quadratic problems. It was the most efficient solver in almost all piecewise quadratic problems when comparing the CPU time elapsed and superior when comparing the effectiveness of performance index of RISCO. Further details of comparisons for performance improvements can be observed in Tables 10 and 11 together with Figs. 8, 9, 10 and 11 in the sense of relative decrement in risk exposure, travel cost, CPU time, and PI values when compared to those did the LMILP.

Results in Tables 10 and 11 with solvers: CP, PMC, and PBM for weights ranging from 0.05 to 1.5 are displayed in four columns, with the risk reduction rate: risk %, the cost reduction rate: cost %, the CPU time reduction rate: CPU % and the PI reduction rate: PI %, respectively. As seen in Table 10, when compared to LMILP, CP achieved slightly less reduction than those of PMC in terms of risk %, cost % and PI % except CPU % as the size of grid network grows. In terms of CPU time, PMC was systematically slower than LMILP, a known fact in CNDP, due to all-or-nothing traffic assignments. Accordingly, PMC was incurred with rather more intensive CPU times elapsed than those of CP when compared to LMILP. As observed in Table 10, the average CPU time elapsed with PMC was increased from 66 to 80 % as compared to those of LMILP when the size of grid network grows from 5 * 5 to 15 * 15. On the contrary, CP exhibited a significant superiority on average CPU time reduction around 76 % as compared to those of LMILP. As observed in Table 11, the risk exposure reduction of PBM was steadily increased from 5.8 to 9.5 %. Also the travel cost reduction of PBM was significantly increased from 3.7 to 12.3 %. Accordingly, the PI value reduction of PBM was increased from 5.1 to 11.5 %. In particular, PBM achieved a significantly superior advantage on CPU time elapsed over those did alternatives: the reduction in CPU time with PBM was varied from 89.6 % to 94.22 %. Similar results obtained with all solvers when compared to LMILP were plotted in Figs. 8, 9, 10 and 11 in terms of risk %, cost %, CPU % and PI % along various sizes of general grid networks. We observed a good performance improvement for all solvers when compared to LMILP, with a significant advantage of PBM over the other comparable ones. Based on the numerical results, we can conclude the superiority of PBM when comparing the computation times: in all cases, they used significantly least CPU time of all solvers.

Moreover, the growth of computation time with PBM did not grow as fast as those required with LMILP, PMC, and CP when the size of general road networks was increased. Indeed, with large sizes of general grid road networks, PBM needed significantly less CPU time elapsed than did CP. Overall, PBM outperformed all other solvers for CNDP with all performance indicators in terms of risk exposure, travel cost, PI values and CPU time elapsed. In particular, PBM enjoyed greatest advantage on the CPU time elapsed over all other solvers, especially as the size of general grid network expands.

Table 10 Improvement rate of bundle-like solvers at 3rd example network

Grid size	σ	CP				PMC			
		Risk %	Cost %	CPU %	PI %	Risk %	Cost %	CPU %	PI %
5 × 5	0.05	−1.3	−1.0	−71.0	−1.0	−1.5	−1.4	95.9	−1.4
	0.25	−2.7	−0.9	−69.7	−1.6	−2.9	−1.2	84.6	−1.8
	0.5	−3.2	−0.9	−69.2	−2.0	−3.0	−1.2	85.0	−2.1
	0.75	−3.8	−0.9	−70.7	−2.6	−−3.7	−1.2	63.8	−2.7
	1.0	−2.3	−0.9	−76.2	−1.8	−2.6	−1.1	47.7	−2.1
	1.25	−2.5	−1.1	−76.7	−2.1	−2.8	−0.7	43.8	−2.1
	1.5	−3.0	−1.1	−74.7	−2.5	−3.1	−0.7	54.1	−2.4
	Average	**−2.7**	**−1.0**	**−72.6**	**−1.9**	**−2.8**	**−1.1**	**67.8**	**−2.1**
7 × 7	0.05	−1.1	−0.4	−75.3	−0.4	−1.5	0.0	73.8	−0.1
	0.25	−1.5	−0.6	−75.1	−0.9	−1.4	−0.4	71.1	−0.7
	0.5	−2.0	−0.6	−74.5	−1.3	−2.3	−0.4	61.9	−1.3
	0.75	−2.2	−0.6	−73.4	−1.5	−2.1	−0.7	71.0	−1.5
	1.0	−2.8	−0.4	−73.0	−1.9	−2.4	−0.6	69.9	−1.8
	1.25	−2.8	−0.4	−76.1	−2.0	−3.0	−0.5	70.8	−2.2
	1.5	−2.3	−0.4	−76.3	−1.7	−2.6	−0.4	54.3	−2.0
	Average	**−2.1**	**−0.5**	**−74.8**	**−1.4**	**−2.2**	**−0.4**	**67.6**	**−1.4**
9 × 9	0.05	−0.9	−5.4	−76.5	−5.0	−1.1	−5.5	89.9	−5.1
	0.25	−1.8	−5.2	−76.2	−4.1	−2.1	−5.3	86.8	−4.3
	0.5	−2.5	−4.2	−77.1	−3.4	−2.4	−4.5	87.5	−3.5
	0.75	−2.6	−4.7	−76.9	−3.5	−2.9	−5.1	87.6	−3.9
	1.0	−2.7	−5.0	−77.0	−3.7	−2.6	−5.2	93.0	−3.7
	1.25	−2.5	−5.0	−76.9	−3.5	−2.7	−5.3	93.4	−3.7
	1.5	−3.1	−4.9	−76.8	−3.7	−3.3	−5.2	90.7	−4.0
	Average	**−2.3**	**−4.9**	**−76.8**	**−3.8**	**−2.5**	**−5.2**	**89.8**	**−4.0**
13 × 13	0.05	−0.8	−3.6	−81.5	−3.3	−1.1	−3.4	98.3	−3.2
	0.25	−0.7	−3.4	−81.0	−2.6	−0.9	−3.2	89.7	−2.5
	0.5	−1.3	−2.9	−81.3	−2.2	−1.6	−3.2	87.0	−2.5
	0.75	−2.0	−4.1	−81.2	−3.1	−2.2	−4.5	87.1	−3.3
	1.0	−2.9	−5.4	−80.4	−4.0	−3.2	−5.3	84.5	−4.2
	1.25	−3.2	−4.6	−80.4	−3.8	−3.7	−4.8	81.7	−4.2
	1.5	−3.4	−4.1	−80.2	−3.6	−3.9	−4.2	80.9	−4.0
	Average	**−2.0**	**−4.0**	**−80.9**	**−3.2**	**−2.4**	**−4.1**	**87.0**	**−3.4**
15 × 15	0.05	−1.0	−5.1	−80.9	−4.7	−0.9	−5.7	117.2	−5.2
	0.25	−3.3	−5.4	−81.3	−4.7	−3.5	−5.5	117.3	−4.8
	0.5	−2.4	−4.4	−81.0	−3.4	−2.7	−4.2	116.9	−3.5
	0.75	−3.0	−2.8	−80.4	−3.0	−3.2	−3.3	116.0	−3.3
	1.0	−3.0	−2.8	−80.0	−2.9	−2.9	−3.4	119.2	−3.1
	1.25	−3.3	−2.5	−80.1	−3.0	−3.7	−3.0	116.7	−3.4
	1.5	−3.4	−3.0	−80.7	−3.2	−3.7	−3.7	115.3	−3.7
	Average	**−2.8**	**−3.7**	**−80.6**	**−3.6**	**−3.0**	**−4.1**	**116.9**	**−3.9**

Table 11 Improvement rate of bundle-like solvers at 3rd example network (cont)

Grid size	σ	PBM			
		Risk %	Cost %	CPU %	PI %
5 × 5	0.05	−3.3	−3.8	−85.7	−3.8
	0.25	−4.8	−3.8	−90.5	−4.1
	0.5	−5.4	−3.7	−88.0	−4.6
	0.75	−6.6	−3.7	−89.0	−5.4
	1.0	−6.7	−3.7	−90.8	−5.6
	1.25	−6.8	−3.6	−91.5	−5.8
	1.5	−7.3	−3.6	−90.7	−6.3
	Average	**−5.8**	**−3.7**	**−89.5**	**−5.1**
7 × 7	0.05	−3.4	−5.3	−88.6	−5.2
	0.25	−4.3	−4.9	−92.2	−4.7
	0.5	−5.5	−4.9	−90.3	−5.2
	0.75	−5.9	−4.8	−90.0	−5.4
	1.0	−6.7	−4.6	−90.7	−5.9
	1.25	−7.3	−4.4	−91.1	−6.4
	1.5	−7.6	−4.2	−91.2	−6.6
	Average	**−5.8**	**−4.7**	**−90.6**	**−5.6**
9 × 9	0.05	−4.2	−11.3	−90.1	−10.7
	0.25	−5.2	−10.9	−91.5	−9.1
	0.5	−5.9	−9.9	−91.0	−8.1
	0.75	−6.4	−9.9	−92.2	−8.0
	1.0	−6.7	−9.8	−92.0	−8.0
	1.25	−7.3	−9.8	−92.1	−8.3
	1.5	−8.6	−9.5	−91.7	−8.9
	Average	**−6.3**	**−10.2**	**−91.5**	**−8.7**
13 × 13	0.05	−5.1	−11.8	−90.3	−11.3
	0.25	−5.2	−10.6	−90.0	−9.0
	0.5	−6.6	−10.0	−90.1	−8.6
	0.75	−7.4	−9.8	−90.2	−8.6
	1.0	−8.5	−9.9	−90.0	−9.1
	1.25	−9.1	−9.1	−90.2	−9.1
	1.5	−9.3	−8.5	−90.1	−9.0
	Average	**−7.3**	**−10.0**	**−90.1**	**−9.2**
15 × 15	0.05	−6.1	−14.9	−91.6	−14.0
	0.25	−8.7	−14.6	−91.4	−12.6
	0.5	−8.9	−13.4	−91.0	−11.3
	0.75	−9.7	−11.4	−91.3	−10.5
	1.0	−10.4	−11.2	−90.9	−10.7
	1.25	−11.1	−10.4	−91.1	−10.8
	1.5	−11.5	−10.1	−91.0	−11.0
	Average	**−9.5**	**−12.3**	**−91.2**	**−11.5**

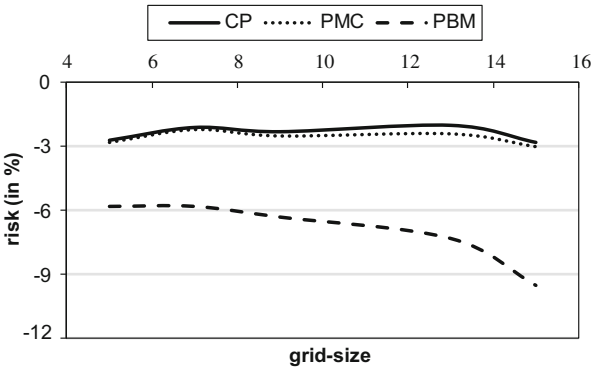


Fig. 8 Risk reduction rate of comparative solvers at 3rd example networks

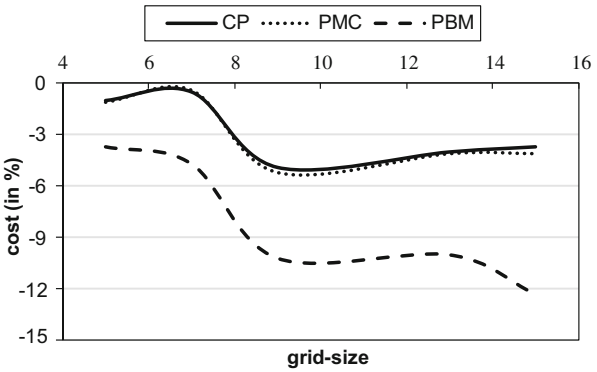


Fig. 9 Cost reduction rate of comparative solvers at 3rd example networks

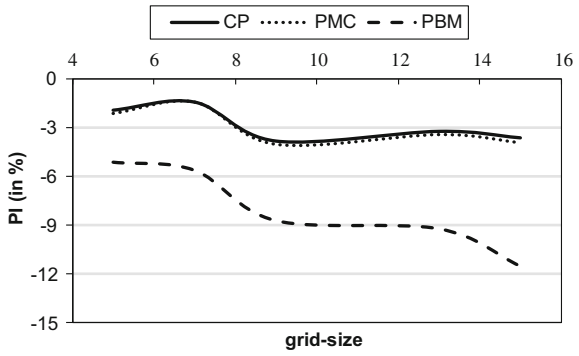


Fig. 10 PI reduction rate of comparative solvers at 3rd example networks

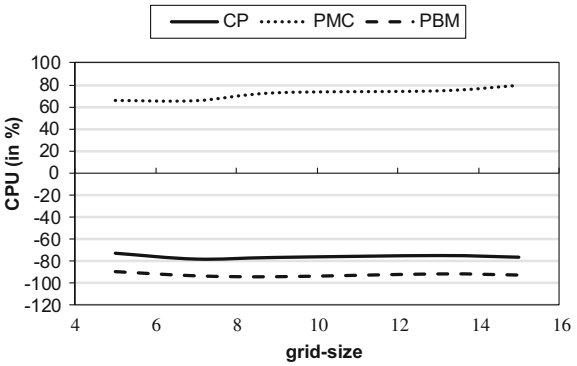


Fig. 11 CPU time reduction rate of comparative solvers at 3rd example networks

5 Conclusions and Discussions

We have presented an efficient algorithm PBM that is designed to work on a big data network design with risk-averse signal control optimization (RISCO). An equilibrium-based bi-level network design model (EQ-based) was presented to solve the RISCO. A weighted sum risk equilibrium model (WSM) was established to determine the maximum risk exposure at the lower level where Wardrop's first principle is respected. Due to kinky structure of RISCO, a bundle-like PBM algorithm was proposed to stabilize solutions of RISCO. PBM takes advantage of good properties of the utilization of sub-differential, aggregation of collective sub-gradients, and the sub-gradient locality measures to compute the search direction iteratively in a single-level problem. The proposed PBM appears to perform well in computation, as illustrated in Sect. 4 using three example road networks when compared to other solvers for RISCO. Particularly, numerical results show that PBM compares considerably favorably with two recent proposed algorithms PMC and LMILP for CNDP using general grid networks with large number of variables. Based on the numerical results, we can conclude the superiority of PBM when comparing the computation times: in all cases, they used significantly least CPU time of all solvers. Moreover, the growth of computation time with PBM did not grow as fast as those required with LMILP, PMC, and CP when the size of general road networks was increased. Indeed, with large sizes of general grid road networks, PBM needed significantly less CPU time elapsed than did CP. Overall, PBM outperformed all other solvers for CNDP with all performance indicators in terms of risk exposure, travel cost, PI values and CPU time elapsed. In particular, PBM enjoyed greatest advantage on the CPU time elapsed over all other solvers, especially as the size of general grid network expands.

Another advantage, not explored completely in this paper, is that PBM provides potential applicability in road network design with mixed decision variables. In particular, the proposed PBM can be applied to the problem of designing urban road networks in a multi-objective decision making framework where discrete decision variables like street orientation or link addition are taken into account as recently

studied in [54, 55]. Concerning the solution methods, as mentioned in the literature [56–58] that the combinatorial nature of the mixed network design or discrete network design problems makes them even much more difficult to solve than the CNDP considered in this paper. Applications of PBM to various types of big data network design problems and computation comparisons with recent proposed meta-heuristics in [54–56] are being investigated. We will discuss these issues of interest in subsequent papers.

Acknowledgements The author is grateful to editors for their kind comments in earlier version of this manuscript. The work reported here has been financially sponsored by Taiwan Science council via grant MOST 104-2221-E-259-029-MY3.

References

1. Erkut E, Alp O (2007) Designing a road network for hazardous materials shipments. Comp Oper Res 34:1389–1405
2. Erkut E, Gzara F (2008) Solving the hazmat transport network design problem. Comp Oper Res 35:2234–2247
3. Marcotte P, Mercier A, Savard G, Verter V (2009) Toll policies for mitigating hazardous materials transport risk. Transp Sci 43:228–243
4. Bianco L, Caramia M, Giordani S, Piccialli V (2013) Operations research models for global route planning in hazardous material transportation. In: Batta R, Kwon C (eds) Handbook of OR/MS models in hazardous materials transportation, International series in operations research & management science 193. Springer Science + Business Media, New York, pp 49–101
5. Gzara F (2013) A cutting plane approach for bilevel hazardous material transport network design. Oper Res Lett 41:40–46
6. Zhang X, Huang G (2013) Optimization of environmental management strategies through a dynamic stochastic possibilistic multiobjective program. J Hazard Mater 246–247:257–266
7. Zhao J, Verter V (2015) A bi-objective model for the used oil location-routing problem. Comp Oper Res 62:157–168
8. Gang J, Tu Y, Lev B, Xu J, Shen W, Yao L (2015) A multi-objective bi-level location planning problem for stone industrial parks. Comp Oper Res 56:8–21
9. Samanlioglu F (2013) A multi-objective mathematical model for the industrial hazardous waste location-routing problem. EJOR 226:332–340
10. Belhoul L, Galand L, Vanderpooten D (2014) An efficient procedure for finding best compromise solutions to the multi-objective assignment problem. Comp Oper Res 49:97–106
11. Angulo E, Castillo E, Garcia-Rodenas R, Sanchez-Vizcaino J (2014) A continuous bi-level model for the expansion of highway networks. Comp Oper Res 41:262–272
12. Kara BY, Verter V (2004) Designing a road network for hazardous materials transportation. Transp Sci 38:188–196
13. Veter V, Kara BY (2008) A path-based approach for hazardous transport network design. Manage Sci 54:29–40
14. Alp E (1995) Risk-based transportation planning practice: overall methodology and a case example. INFOR 33:4–19
15. Jin H, Batta R (1997) Objectives derived from viewing hazmat shipments as a sequence of independent Bernoulli trials. Transp Sci 31:252–261
16. Erkut E, Ingolfsson A (2000) Catastrophe avoidance models for hazardous materials route planning. Transp Sci 34:165–179

17. Erkut E, Ingolfsson A (2005) Transport risk models for hazardous materials: revisited. Oper Res Lett 33:81–89
18. Bell MGH (2007) Mixed routing strategies for hazardous materials: decision-making under complete uncertainty. Int J Sus Transp 1:133–142
19. Luo ZQ, Pang JS, Ralph D (1996) Mathematical programs with equilibrium constraints. Cambridge University Press, Cambridge, New York
20. Outrata J, Kocvara M, Zowe J (1998) Nonsmooth approach to optimization problems with equilibrium constraints. Kluwer Academic Publishers, Dordrecht, The Netherlands
21. Dempe S (2003) Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. Optimization 52:333–359
22. Wang D, Lo HK (2010) Global optimum of the linearized network design problem with equilibrium flows. Transp Res Part B 44:482–492
23. Luathep P, Sumalee A, Lam WHK, Li ZC, Lo HK (2011) Global optimization method for mixed transportation network design problem: a mixed-integer linear programming approach. Transp Res Part B 45:808–827
24. Li C, Hai Y, Zhu D, Meng Q (2012) A global optimization method for continuous network design problems. Transp Res Part B 46:1144–1158
25. Wang G, Gao Z, Xu M, Sun H (2014) Models and a relaxation algorithm for continuous network design problem with a tradable credit scheme and equity constraints. Comp Oper Res 41:252–261
26. Erkut E, Verter V (1998) Modeling of transport risk for hazardous materials. Oper Res 46:625–664
27. Vincent RA, Mitchell AI, Robertson DI (1980) User Guide to TRANSYT, LR888. TRRL, Crowthorne
28. Chiou SW (2003) TRANSYT derivatives for area traffic control optimisation with network equilibrium flows. Transp Res Part B 37:263–290
29. Karmitsa N, Bagirov A, Makela MM (2012) Comparing different nonsmooth minimization methods and software. OMS 27:131–153
30. Sagastizbal C (2013) Composite proximal bundle method. Math Prog 140:189–233
31. Bonnans J, Gilbert J, Lemarechal C (2006) Sagastiźabal C. Numerical optimization. Theoretical and practical aspects, Universitext. Springer, Berlin
32. Lemarechal C (2001) Lagrangian relaxation. In: Computational combinatorial optimization. Lecture notes in computer science, vol 2241. Springer, Berlin, pp 112–156
33. Hare W, Sagastizabal C (2010) A redistributed proximal bundle method for nonconvex optimization. SIAM J Opt 20:2442–2473
34. Hiriart-Urruty JB, Lemarechal C (1993) Convex analysis and minimization algorithms II. Springer, Berlin
35. Makela M, Neittaanmaki P (1992) Nonsmooth optimization: analysis and algorithms with applications to optimal control. World Scientific Publishing Co., Singapore
36. Makela M (2002) Survey of bundle methods for nonsmooth optimization. OMS 17:1–29
37. Kiwiel KC (1985) Methods of descent for nondifferentiable optimization. Lecture notes in mathematics, vol 1133. Springer, Berlin
38. Kiwiel KC (1990) Proximity control in bundle methods for convex nondifferentiable minimization. Math Prog 46:105–122
39. Cheney EW, Goldstein A (1959) Newton's method for convex programming and Tchebycheff approximation. Numer Math 1:253–268
40. Shor N (1998) Nondifferentiable optimization and polynomial problems. Kluwer Academic Publishers, Boston
41. Colson B, Marcotte M, Savard G (2007) An overview of bilevel optimization. Ann Oper Res 153:235–256
42. Kelley JE (1960) The cutting plane method for solving convex programs. J SIAM 8:703–712
43. Patriksson M, Rockafellar RT (2003) Sensitivity analysis of aggregated variational inequality problems, with application to traffic equilibrium. Transp Sci 37:56–68

44. Dontchev AL, Rockafellar RT (2002) Ample parameterization of variational inclusions. SIAM J Opt 12:170–187
45. Kiwiel KC (1995) Proximal level bubdle methods for convex nondiferentiable optimization, saddle-point problems and variational inequalities. Math Prog 69:89–109
46. Kiwiel KC (1996) Restricted step and Levenberg–Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. SIAM J Optim 6:227–249
47. Kiwiel KC (2006) A proximal bundlemethod with approximate subgradient linearizations. SIAM J Optim 16:1007–1023
48. Suwansirikul C, Friesz TL, Tobin RL (1987) Equilibrium decomposed optimization: a heuristic for the continuous equilibrium network design problem. Transp Sci 21:254–263
49. Giannikos I (1998) A multiobjective programming model for locating treatment sites and routing hazardous wastes. EJOR 104:333–342
50. Current J, Ratick S (1995) A model to assess risk, equity and efficiency in facility location and transportation of hazardous materials. Loc Sci 3:187–201
51. Gopalan R, Batta R, Karwan M (1990) The equity constrained shortest path problem. Comp Oper Res 17(3):297–307
52. Marianov V, ReVelle C (1998) Linear non-approximated models for optimal routing in hazardous environments. JORS 49(2):157–164
53. Kang Y, Batta R, Kwon C (2014) Generalized route planning model for hazardous material transportation with VaR and equity considerations. Comp Oper Res 43:237–247
54. Miandoabchi E, Daneshzand F, Szeto WY, Farahani R (2013) Multi-objective discrete urban road network design. Comp Oper Res 40:2429–2449
55. Wang S, Meng Q, Yang H (2013) Global optimization methods for the discrete network design problem. Transp Res Part B 50:42–60
56. Cantarella GE, Pavone G, Vitetta A (2006) Heuristics for urban road network design: lane layout and signal settings. EJOR 175(3):1682–1695
57. Miandoabchi E, Farahani R (2011) Optimizing reserve capacity of urban road networks in a discrete network. Adv Eng Soft 42:1041–1050
58. Farahani R, Miandoabchi E, Szeto WY, Rashidi H (2013) A review of urban transportation network design problems. EJOR 229:281–302

Evaluation of Evacuation Corridors and Traffic Big Data Management Strategies for Short-Notice Evacuation

Lei Bu and Feng Wang

Abstract The chapter presents a simulation study of the large-scale traffic data under a short-notice emergency evacuation condition due to an assumed chlorine gas spill incident in a derailment accident in the Canadian National (CN) Railway's railroad yard in downtown Jackson, Mississippi by employing the dynamic traffic assignment simulation program DynusT. In the study, the effective evacuation corridor and traffic management strategies were identified in order to increase the number of cumulative vehicles evacuated out of the incident-affected protective action zone (PAZ) during the simulation duration. An iterative three-step study approach based on traffic control and traffic management considerations was undertaken to identify the best strategies in evacuation corridor selection, traffic management method, and evacuation demand staging to relieve heavy traffic congestions for such an evacuation.

Keywords Large-scale emergency evacuation · Computer simulation · Dynamic traffic assignment · Protective action zone · Traffic control and management

1 Introduction

Short-notice incident is one that occurs unexpectedly or with minimal warning which does not allow emergency responders sufficient time to prepare for it. The incident may be natural or manmade which could be localized or widespread with a variety of primary and secondary consequences. The majority of incidents that precipitate a short-notice evacuation occur within a very local area, most often in

L. Bu (✉) · F. Wang
Institute for Multimodal Transportation and Department of Civil and Environmental Engineering, Jackson State University, 1400 Lynch Street, PO Box 17068, 39217-0168 Jackson, MS, USA
e-mail: leibu04168@gmail.com

F. Wang
e-mail: feng.wang@jsums.edu

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_9

urbanized locations, such as structure fires, gas leaks, chemical spills, transportation accidents, and terrorist attacks involving conventional explosives, while larger incidents may affect an entire city or region. Examples of wide-scale, short-notice incidents that would likely require a sizable evacuation include a tsunami, chemical release that results in a large moving toxic cloud, explosion at a specialized site such as liquid natural gas facility, and terrorist attack using unconventional explosives [1]. Evacuations that result from either a localized or widespread incident will likely involve a tremendous number of evacuees, primarily in the format of vehicular traffic, possibly from more than one community or jurisdiction, who need to move away from the at-risk area. This will require intensive efforts on the role of emergency managers, first responders, law enforcement officers, and traffic management professionals to coordinate, guide, and shelter the affected population.

The Greater Jackson is the largest metropolitan area in Mississippi with a total population of 539,000. In downtown Jackson, there are social-economically important facilities including state governments, university campuses, football stadium, and medical centers. When a chemical spill incident happened, even an evacuation of only one or two blocks would involve the movements of thousands of people, and on-scene emergency responders may be needed at the site from which at-risk people are evacuated. How to deploy traffic control and/or traffic management effectively on the local traffic network around the incident location and select the evacuation route(s) with the best evacuation performance has become a quite realistic question faced by the state and local transportation agencies.

This study aims to moving a large disaster affected population through a transportation network toward safer areas quickly and efficiently for a no-notice or short-notice emergency evacuation by selecting the best evacuation corridor and deploying the most effective traffic management strategies including contra-flow and evacuation demand staging strategy. Traffic simulation tool DynusT based on Dynamic Traffic Assignment (DTA) will be applied to test the performances of scenarios assumed.

2 Literature Review

In emergency and critical infrastructure management, the adoptions of optimal evacuation routes and effective traffic management strategies could well improve the evacuation performance by relieving heavy traffic congestions generated by the sudden surge of the evacuation traffic demand [2, 3]. The planning and design of effective emergency evacuation traffic operations have been rigorously investigated [4–6] since the early studies dealing with traffic management under emergency conditions. The previous research has identified the most feasible operational strategies to maximize the efficiency and performance of transportation infrastructure under evacuation conditions to be: (1) contra-flow operation [2, 7–10], and (2) demand loading and staging strategy [11–16].

2.1 Contra-Flow Operation

The contra-flow configurations and flow rates under full contra-flow operations (i.e., one-way-out) were studied and the maximum evacuation flow was identified by presenting the relevance of a practical maximum sustainable evacuation traffic flow rate based on the speed-density characteristics of the evacuation traffic observed during the major hurricanes in the past ten years [2, 17]. Comparative analyses on contra-flow evacuation traffic streams were also conducted at same locations but at different degrees of contraflow. A quantitative evaluation of the performance of alternative evacuation strategies was assessed using the dynamic traffic assignment model DynusT, with four supply scenarios, namely, the base case with the existing supply conditions with no contra-flow, the full contra-flow scenario with maximized roadway capacities and paved shoulders, the evaculane scenario (exclusive evacuation lanes), and the partial contra-flow scenario representing the intermediate supply levels between the two extremes [18]. It is noticeable that the contra-flow strategy was mainly deployed on the freeways in most of the previous studies. As a matter of fact a contra-flow deployment on arterial streets would be more complicated and challenging than on freeways due to the access disturbance at intersections.

2.2 Demand Loading and Staging Strategy

For the evacuation demand loading and staging strategy, an in-depth analysis for the selected evacuation demand generation and loading models, including S-curve model, Rayleigh distribution model, and sequential logit model was conducted [15]. The problem of scheduling evacuation trips between a selected set of origin nodes and safety destinations was studied to minimize network clearance time [16]. Meanwhile, models for trip generation and trip distribution were proposed and departure rate was estimated based on the Rayleigh cumulative density function [14]. A spatio-temporally optimal model for evacuation was presented with the evacuation performance compared to both the simultaneous evacuation and the temporally optimal evacuation cases [19]. In addition, three different staging strategies for the departure-timing shift, namely, half strategy, quarter strategy and split strategy were identified, and the stage timing was based on the network quartile division [13]. In these previous studies, staged evacuation departures were proved to be effective in reducing congestions due to queue buildup and dissipation. A staging strategy seemed quite reasonable for evacuations due to hurricanes or floods where there was enough deployment time for a staging operation. However, the benefits of distributing evacuation trip demand with acceptable temporal patterns would potentially improve short-notice or no-notice evacuations as well. The S-curve staging strategy of evacuation demand will be applied in this study to

compare the evacuation performances with that of the demand even distribution in the different period of time.

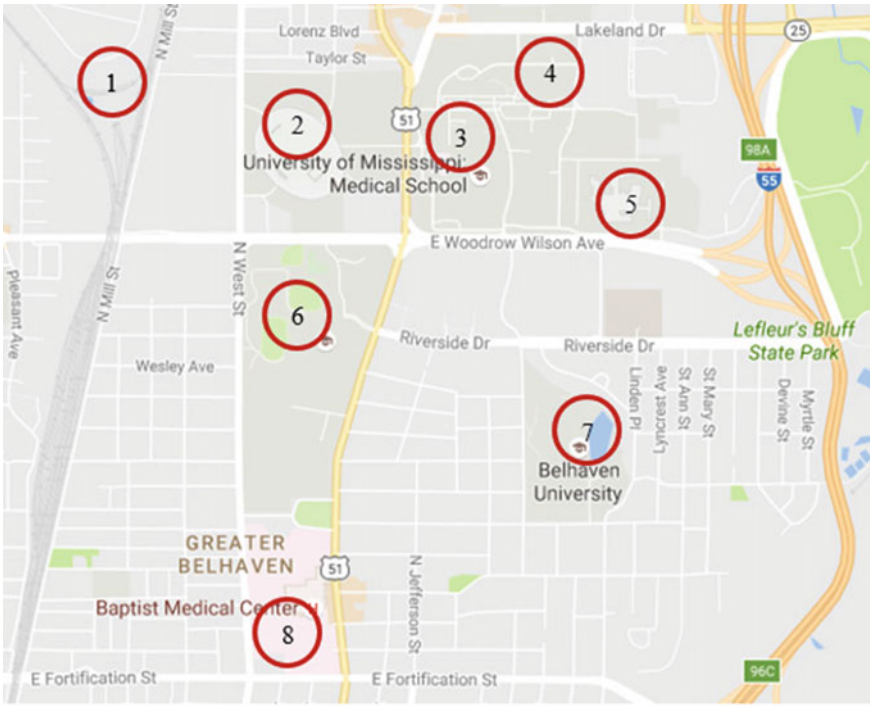
2.3 Traffic Simulation Tool DynusT

In recent years, the parallel computing and genetic algorithm based mesoscopic Dynamic Traffic Assignment (DTA) traffic simulation tool DynusT was used for the simulations of several complex and large-scale evacuation studies, and the key observations in these studies concluded that an optimized spatio-temporal (dynamic) evacuation could achieve a faster evacuation with less time spent in the system and less number of evacuees stuck in the network [18, 20, 21]. The DTA nature of the DynusT program also made it capable of testing the effect of the deployment of Intelligent Transportation System (ITS) devices such as variable/dynamic message signs (VMS or DMS) that provided real-time traffic information to the evacuees [22]. For an urban environment in the US, ITS technologies are quite popular and as a result real-time dynamic traffic information has been effectively disseminated to and used by the road users, therefore a dynamic traffic assignment program would be appropriate to be used to model the dynamic driver behaviors in response to real-time traffic information during an evacuation trip in an urban environment.

3 Background Description for Simulation

Jackson is not only the capital city and the most metropolitan area of Mississippi, but also a railroad town with the presences of Canadian National (CN), Kansas City Southern (KCS), and Amtrak. As early as in 1870s, the Illinois Central (IC), as the predecessor of the present Canadian National had started railroad transportation service in Jackson and Mississippi. The CN North Jackson Yard on N. Mill St is located in the heart of the Greater Jackson area, only a few miles to each of the densely populated cities Jackson, Ridgeland, Flowood, and Madison. In the very near vicinity (2.0 miles) of the railroad yard are several social-economically important locations that include Jackson State University's football stadium or Mississippi Veterans Memorial Stadium (MVMS), the University of Mississippi Medical Center campus (UMMC), Millsaps College, Belhaven University, Baptist Hospital, St Dominic's Medical Center, G.V. "Sonny" Montgomery VA Medical Center, and the state government offices. Figure 1 shows some of the important locations near the CN railroad yard. The map also shows the existence of Interstate highway 55, and three major arterial streets near the CN railroad yard, which are N West Street, N State Street, and Woodrow Wilson Avenue (WWA).

In this study, a spill of the highly toxic chlorine gas as a freight train with 30,000 gallon tanker was assumed to have derailed at the CN railroad yard near N. Mill

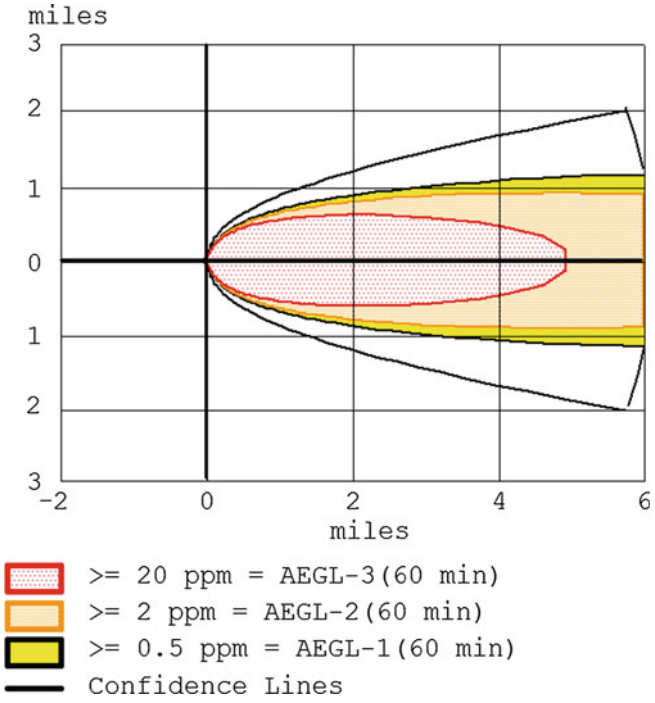


1- CN North Yard;
2-Mississippi Veterans Memorial Stadium (MVMS);
3- University of Mississippi Medical Center (UMMC);
4- St Dominic's Medical Center;
5-Montgomery VA Medical Center;
6-Millsaps College;
7-Belhaven University;
8-Baptist Hospital.

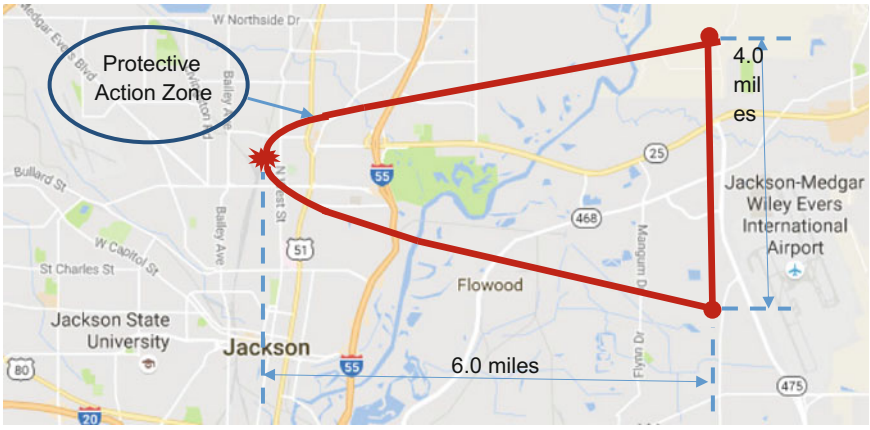
Fig. 1 CN north yard and social-economically important locations

Street in Jackson downtown close to the Jackson State University MVMS stadium during the afternoon peak hours and the worst prevalent wind intensity and direction, the game audience in the stadium, and the afternoon peak hour traffic condition were all to be considered. Using DynusT, a DTA simulation tool in an iterative study approach developed for the assumed emergency incident and evacuation, the three major arterial streets located near the stadium and UMMC campus were tested to identify a most effective evacuation corridor and then effective traffic management and control strategies along with possible patterns of evacuation traffic demand distribution were studied by using dynamic traffic assignment based simulations for different evacuation scenarios.

The Areal Locations of Hazardous Atmospheres (ALOHA) program developed by the U.S. Environmental Protection Agency (EPA) and National Oceanic and Atmospheric Administration (NOAA) was adopted for the Protective Action Zone



(a) Computation result by ALOHA



(b) Protective action zone

Fig. 2 Estimation of affected area for evacuation

(PAZ) calculation to determine the zone area affected by the spill incident (NOAA [23]. Figure 2 shows the threat zone due to the chlorine spill, calculated by the ALOHA software.

As shown in Fig. 2a, the PAZ is outlined by a confidence line, which includes uncertainty in a safety manner. The downwind distance of the outlined zone is 6.0 miles and the widest span perpendicular to the wind direction is about 4.0 miles. The outermost area (yellow) inside the outlined zone is at Acute Exposure Guideline Level 1 (AEGL-1) by the 60th minute after the spill incident. Within the AEGL-1 line, the concentration of chlorine gas is between 0.5 ppm and 2 ppm. The next area (orange) is at AEGL-2 by the 60th min. Within the AEGL-2 line, the concentration of chlorine gas is between 2 ppm and 20 ppm. The innermost (red) area is at AEGL-3 in 60 min. Within the AEGL-3 line, the concentration of chlorine gas is equal to or larger than 20 ppm. To have maximum safety, the area outlined by the confidence line with a downwind distance of 6.0 miles and a span of 4.0 miles at the wide end was taken as the Protective Action Zone or the evacuation zone, totally 17.3 square miles, which is shown in Fig. 2b.

4 Evacuation Trip Demand Modeling

The simulation program dynamically assigned each vehicle a "shortest path" over the highway network with nodes and links, traffic control devices and traffic management deployments, and real-time traffic information, for each of the OD (origin and destination) pairs due to the trip demands for routine transportation needs and evacuation needs in the Greater Jackson area. The study area consists of three counties: Hinds, Rankin, and Madison that encompass the Greater Jackson Area which is the most metropolitan area of Mississippi. Social-economic data, traffic data, and highway network data for the TAZ zones of the study area were provided by the Mississippi Department of Transportation (MDOT). Therefore, the simulation program needed two major data inputs which were the highway network model and the OD demand tables. The highway network model was provided by MDOT's Transportation Planning Division. The MDOT network model geographically divided the Greater Jackson Area into 691 traffic analysis zones (TAZ) with 4,607 nodes and 10,288 links which is shown in Fig. 3.

Characteristics data for each node and link, and social-economic attributes for each TAZ zone were also contained in the model. The trip production, trip attraction, and trip distribution for every TAZ were calculated by using the local MPO models based on the NCHRP Report No. 365 [24]. The TransCAD [25] program was used to generate the 24 h O-D demand tables for the study. The steps for preparing the trip demand data are discussed in the following paragraphs.

Fig. 3 Network for the simulation

4.1 Background and Evacuation Trip Demands

The evacuation demand includes two parts. One part is a reduced background demand, and the other a modified impact demand. The trip demand due to daily and routine transportation needs is called background traffic. The order of the emergency evacuation should have a reduction effect on the background demand. From the moment when the evacuation operation were in effect, all newly generated trips with destinations located within the PAZ should be cancelled due to the chlorine spill and the evacuation order. On the other hand, the modified impact demand is the evacuation traffic that moves evacuees away from the PAZ. During the evacuation operation in the simulation duration, the modified impact evacuation traffic demand and the reduced background demand occurred concurrently.

The DynusT program was set up to simulate for 4 h of traffic operation including the time before and after the evacuation was ordered. At the beginning of the first hour, a normal background demand table was loaded to the road network. And since the beginning of the second hour, a reduced background demand table and a modified impact demand table were loaded for each of the three evacuation hours. The summation of the two demand tables accounted for the total evacuation trips which were loaded to the simulation program where the dynamic traffic assignments of the evacuation/background trips over the network were conducted.

4.2 Evacuation Trip Production and Attraction

The total impact evacuation demand was estimated at 55,281 passenger car vehicles in the normal-day which includes the population of centers such as hospitals and colleges mentioned. There was 193 bus vehicles by public transportation when there was not an activity in the MVMS stadium. If a game activity were going on in the football stadium when the chlorine spill happened, more evacuation trips and vehicles would be expected. The capacity of the stadium is 60,492 seats. With a stadium activity considered, additional trip production would be added from the audience of the activity in Zone #149 where the stadium is located. The number of evacuation vehicles from the stadium was assumed to be in four different levels as 0 (baseline), 20,000, 40,000, and 60,492 vehicles, respectively. On the other hand for trip attraction, TAZ zones with major residency areas and public places as temporary sheltering were considered destination zones for background and evacuation trips.

4.3 Evacuation Trip Distribution and O-D Demand Table

The trip distributions and O-D demand tables for the reduced background trip demand and the modified impact evacuation trip demand were generated by applying the impact evacuation trip production and attraction data to the gravity model and friction factors (based on the NCHRP Report No. 365) using the TransCAD software. The OD demand table (matrix) of the reduced background demand and the OD demand table (matrix) of the impact demand for each hour were combined and input to the simulation program DynusT to conduct dynamic traffic assignments to generate trip trajectories for each trip OD pair in traffic simulation runs. The evacuation performance measurements were extracted from the simulation results for the evacuation traffic demands and traffic management strategies over the network to evaluate and identify the effective traffic management strategies.

5 Traffic Management Strategy Development

Before the order of evacuation, law enforcement personnel, traffic control experts and devices were assumed to be in place at key highway locations such as intersections and freeway exits/entrances. Traffic control and management strategies were developed based on knowledge of the existing practices and consideration of the local situation. A calibration process comparing the real data collection of traffic volumes with the results of the simulation program for various network locations was conducted to make necessary adjustments to the simulation setup.

5.1 Baseline Traffic Management

After the order of evacuation, traffic controls were (with law enforcement) placed at upstream links of relevant highways or streets to block inbound traffic from entering the PAZ zone. For example, traffic on I-55 SB and I-55 NB were forced to exit the freeway before the Meadowbrook interchange and E Fortification Street interchange, respectively. However, the freeway segment between the Meadowbrook interchange and E Fortification Street interchange remained open for the evacuation traffic to leave the PAZ zone.

5.2 Advanced Traffic Management Strategies

In the study, the effectiveness of deploying contra-flow operation and variable massage signs (VMS) was evaluated for three arterial streets. The reversal use of inbound traffic lanes with light traffic volumes for the heavily used outbound direction has been proved effective in many studies and successfully implemented in practices as well. As an Intelligent Transportation System facility, a VMS device can be connected to and remotely controlled by a traffic management center (TMC) through fiber optic cables or wireless technologies to display real-time dynamic traffic information such as congestion, incident, or detour information to assist the road users to avoid the congested routes or locations and therefore increase mobility for the whole network. In addition, the possible effect of a staging controlled demand on the evacuation performance was also tested using simulations

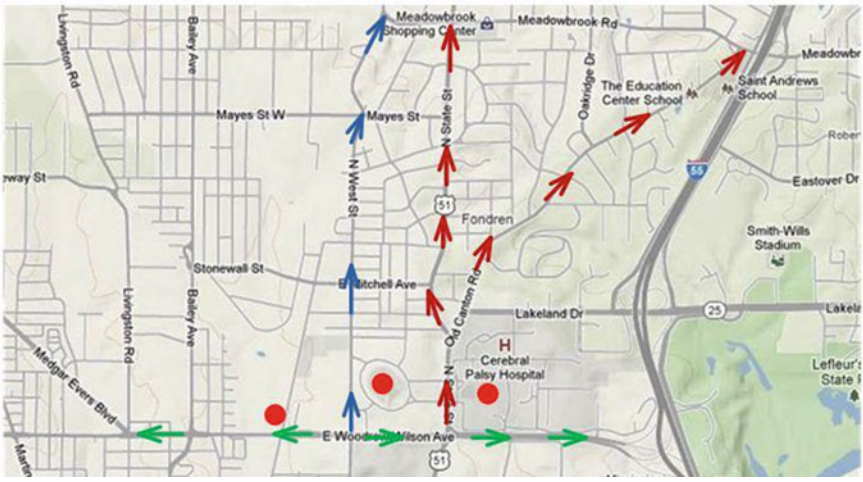


Fig. 4 Contra-flow deployments on candidate corridors

in the study to quantify the degrees of improvement by applying a staging strategy to such an evacuation situation.

As shown in Fig. 4, the contra-flow operation on each of the three candidate arterial streets is described in the following: (1) The simulated contra-flow operations on Woodrow Wilson Avenue were for both westbound and eastbound directions which were separated by a segment between Mill Street and N West Street where the railroad yard is located. The contra-flow operation on the west segment of Woodrow Wilson Avenue was from Mill Street until Medgar Evers Blvd and the contra-flow operation on the east segment was from N West Street to I-55. The west and east segments were made westbound and eastbound respectively moving traffic away from the PAZ. (2) The simulated contra-flow operation on N State Street was on the north segment by reversing the direction of the southbound traffic lanes to northbound over that segment, beginning at Woodrow Wilson Avenue and ended at Meadowbrook Road. The Old Cannon Road is diverged from N State Street and was also deployed northbound contra-flow up to I-55. (3) N West Street runs parallel to N State Street and is on the other side of the MVMS stadium. A simulated contra-flow operation was deployed on the north segment of the street beginning at Woodrow Wilson Avenue and ended at Meadowbrook Road.

5.3 Traffic Signal Consideration

The detailed traffic signal timing data for emergency evacuation operation were not available from MDOT or Jackson City, and therefore traffic signal control data were not included in this study. To enable a relatively realistic simulation process, the study assumed that all signalized intersections were equipped with actuated signal controllers, which could automatically adjust green time allocations for all the different approaches at an intersection depending on the incoming time-varying traffic volumes of the approaches. Although this modeling strategy might slightly overestimate the available traffic capacity on the arterial streets, the method served as the best available approximation to normalize the effects of the traffic signal control, and therefore provided a platform for the study to evaluate and compare the effects of interest for the advanced traffic management strategies and other contributing factors under the normalized signalization traffic control condition.

5.4 Simulation Calibration

To evaluate the accuracy of the dynamic traffic assignment model in the DynusT software and to determine input parameters for hourly volume percentages, a simulation calibration procedure was followed. This was an iterative process of adjusting the input parameters of the simulation software so that the traffic volumes computed by the DynusT software at specified network locations for the peak hours matched

well with the collected actual volumes satisfying the predefined error criteria. With the iterative process, the percentages of average daily traffic (ADT) distributed in the 4 h at the afternoon peak from 2:00 to 6:00 p.m. were determined as 6.2 %, 6.8 %, 7.9 %, and 8.8 %, respectively. Four hourly OD demand matrices were produced by multiplying the initial hourly traffic percentages obtained by borrowing from a known traffic distribution pattern to the total ADT based background demand matrix and then imported to the DynusT program for calibration simulation runs. Traffic counts were manually conducted for selected highway locations during the afternoon peak hours. In addition, the traffic volume data collected during a previous evacuation event were provided by MDOT's Traffic Division and used for the simulation calibration. The calibration was an iterative trial and error process. The standard measurement method of GEH (after Geoffrey E. Havers) statistic was used for the traffic volume calibration [26] and the formula is shown as follows:

$$\mathrm{GEH} = \sqrt{\frac{2(\mathrm{M}-\mathrm{C})^2}{\mathrm{M}+\mathrm{C}}} \tag{1}$$

where M is the model estimated hourly traffic volume and C is the field observed hourly traffic volume.

After more than 50 simulation trials, the calibrated input parameters for hourly traffic percentages for the four peak hours and link capacities of the arterials were adjusted, satisfying all FHWA limits [26, 27]. The difference between the overall assignment volume for all links in the network and the traffic volumes provided by MDOT was within the ±5 % limit. The difference between the total assigned freeway volume and the observed total was within the ±7 % limit. The errors on specific links between the assigned hourly volumes and the observed ones were all within the ±12 % limit.

6 Simulation Results and Analysis

To quantify the evacuation performance, the following Measures of Effectiveness (MOEs) were adopted: cumulative vehicle evacuation percentage, vehicle throughput, average trip travel time, and total evacuation time. The cumulative vehicle evacuation percentage is defined as the percentage of the total number of vehicles evacuated out of the PAZ zone over the total number of vehicles generated in the PAZ along with time in 10 min increments over the simulation period; Vehicle throughput is defined as the number of vehicles that arrive at destinations in every 10-min interval over the simulation period; The average trip travel time is the average time in minutes elapsed per vehicle trip from origin to destination; The total evacuation time is the length of time in minutes for the evacuation of 90 % of the all generated vehicles out of the PAZ. The simulation results are presented in three steps: Step 1 Evacuation Corridor Selection; Step 2 Traffic Management Strategy

Evaluation; and Step 3 Evacuation Demand Staging Effect. Four levels of stadium traffic volumes of 0, 20,000, 40,000, and 60,492 (at the stadium's capacity of 60,492 seats) vehicles were simulated and denoted as scenarios 1, 2, 3, and 4 respectively.

6.1 Step 1-Evacuation Corridor Selection

For each of the three candidate corridors of N State Street, N West Street, and Woodrow Wilson Avenue, two different management strategies: portable VMS deployment and/or contra-flow operation of traffic lanes on the candidate corridor segment were tested. Evacuation demand was loaded over 2 h from the 60th minute to the 180th minute with an equal demand (in aggregate) in each 60-minute interval. The evacuation performance was assessed by counting the total number of cumulative evacuation vehicles evacuated out of the PAZ at the end of the simulation period. Table 1 lists the numbers of cumulative evacuation vehicles (CEV) for the three candidate arterial corridors with or without the traffic management strategies of contra-flow and VMS deployments. The following paragraphs are the observations and discussions on the CEV numbers presented in Table 1.

At stadium volume of 0 vehicles, deploying VMS signs only on each of the three corridor streets could increase the total CEV number compared with that without deploying any traffic management strategy. Deploying contra-flow on N West Street could increase the CEV number no matter whether the VMS signs were deployed or not. The contra-flow deployments reduced CEVs on the other two streets, even with the deployment of VMS signs, probably because of the access disturbance from the university campuses along the two streets. Deploying both contra-flow and VMS on N West Street could obtain the largest CEV number among all the three streets.

Table 1 Cumulative evacuated vehicle volumes in simulations

Demand scenario	Arterial street	Without contra-flow without VMS	VMS without contra-flow	Contra-flow without VMS	Contra-flow with VMS
#1	West st	50580	51643	51468	54495
	State st	50580	50788	49000	49294
	WWA	50580	50628	50330	50468
#2	West st	63735	65539	65763	67358
	State st	65880	66023	64222	66162
	WWA	60487	62431	63596	65337
#3	West st	65134	70311	70913	70758
	State st	74299	74535	75686	76449
	WWA	68327	68415	71395	74898
#4	West st	62390	67819	70995	71485
	State st	75728	76076	77800	78431
	WWA	68479	69779	77206	78544

At stadium volume of 20,000 vehicles, the traffic management strategy deployments on each street could increase the total CEV numbers except for the contra-flow deployment on N State Street without VMS signs. N State Street and N West Street show larger CEV numbers by deploying both contra-flow and VMS.

At stadium volume of 40,000 vehicles, the traffic management deployments on each street could increase the total CEV numbers and N State Street could handle the largest number among the three streets by deploying both contra-flow and VMS.

At stadium volume of 60,492 vehicles, the traffic management deployments on each street could increase the total CEV numbers. N State Street and E Woodrow Wilson Avenue could obtain larger numbers by deploying both contra-flow and VMS, probably because of the advantages of connections to the freeway I-55 by the Woodrow Wilson Avenue and the N State Street/Old Canton Road corridors.

As shown in Table 1, by deploying both contra-flow and VMS strategies, the evacuation performance on each of the three streets can be improved compared with the case without any traffic management strategy, except for N State Street when there is not a stadium volume. This is mainly because the link features of the N State Street are different from those of the other two streets. The three university campuses, two hospitals, and other population concentrations are all close to the State Street, which may mean that N State Street is more prone to the disturbance from a contra-flow deployment than the other two streets. Meanwhile, the evacuation performance for deploying both contra-flow and VMS strategies are generally better than that of deploying only one of the two strategies. At a low evacuation traffic demand level, the capacity of an arterial street is about enough to accommodate the evacuation traffic demand, and the contra-flow deployment may cause disturbance to normal traffic and decrease the lane capacity. The benefit of the gained total capacity due to the contra-flow might be canceled out by the lane capacity loss due to the disturbance.

Generally N State Street shows the best evacuation performance among all the three streets in all the tested evacuation scenarios by achieving the largest numbers of cumulative vehicles evacuated out of the spill affected PAZ zone during the simulation period. N West Street would be ideal for a contra-flow deployment at low evacuation traffic demands because of its relatively small access disturbance. In contrast, both N State Street and Woodrow Wilson Avenue have good evacuation performance at high evacuation traffic demands simply because the two routes both have direct network connectivity to the freeway I-55. Based on the simulation results, N State Street was selected as the best evacuation corridor for the further study steps.

6.2 Step 2-Traffic Management Strategy Evaluation

In order to test and evaluate the performance of deploying traffic management strategy on N State Street, there were not evacuation strategies deployed on the other evacuation corridors. Possible traffic management strategies available for deployment on N State Street would be full contra-flow, partial contra-flow, adding a lane, and staged departure of evacuation demand. Portable ITS/VMS signs were always

deployed in Step 2 for their proven effectiveness. The following iterative traffic management strategies were tested in Step 2: (1) Full contra-flow. In this strategy, both of the two inbound lanes were reversed so that all the lanes could be used by the evacuation vehicles to leave the PAZ; (2) Partial contra-flow. In this strategy, one of the two inbound lanes was reversed and the other lane retained the inbound lane function so that the emergency ambulance and emergency management personnel may enter the PAZ using the inbound lane; (3) Adding a lane in each direction. In this strategy, a shoulder lane was used in each direction; (4) Staging without any contra-flow or lane addition; and (5) Staging with partial contra-flow.

In strategies (4) and (5), demand staging was initially tested by evenly distributing evacuee departures in 2 h from the 60th min to the 180th min in 15 min intervals. Evacuation demand temporal distribution for each scenario is shown in Fig. 5 and evacuation departure curve for each scenario is shown in Fig. 6.

Figure 7 shows the cumulative percentages of vehicles evacuated out of the PAZ and Fig. 8 shows the vehicle throughputs for the five different traffic management strategies with time in the simulation period.

The observations of the simulation results in Step 2 are as follows.

The cumulative evacuation percentage results in Fig. 7 show that using staged evacuation demand load strategy could well improve the evacuation performance by increasing the cumulative evacuation percentage in each of the four demand scenarios. Especially the strategy deploying partial contra-flow with staged evacuation demand load could evacuate more than 90 % of evacuation traffic out of the affected area in the simulation period. This strategy uses the shortest evacuation time among all the five strategies tested and for all the four demand scenarios, which are shown in Table 2.

The vehicle throughput results in Fig. 8 show that at 0 and 20,000 stadium volumes, adding a lane, contra-flow, and partial contra-flow are among the best throughputs in the first evacuation hour, but in the last 2 h surpassed by the two strategies with staging. At stadium volumes of 40,000 and 60,492 vehicles, deploying the staged partial contra-flow could achieve the best throughput performance throughout all the three evacuation hours.

In Step 2, the staging strategy involved an even distribution of the total evacuation demand in 2 h, namely, by loading 12.5 % of the total evacuation demand onto the network in each of the eight 15-minute intervals from the 60th minute to the 180th min of the simulation period of 4 h. However, in a short-notice incident, evenly distributing evacuation traffic demand may be restricted by the emergency situation. Therefore, other formats of staging are further studied in Step 3.

6.3 Step 3-Evacuation Demand Staging Effect

In Steps 3, three different staging distributions combined with the deployment of partial contra-flow strategy were tested and simulated for N State Street to compare the evacuation performance. In the test, only evacuation demand generated from the stadium was applied for the temporal staging distributions to relieve the burst heavy

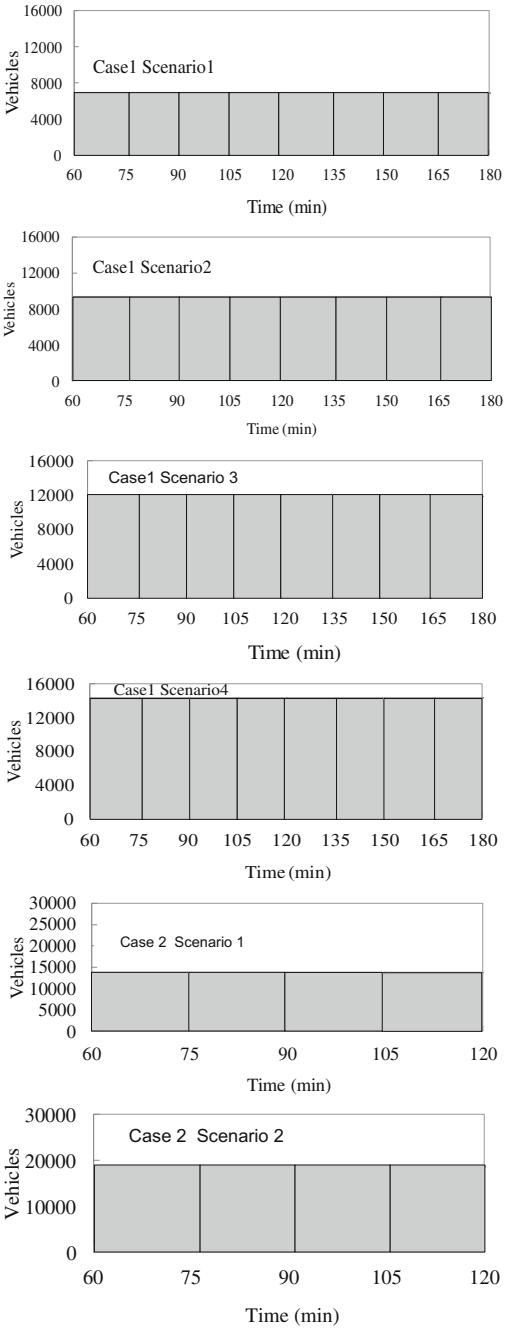


Fig. 5 Evacuation demand temporal distribution in Step 3

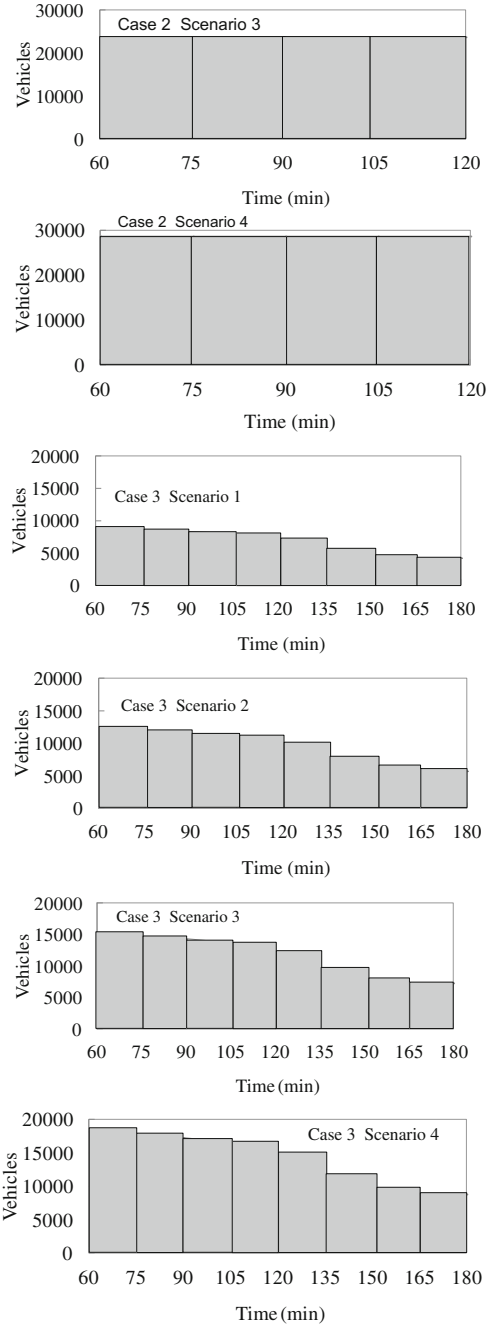


Fig. 5 (continued)

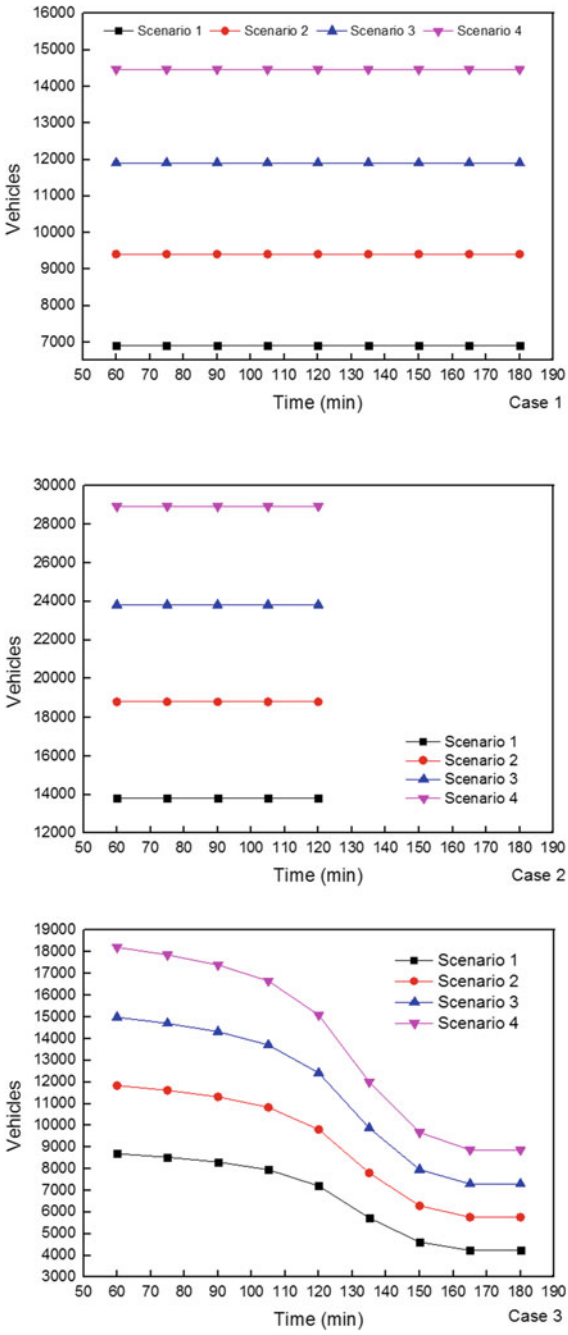


Fig. 6 Evacuation departure curves in step 3

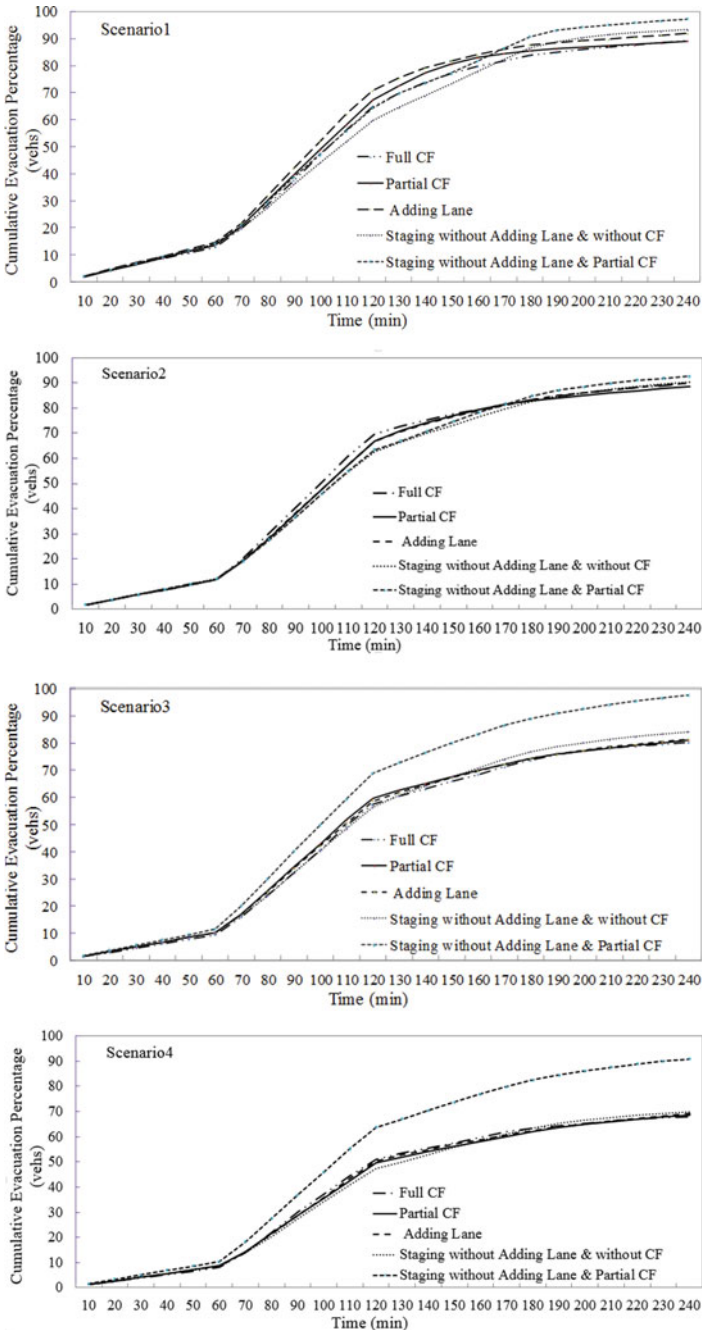


Fig. 7 Simulation results in step 2—cumulative evacuation percentages

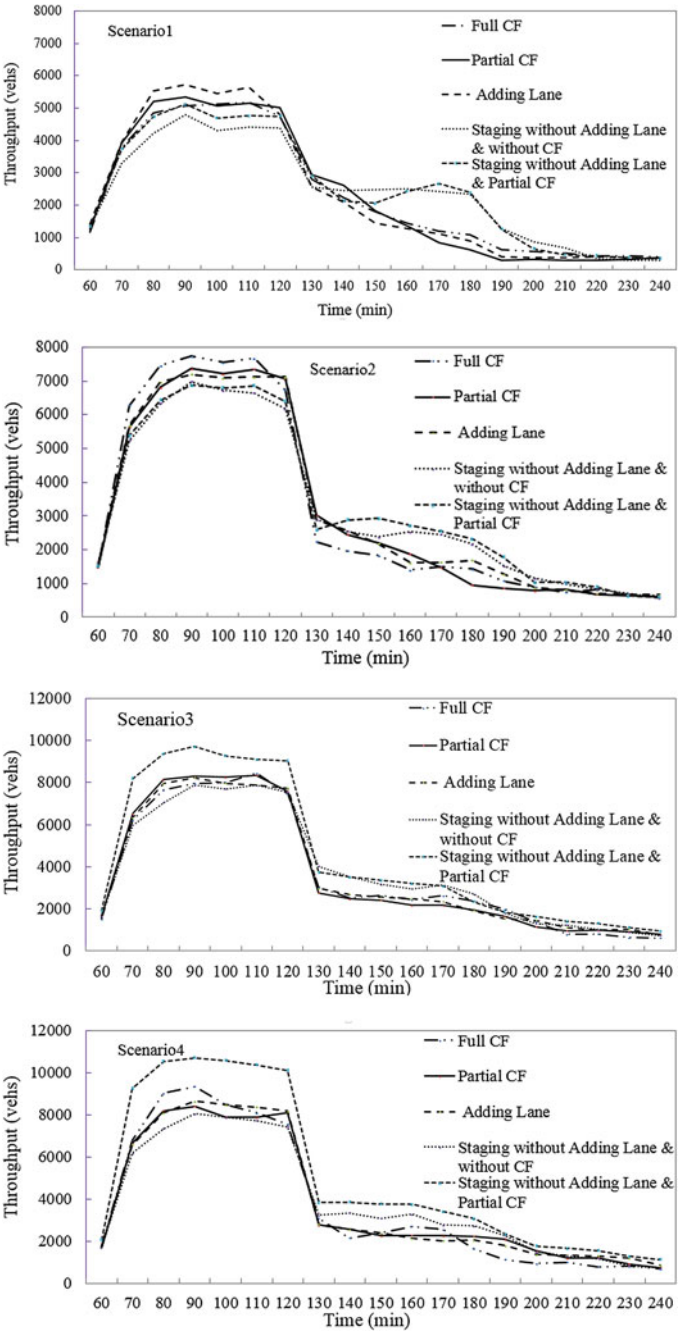


Fig. 8 Simulation results in step 2—vehicle throughputs

Table 2 Evacuation time and average travel time (min) for management strategies

Demand scenario	Performance	Full CF	Partial CF	Adding lane	Staging	Staging with partial CF
1	Evacuation time	>180	>180	150	135	115
	Average travel time	27.16	26.79	27.40	25.68	25.47
2	Evacuation time	>180	>180	>180	175	155
	Average travel time	27.48	27.01	28.15	26.43	26.17
3	Evacuation time	>180	>180	>180	>180	125
	Average travel time	28.56	28.15	29.18	27.19	27.04
4	Evacuation time	>180	>180	>180	>180	175
	Average travel time	29.53	29.04	29.23	27.86	27.80

congestion around the stadium. Table 3 lists the three cases of evacuation demand departure rates in terms of percentage of the total evacuation traffic demand with time.

As shown in Table 3, Case 1 is the staging by loading 12.5 % of the evacuation demand in eight 15-min intervals from the 60th min to the 180th min of the simulation period. Case 2 is loading 25 % of the evacuation demand in four 15-min intervals from the 60th min to the 120th. Case 3 is the s-curve distribution of the evacuation demand over the 60th min to the 180th min.

The vehicle throughputs with time for the three staging distribution cases and four evacuation demand scenarios are shown in Fig. 9.

The evacuation times (min) and average trip times for the three staging distribution cases and four evacuation demand scenarios are listed in Table 4.

The cumulative evacuation percentage results and the throughput values indicate that at demand scenario 1 with 0 vehicles exiting from the football stadium, all the staging strategies could evacuate 90 % of the evacuation demands out of the PAZ in the 3 h of evacuation operation in the simulations. At demand scenario 2 with

Table 3 Evacuation demand departure rates in percentage

Distribution case	Time range in simulation (min)							
	60–75	75–90	90–105	105–120	120–135	135–150	150–165	165–180
1	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5
2	25	25	25	25	0	0	0	0
3	15.73	15.43	15.03	14.39	13.03	10.37	8.36	7.66

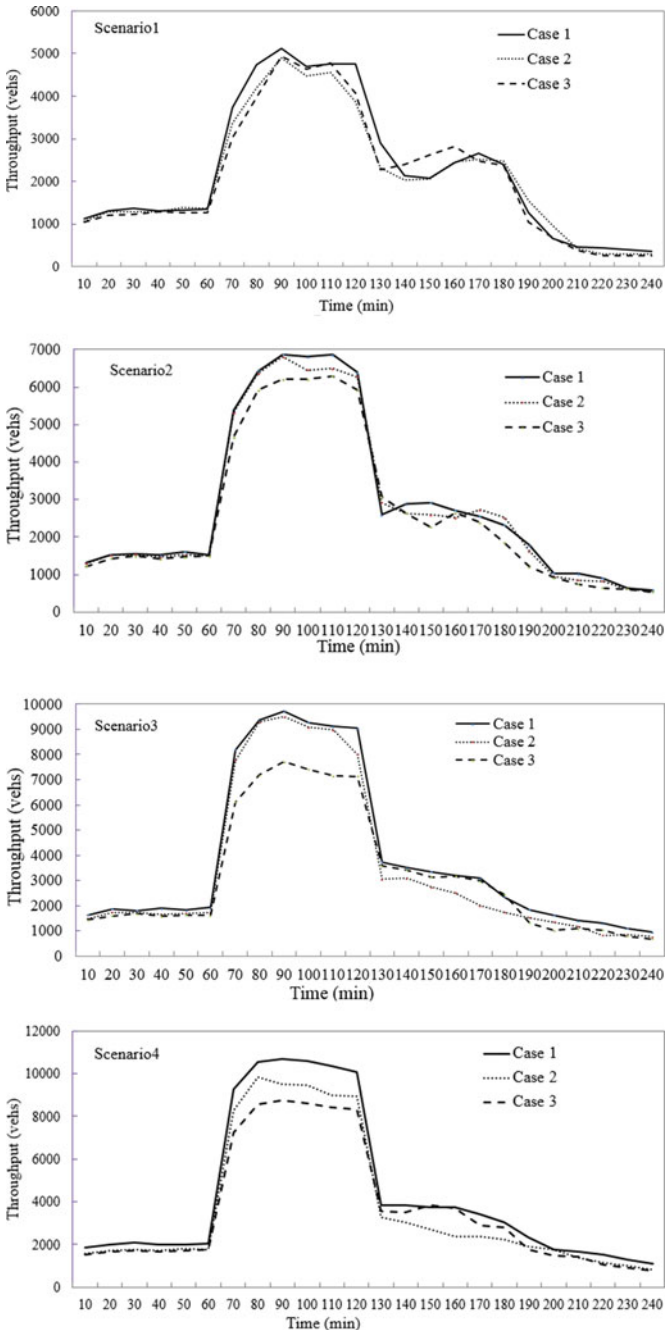


Fig. 9 Simulation results in step 3—vehicle throughputs

Table 4 Evacuation time and average travel time for staging

Demand scenario	Performance	Case 1	Case 2	Case 3
1	Evacuation time	115	150	150
	Average travel time	25.47	26.42	27.23
2	Evacuation time	155	175	>180
	Average travel time	26.17	28.35	27.25
3	Evacuation time	125	>180	>180
	Average travel time	27.04	29.75	27.04
4	Evacuation time	175	>180	>180
	Average travel time	27.80	30.28	27.80

20,000 vehicles in the stadium, the s-curve staging of evacuation demand could not accomplish 90 % of the evacuation traffic by the end of the simulation. At the other two demand scenarios, both the staging distributions Case 2 and Case 3 fail to evacuate 90 % of the evacuation traffic out of the PAZ by the end of the simulation. The simulation results may suggest the importance of staggered departures of evacuation trips even with a short-notice emergency evacuation.

7 Conclusions

In this study, a chlorine gas spill incident in Jackson downtown as a freight train with 30,000 gallon tanker derailed at Canadian National's North Jackson railroad yard was considered to justify an emergency evacuation. The traffic operations during the emergency evacuation over the road network for the affected population to leave the PAZ were simulated using the Dynamic Traffic Assignment (DTA) based program DynusT. Various traffic management strategies such as law enforcement for traffic controls, adding shoulder lanes, full and partial contra-flow operations, ITS/VMS deployment, and various staged evacuation demands were utilized and evaluated in order to improve the evacuation performance. The following conclusions can be drawn based on the simulation results:

(1) Input parameters were adjusted by the traffic volume calibration to satisfy all FHWA limits. The difference between the overall assignment volume for all links in the network and the traffic volumes provided by MDOT was within the ±5 % limit. The difference between the total assigned freeway volume and the observed total was within the ±7 % limit. The errors on specific links between the assigned hourly volumes and the observed ones were all within the ±12 % limit.

(2) For the three candidate arterial corridors with or without the traffic management strategies of contra-flow and VMS deployments, N State Street showed the best evacuation performance among all the three streets in all the tested evacuation scenarios by achieving the largest numbers of cumulative vehicles

evacuated out of the spill affected PAZ zone during the simulation period and was selected as the best evacuation corridor.

(3) For N State Street, deployments of full contra-flow, partial contra-flow, adding a lane, and staged departures of evacuation demands showed that the partial contra-flow deployment with a staged evacuation demand could achieve the best performance by decreasing the evacuation time and the average travel time at all the four evacuation demand scenarios.

(4) For N State Street, different staging distributions combined with the deployment of partial contra-flow strategy simulation results suggested that the importance of staggered departures of evacuation trips even with a short-notice emergency evacuation.

Acknowledgments The authors are grateful to the Mississippi Department of Transportation (MDOT) for funding the research study. MDOT engineers Acey Roberts and Wes Dean are thanked for their valuable support and guidance. The project was also partially funded by the Institute for Multimodal Transportation (IMTrans) at JSU through the UTC program of the USDOT.

References

1. Zimmerman C, Brodesky R, Karp J (2007) Using highways for no-notice evacuations: routes to effective evacuation planning primer series. Report No. FHWA-HOP-08–003, Federal Highway Administration, USDOT, Washington DC
2. Wolshon B (2001) One-way-out: contraflow freeway operation for hurricane evacuation. Nat Hazards Rev 2(3):105–112
3. Murray-Tuite P, Wolshon B (2013) Assumptions and processes for the development of no-notice evacuation scenarios for transportation simulation. Int J Mass Emerg and Disasters 31(1):78–97
4. Hardy M (2010) Structuring, modeling, and simulation analyses for evacuation planning and operations. Paper presented at the 89th transportation research board meeting, Washington DC, 10–14 Jan 2010
5. Qiao F, Ge R, Yu L (2009) Evacuation modeling in small and dense area. Paper presented at the 88th transportation research board meeting, Washington DC, 11–15 Jan 2009
6. Urbina E, Wolshon B (2003) National review of hurricane evacuation plans and policies: a comparison and contrast of state policies. Transport Res A-POL 37(3): 257–275. Wikimapia. CN/IC North Yard. http://wikimapia.org/6581747/CN-IC-North-Yard. Accessed 12 May 2013
7. Kim S, Shekhar S, Min M (2008) Contra-flow transportation network reconfiguration for evacuation route planning. IEEE T Knowl Data En 20(8):1115–1129
8. Kirchsteiger C (2006) Current practices for risk zoning around nuclear power plants in comparison to other industry sectors. J Hazard Mater 136(3):392–397
9. Theodoulou G, Wolshon B (2004) Modeling and analyses of freeway contraflow to improve future evacuations. Paper presented at the 83th Transportation Research Board Meeting, Washington DC, 11–15 Jan 2004
10. Wolshon B, Lambert L (2006) Planning and operational practices for reversible roadways. ITE J 76(8):38–43
11. Chien SI, Korikanthimath VV (2007) Analysis and modeling of simultaneous and staged emergency evacuations. J Transp Eng 133(3):190–197

12. Dixit VV, Radwan E (2009) Hurricane evacuation: origin, route and destination. J Transp Safety and Security 1(1):74–84
13. Mitchell SW, Radwan E (2006) Heuristic priority ranking of emergency evacuation staging to reduce clearance time. Transport Res Rec 1964:219–228
14. Noh H, Chiu Y, Zheng H et al (2009) An approach to modeling demand and supply for a short-notice evacuation. Paper presented at the 88th transportation research board meeting, Washington DC, 11–15 Jan 2009
15. Ozbay K, Yazici MA, Chien S I-Jy (2006) Study of the network-wide impacts of various demand generation methods under hurricane evacuation conditions. Paper presented at the 85th transportation research board meeting, Washington DC, 22–26 January 2006
16. Sbayti H, Mahmassani HS (2006) Optimal scheduling of evacuation operations. Transport Res Rec 1964:238–246
17. Dixit V, Wolshon B, Montz T (2011) Evacuation traffic dynamics and development of maximum sustainable evacuation traffic flow rates. Paper presented at the 90th transportation research board meeting, Washington DC, 23–27 Jan 2011
18. Songchitruksa P, Henk R, Venglar S et al (2012) Dynamic traffic assignment evaluation of hurricane evacuation strategies for the Houston-Galveston, Texas, Region. Transp Res Rec 2312:108–119
19. Abdelgawad H, Abdulhai B (2009) Optimal spatio-temporal evacuation demand management: methodology and case study in Toronto. Paper presented at the 88th transportation research board meeting, Washington DC, 11–15 Jan 2009
20. Chiu YC, Zhou L, Song HB (2010) Development and calibration of anisotropic mesoscopic simulation model for uninterrupted flow facilities. Transport Res B-Meth 44(1):152–174
21. Zheng H, Chiu YC, Mirchandani PB et al (2010) Modeling of evacuation and background traffic for optimal zone-based vehicle evacuation strategy. Transp Res Rec 2196:65–74
22. Zhou X, Taylor J (2012) DTALite. https://code.google.com/p/nexta/. Accessed 2 Jan 2012
23. NOAA. http://response.restoration.noaa.gov/aloha. Accessed 12 May 2013
24. William AM, McGuckin NA (1998) NCHRP report 365: travel estimation techniques for urban planning. Trans Res B Washington DC
25. Corporation Caliper (2012) The user's guide to TransCAD 5 transportation planning software. Caliper Corporation, Newton MA
26. The Federal Highway Administration (2010) Calibration, Traffic Analysis Toolbox Volume IV: Guidelines for Applying CORSIM Microsimulation Modeling Software. http://ops.fhwa.dot.gov/trafficanalysistools/tat_vol4/sec5.htm. Accessed 25 May 2012
27. Li C, Wang F (2011) Emergency evacuation study for the greater jackson area: evacuation traffic from New Orleans. Research Report for State Study 210, Mississippi Department of Transportation, Jackson, Mississippi

Analyzing Network Log Files Using Big Data Techniques

Víctor Plaza-Martín, Carlos J. Pérez-González, Marcos Colebrook, José L. Roda-García, Teno González-Dos-Santos and José C. González-González

Abstract The IT Department of the Universidad de La Laguna (ULL, Tenerife, Spain) provides service to 26 buildings with more than 1,000 network devices (wireless and wired), and access to more than 10,000 devices (computers, tablets, smartphones, etc.) which generate around 200 MB/day of data that is stored mainly in the DHCP log, the Apache HTTP log, and the WiFi log files. Within this context, the chapter addresses the design and development of an application that uses Big Data techniques to analyze those log files in order to track information on the device (date, time, MAC address, and georeferenced position), as well as the number and type of network accesses for each building. In a near future, this application will help the IT Department to analyze all these logs in real time.

Keywords Aerospace industry · Big Data · Earned value management · Hadoop · Log file analysis · R project · TOGAF

V. Plaza-Martín · T. González-Dos-Santos
Grado de Ingeniería Informática, Universidad de La Laguna, Tenerife, Spain
e-mail: plazamartin.victor@gmail.com

T. González-Dos-Santos
e-mail: tenoglez@gmail.com

C.J. Pérez-González
Departamento de Matemáticas, Investigación Operativa y Computación,
Universidad de La Laguna, Tenerife, Spain
e-mail: cpgonzal@ull.edu.es

M. Colebrook (✉) · J.L. Roda-García
Departamento de Ingeniería Informática y de Sistemas,
Universidad de La Laguna, Tenerife, Spain
e-mail: mcolesan@ull.edu.es

J.L. Roda-García
e-mail: jlroda@ull.edu.es

J.C. González-González
Servicio de Tecnologías de la Información y de la Comunicación,
Universidad de La Laguna, Tenerife, Spain
e-mail: jgonzal@ull.edu.es

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_10

1 Introduction

By the time you read this chapter, over 16 TB of data have been generated in the world every second, as shown in [1]. This represents that in 2016, global IP traffic will reach 1.1 zettabytes (ZB) per year, or 88.4 exabytes (EB, nearly one billion gigabytes) per month, or nearly 3 EB per day, which is certainly a huge amount of data.

According to [2], in 2013 four zettabytes (1 ZB = 10^9 TB = 10^{21} bytes) of data were created by digital devices. In 2017, it is expected that the number of connected devices will reach three times the number of people on earth.

Hence, the main topic of the book, namely Big Data, is justified based on the current technological situation in which data is generated at a higher speed than can actually be processed, and large companies are facing the problem of deleting data due to the impossibility to store it, thus losing useful information.

In fact, one of the main sources of data are the log files, which record either events that occur in an operating system or other software runs, or messages between different users of a communication software.

Thus, an opportunity arose to work collaboratively with the Information Technology and Communication Department (STIC in Spanish) at the Universidad de La Laguna. The STIC provides service to 26 buildings with more than 1,000 network devices (wireless and wired), and renders access to more than 10,000 devices (computers, tablets, smartphones, etc.), which generate around 200 MB/day of data that is stored mainly in the DHCP (Dynamic Host Configuration Protocol) log, the Apache HTTP log, and the WiFi log files.

The key problem was the need to monitor the access made from a single device, or the user's activities all around the campus. In this sense, they wanted to infer usage patterns throughout the day in order to strengthen the WiFi network at certain points.

Within this context, the chapter addresses the design and development of an application that uses Big Data techniques to analyze those log files in order to track information on the device (date, time, MAC address, and georeferenced position), as well as the number and type of network accesses for each building. In a near future, we believe that this application will help the STIC to analyze all these logs in real time.

Although there are several applications focused on log management in the market (see Table 1), the STIC wanted a custom tailored application since their log files were not in the standard format, as we explain in the next sections. For a detailed comparison of open-source log management solutions, the reader is referred to [3].

The remainder of the chapter is organized as follows. In Sect. 2, we provide the definition of Big Data, its influence and relevance nowadays, as well as a description of the Hadoop framework. Section 3 presents the problem description using TOGAF. The project development and the working methodology are

Table 1 List of log management and analysis tools

Tool	Description	Type of license
ELK Stack	Set of applications and utilities (Logstash, Elasticsearch, Kibana) to create a powerful search and analytics platform	Free
Graylog2	Log management system with server and a web interface	Free
Logentries	Real-time log management and analytics	Free/Commercial
Loggly	Cloud-based log management service	Free/Commercial
Logscape	Allows searching, visualizing and analyzing log files from a central dashboard	Free/Commercial
Splunk	Industry-leading platform that automatically indexes log data	Commercial
Logtrust	Turns machine data into business insights	Commercial

described in Sect. 4. In Sect. 5, we present and discuss the results for each developed task, along with some charts depicting the Hadoop cluster indicators. Finally, the conclusions are provided in Sect. 6.

2 Big Data State-of-the-Art

As we stated in the introduction, every day nearly 3 EB (1 EB = 10^6 TB) of data is generated [1]. IBM [4] pointed out that the main sources for this entire data arise from the following:

- Web data and social media
  - Web content
  - Twitter feeds
  - Facebook postings
  - Clickstreams (data from user navigation)
- Data generated from M2M (machine-to-machine) communication
  - GPS signals
  - RFID readings
  - Intelligent readers
- Great data transactions
  - Government
  - Business
- Biometrics
  - Facial recognition
  - Genetics

- Data generated directly from human beings
  - Emails
  - Voice recordings
  - Records of all kinds

Another interesting approach is given by [5], who associated the data types (structured or unstructured) to the data source (internal or external), asserting that the unstructured external data is the largest area of opportunity for the enterprise (see Fig. 1).

Bloem et al. [6] summarized these sources by relating the increasing data variety and complexity with its volume (in bytes), from the ERP systems to the current Big Data scenario, going through the CRM and Web systems (see Fig. 2). As a consequence, Big Data is defined as a variable equal to:

$$Transactions + Interactions + Observations$$

Nevertheless, the definition of Big Data that is the most widely accepted was stated by Doug Laney, industry analyst from Gartner, in 2001 when he established the three V's of the Big Data [7]: Volume, Velocity and Variety.

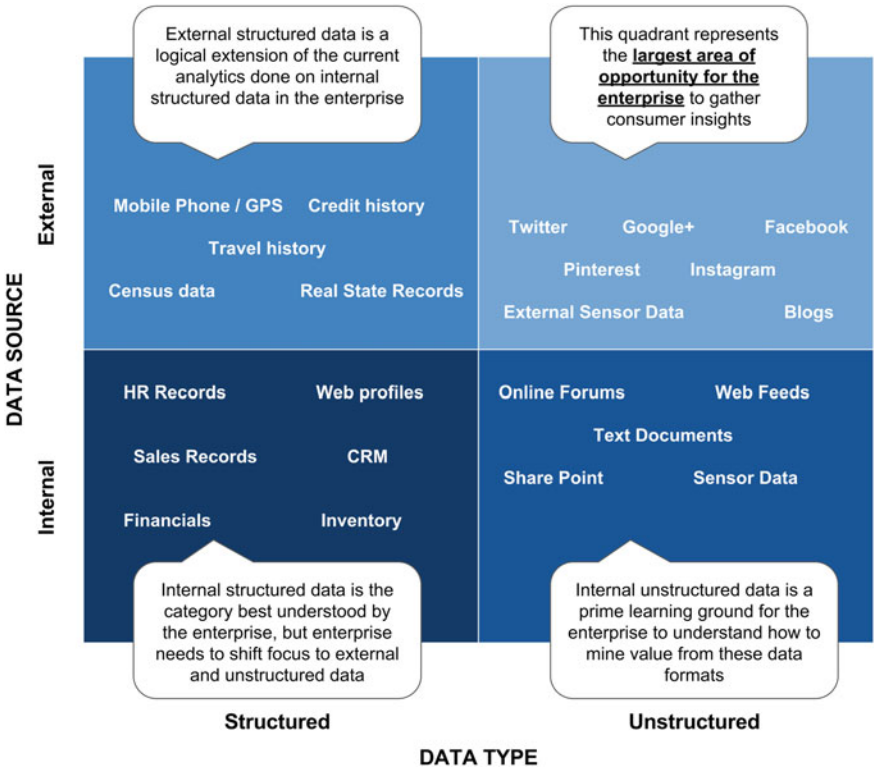


Fig. 1 Data Types versus Data Sources (adapted from [5])

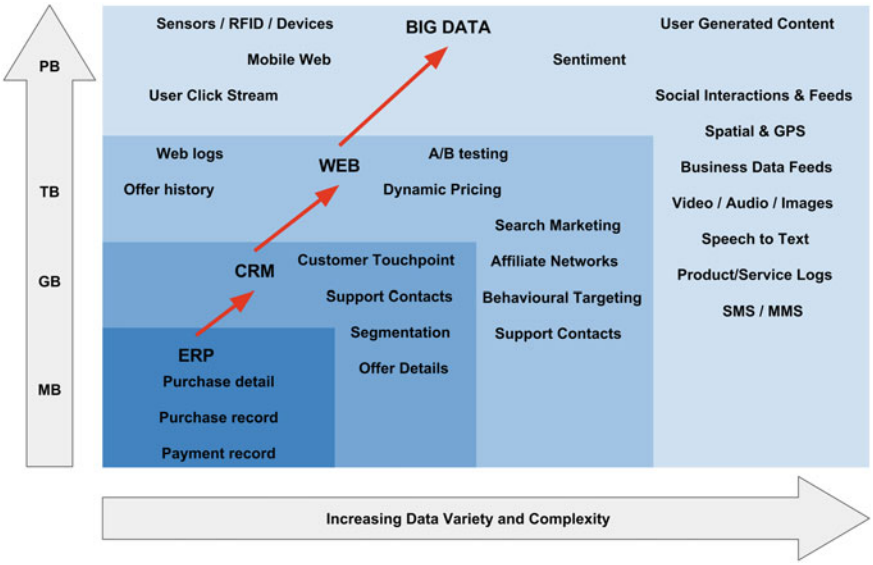


Fig. 2 From ERP to Big Data (adapted from [6])

Volume

The volume refers to the amount (size) of the datasets generated and the difficulty that companies and organizations have to process them. The forecasts are that in 2020, 25 billion devices will be connected to the Internet, making data grow exponentially, multiplying by 10 the amount in a period of just 5 years.

Velocity

The speed is often misunderstood as a real-time analysis, but it also refers to the rate of change (or flow) in the data, linking data coming at different speeds, and also to the activity peaks.

Variety

This is perhaps the most interesting characteristic to analyze, since data can arise from all areas of daily life, including multiple repositories, domains or types. However, most of this data already belongs to organizations, but they make no use of it, becoming what Doug Laney called *dark data*. This data, despite not being useful by itself, can be analyzed and processed to generate useful and valuable information, and hence, opening new business opportunities. Depending on the type of data, we can subdivide it into four main groups:

- Structured
- Semi-structured
- Unstructured
- Complex structured

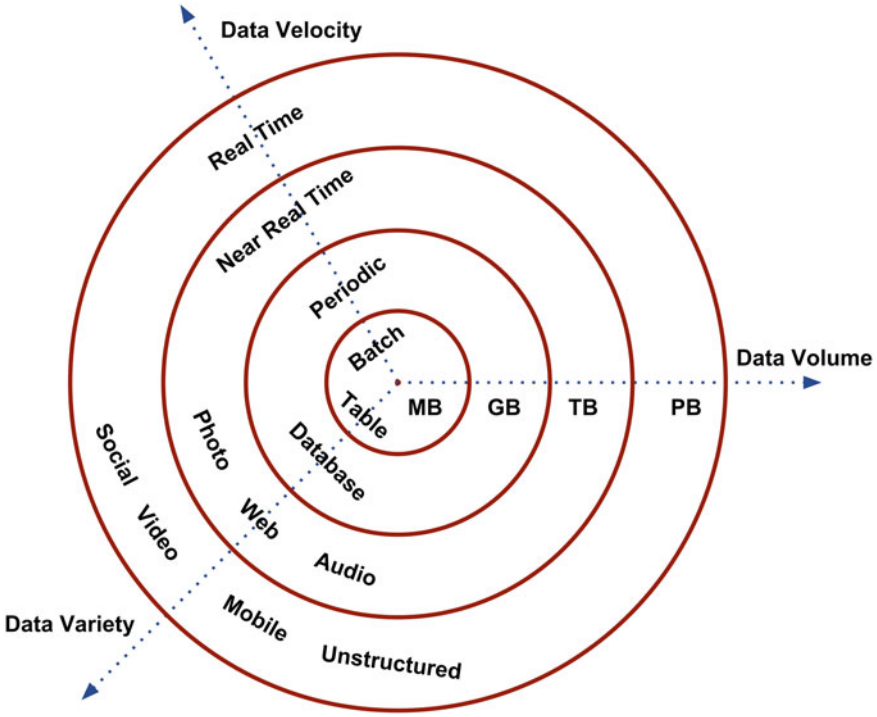


Fig. 3 The three V's that define Big Data (adapted from [8])

Besides, the terms **variability** and **complexity** also arise as possible features concerning Big Data, since data might be generated inconsistently, with huge peaks from multiple sources that makes very difficult to relate them.

The following Fig. 3 adapted from [8] graphically summarizes the three characteristic V's of Big Data.

Yiu [9] also contributed with another definition of Big Data: *"datasets that are too awkward to work with using traditional, hands-on database management tools*. Besides, he also defined the term Big Data Analytics as *the process of examining and interrogating big data assets to derive insights of value for decision making*."

For the TechAmerica Foundation [10], Big Data is a term that describes "*large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information*."

One of the latest definitions comes from the National Institute of Standards and Technology [11], who considers Big Data as "*consisting of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis*."

The main idea behind all these definitions of Big Data is to analyze the data sets to gain insight and valuable information for the final decision making. In this sense,

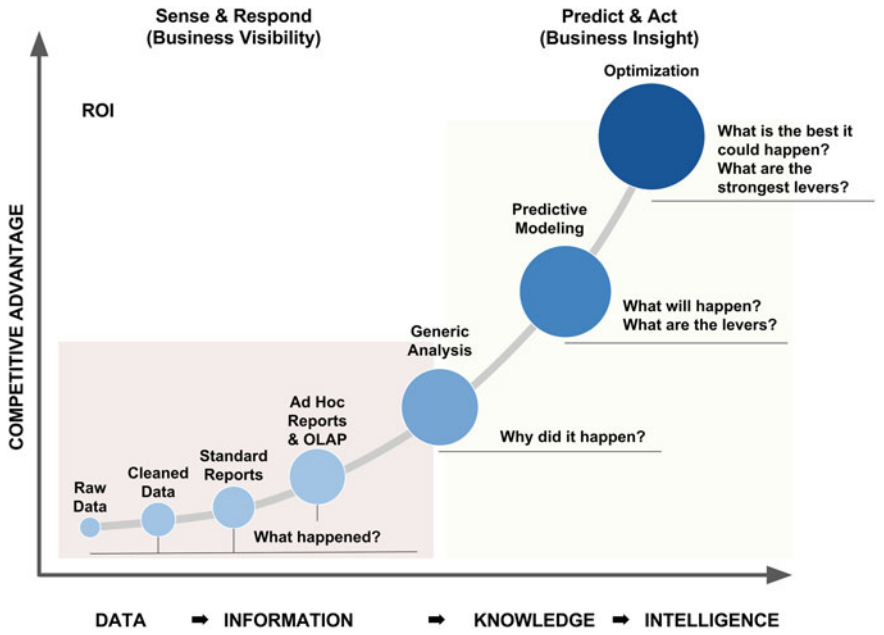


Fig. 4 From "Sense & Respond" to "Predict & Act" (adapted from [13])

Davenport and Harris [12] and SAP [13] suggest that businesses should change as soon as possible from a "Sense & Respond" to a "Predict & Act" strategy, in which Big Data plays an important role in all the steps towards the main goal (see Fig. 4).

Finally, and before presenting the Hadoop framework in the next section, just to remark that the Telecommunications Standardization Sector of the International Telecommunication Union (ITU-T) has recently released "Recommendation Y.3600", which provides requirements, capabilities and use cases of cloud computing based Big Data as well as its system context [14]. This is the first attempt to make an international standard on Big Data, and we believe that it will truly provide solid foundations for the future design and development of applications on this subject.

2.1 The Hadoop Framework

For some time ago, the term Big Data has been directly involved with the use of the Hadoop framework because it allowed analyzing unstructured data easily and quickly. The Hadoop MapReduce framework [15] is a free software solution that supports distributed applications, and is currently the most widely used solution for leading companies such as Yahoo, Ebay, Facebook, IBM, etc., since it allows working with thousands of nodes and petabytes of data.

Hadoop is usually referred as an ecosystem, due to the huge number of projects, services, libraries, APIs, etc., that gathers around. However, its architecture is mainly composed of these elements:

- Hadoop Common: the set of Java libraries and tools required by the different modules to access the filesystem and to start the cluster daemons.
- HDFS (Hadoop Distributed File System): basically a distributed file system, scalable and portable, characterized by its high throughput access to application data and the suitability for applications that have large data sets. An HDFS cluster consists of:
  - A single NameNode, that is a master server that manages the file system namespace and controls the file access by the clients.
  - In addition, there are a number of DataNodes, usually one per physical node in the cluster, which manages the storage attached to them.
- Hadoop YARN: the framework to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. YARN frees the MapReduce engine from cluster resource management tasks, streamlining data processing and the execution of the tasks. This data-computation framework consists of:
  - The ResourceManager: ultimate authority that arbitrates resources among all the applications in the system.
  - The NodeManager: framework agent who is responsible for containers, monitoring their resource usage (CPU, memory, disk, network) and reporting to the ResourceManager/Scheduler.
- MapReduce: a JobTracker where client applications send jobs.
- A set of parallel applications that enhance the Hadoop ecosystem, such as Pig, HBase, Hive, Hue, etc.

To develop this project we used Hadoop 2.0, which is a great update to the architecture of Hadoop 1.0, as shown in Fig. 5.

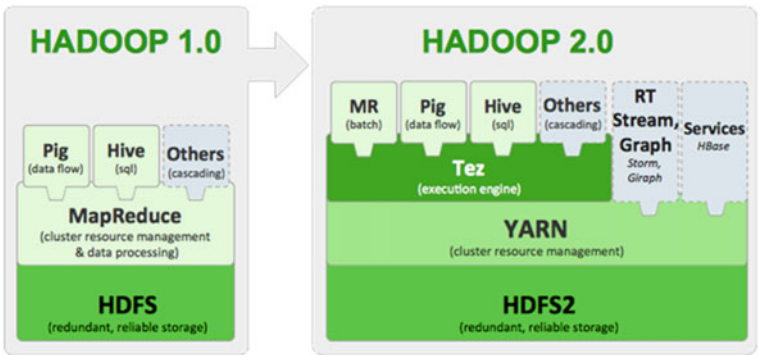


Fig. 5 Architectural update in Hadoop 2.0

It is noteworthy to remark that the HDFS system is specially designed to handle large files. Besides, it is responsible for splitting and distributing the contents of large data files into blocks or chunks of 64 MB (by default), generally across multiple machines (nodes). Its reliability is based on the replication of data blocks across three different hosts (by default). Hence, the hosts can communicate among them and rebalance the flow of data, move data and maintain a high replication within the cluster to ensure the proper performance of the application.

The inner working scheme in Hadoop MapReduce is a sequential process in which the JobTracker partitions large data sets stored in the HDFS system into smaller blocks of 64 MB in the NameNode (master node). Then, these blocks are distributed to the DataNodes (slave nodes) to perform the reading and mapping to generate the output as a key-value tuple that is stored in a temporary destination. At this stage, the Shuffler sorts those tuples by the key and subsequently all elements with the same key in the same node remain together. In the last step, the DataNodes gather all the tuples with other partial results, to generate the final output file. Even though the MapReduce process is completed, one last step is usually added to visualize the data, such as using the statistical software R and the Shiny framework.

For a deeper knowledge using the Hadoop framework for addressing Big Data challenges, the reader is referred to Hu et al. [16]. Once we have introduced the working framework, we are now ready to describe the problem in the next section.

3 Problem Description

As we stated in the introduction, the STIC is in charge of the management and maintenance of the university's network, which comprises more than 100 telematical services accessible from 26 buildings, along with more than 1,000 network devices (wired and WiFi).

Previous to this project, the STIC had an application developed with Pentaho Business Intelligence Community [17]. The main goal of this tool was to combine the information obtained from various sources, mainly the access log from the Apache server, the DHCP log file, and the log files generated by the WiFi device drivers, in order to determine the total number and type of accesses for each building.

Although the application performed relatively well without the chance to apply any filter, the running time spent to get the results for the network's least used time slot for a single day was nearly three days of execution. Thus, we settle the possibility of carrying out a new process using Big Data tools to significantly reduce the running times.

Therefore, and before developing the new solution, we first modeled the log system using architectural framework tools.

3.1 Modeling the WiFi Log System

Current modern ICT (Information and Communications Technology)-intensive institutions need to use business-modeling techniques that relate the business goals with the technology evolution. In the last 20 years, new ways to interconnect company internal systems using new enterprise architecture frameworks and technologies have appeared.

Regarding this matter, The Open Group Architecture Framework (TOGAF) is one of the most important enterprise architecture frameworks developed since the 1990s. In its latest version 9.1, TOGAF [18] provides a method and a set of good practices that allow a good relationship between the overall systems of an enterprise and its stakeholders.

Businesses evolve rapidly to consider new stakeholders and technical requirements. Institution goals and emerging technologies must be aligned accordingly, but daily operational tasks make this issue a hard job to deal with, whereas daily changes make business management a complex activity. The continuous transformation from the current system to the required one is maintained under the TOGAF methods and good practices. Two of the most important concepts in TOGAF are the views and the viewpoints, respectively, which allow the enterprise architect to focus in different parts of the complete business. As stated in TOGAF documentation [18], we can use the framework to model all the business or we can concentrate in different parts of it.

In this chapter, we model the WiFi log information system of the ICT Department (STIC) of the Universidad de La Laguna. We have used two different views to define the main goals: the layered structure of the system and the application solution view.

For this task, we have used Archimate (TOGAF 2011), which is an enterprise modeling language that captures the complexity of different domains and its relations within the organization. Archimate is a modeling tool that will increase the representation and comprehension of the enterprise modeled systems. Hence, the enterprise architect models different parts of the institution focusing on the elements he/she wants to emphasize considering many other aspects of the system. Besides, by using Archimate views the architect is able to analyze a specific problem bearing in mind the complete system.

In Archimate, the service concept plays a central role and can be defined as a unit of functionality that an entity offers to other entities or environment. One of the most important viewpoint diagrams is the Layered viewpoint [19]. The main goal of the Layered viewpoint is to provide an overview in one diagram. This view consists of three layers: Business layer, Application layer and the Technology layer.

The Business layer offers products and services to external customers that run business processes performed by business actors. The Application layer gives support to the business layer with application services that are performed by applications. And finally, the Technology layer offers infrastructure services accomplished by computers and communication hardware and software.

Relationships between layers are formed by the use relation, which shows how the higher levels use the services of the lower layers. A second type of link is the realization relation that shows which element in the lower layer realizes comparable elements in higher layers.

3.2 Solution Achieved

We analyzed the WiFi log system and, although there exist many relationships between different business domains (login, academic, human resources, etc.) we have centered our efforts in the main goals that STIC's staff defined:

- Track all network communication elements (routers, switches, servers, etc.)
- Fast support to network failures
- Track specific MAC or IP elements in the network
- Detect communication bottlenecks
- Geolocated network status visualization
- Provide better network services

3.2.1 Layered Viewpoint

Through the Layered viewpoint diagram, we propose a viewpoint based on layers or views, namely a BAI scheme (Business—Application—Infrastructure). Thus, we can make an approach from different perspectives or viewpoints based on our interests and subsequently, obtaining both a global representation of the solution developed, as well as the current application by only changing the view.

Starting with the infrastructure layer, we can see in Fig. 6 that consists of two basic elements. The first one is a set of computers located in the Computer Science School connected via LAN, which provide hardware support in our Big Data solution. On this infrastructure we have installed Hadoop 2.6.0 for Linux [15], running both as NameNode and DataNode. One of the computer plays the role of master of the cluster while the rest of nodes have the DataNode settings and play the role of slaves.

On the other hand we have a Windows based computer holding an Apache HTTP server, along with R [20] as the statistical analysis software.

Hadoop provides a number of services such as Streaming, HDFS, YARN and JobTracker, which will all be used by the application component MapReduce, whereas R provides RStudio [21] and ggplot services to Shiny [22].

As stated above, at the application level we have two basic components. The first one is the MapReduce component, which comprises the Mapper, Shuffler and Reducer services running sequentially in a pipeline mode. This application component generates a CSV output file that serves as input to the other application component (Shiny), and hence, establishing a collaboration between these two components.

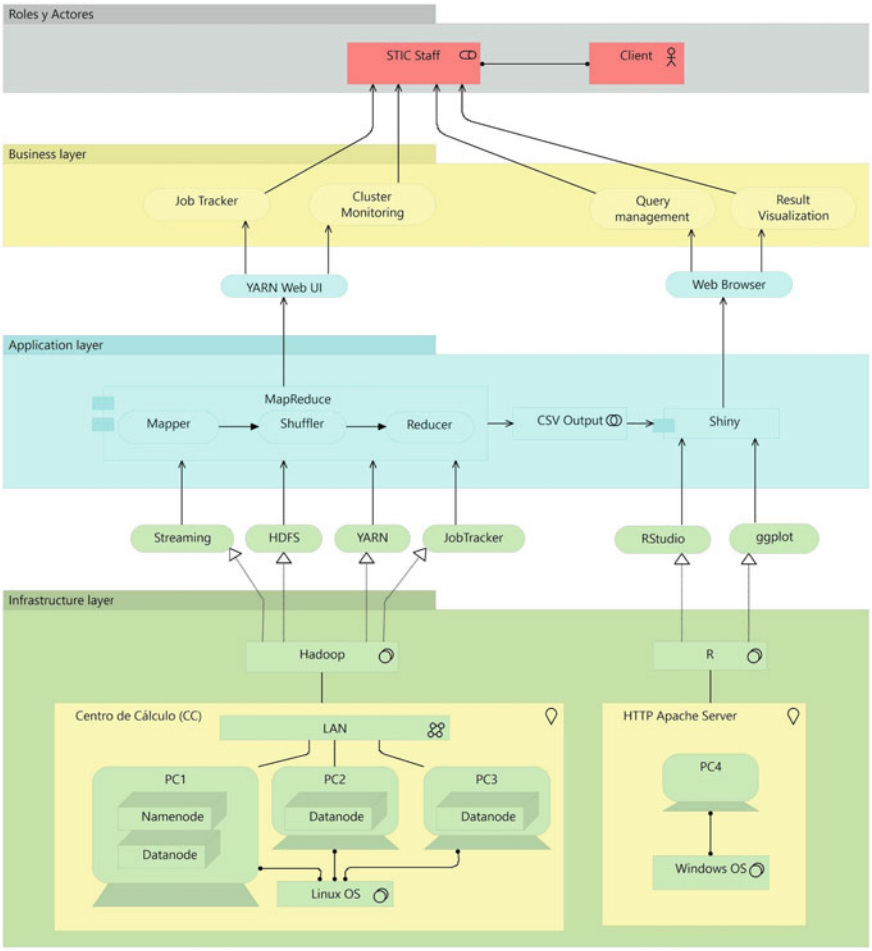


Fig. 6 The layered viewpoint

At the business level, we can see that the user (STIC's staff) is presented with three basic actions: the monitoring of the tasks and the cluster (both through the user interface provided by YARN), and viewing the results and managing the queries through a small dashboard.

3.2.2 Application Behavior Viewpoint

If we now change our view and perform an analysis at the application level focusing on the MapReduce component (see Fig. 7), we can see that is composed by two other MapReduce components: one for analyzing the DHCP logs, whereas the other one integrates this information with the analysis of the Apache logs.

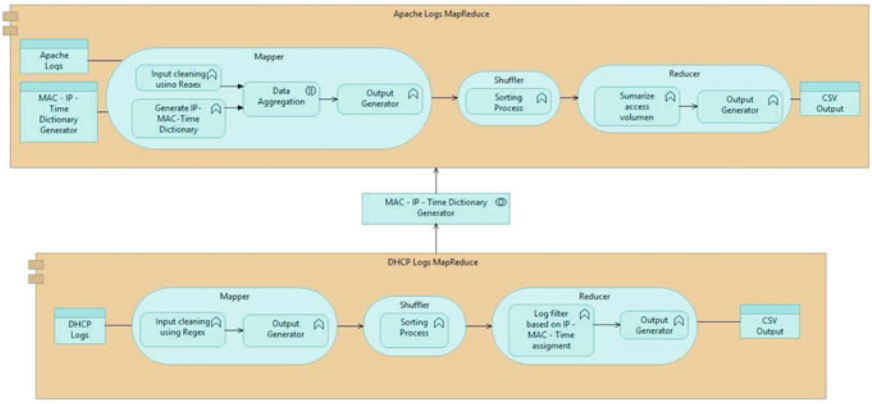


Fig. 7 The application view

The MapReduce process responsible for the analysis of the DHCP logs starts cleaning the input data using a regular expression. Thus, we obtain the set of useful data and, in turn, we filter those log lines that are not needed for the analysis. All this is accomplished by using generators and iterators in Python, so that a certain element is not generated until you actually need it. These programming techniques imply less computational cost and memory usage. Once we have filtered this information, we proceed to generate the data structure with that task, in this case a dictionary within a dictionary, so that the IP is the key of the generic dictionary and its value a dictionary where the key is the time and the MAC is the value. This allows us to clearly identify the use of an IP address at any time and what device is making use of it.

The Shuffler sorts all the information generated on each of the cluster nodes within the mapping process, so as to ensure that all entries with the same key are computed in the same DataNode, and therefore the integrity of the solution is assured.

Now the Reduce process begins to synthesize all the DHCP information to a minimal set, storing a record of each time an IP assigned to MAC changed, instead of the whole entry. Consequently, the output volume is reduced by 58 %, even though it contains the same useful information. This aspect is key since the MapReduce process responsible for processing the Apache Log files will use this file later.

The second MapReduce process is similar to the first one in its concept, because it is based on cleaning the input data through regular expressions to get the smallest unit of information required. Later on, we check the dictionary generated as a result of the above process for the DHCP analysis, and we add both data if needed.

Thus, we relate each useful access to its corresponding MAC address, adding georeferencing information generated by the STIC's staff based on the architecture of their WiFi and wired network.

4 Project Development

After several meetings with the STIC's staff, some restrictions were set, which shaped the structure of this project, emphasizing among them the use of Python as the programming language in order to ease the maintenance of the solution proposed.

Bearing always in mind that performance was the top priority, different alternatives were analyzed and compared in order to ensure the best results were obtained in such diverse aspects as reading/writing speed from both files and the HDFS system, processing time, functional requirements, etc.

Hence, and using as a baseline the comparative generated by Laserson [23], we concluded that the best option was to develop the solution through the Streaming library that has performance ratios and a range of features similar to Java.

4.1 Working Methodology

One of the biggest challenges when working with large amounts of data is to verify the accuracy of the results obtained, and this was a critical aspect for the STIC's staff.

Accordingly, a working methodology was designed in which an incremental development of the solution would allow controlling the results from a large enough input dataset.

Therefore, the development process was split into milestones (or tasks), each of them with a defined, well-documented objective, and a series of data files as inputs for analysis. Thus, it was concluded that the working planning would be as follows (the 'V' before the number stands for the version):

V0.5 Analysis of a log file folder

The system with a single node will compute the data in a directory, traverse all log files inside, and produce a CSV file as the output with the names of the files found and the number of log lines detected in each one.

V1.0 Analysis of a single log file with a single node

For one single log file of a single virtual server from the campus, get the number of accesses grouped by hour.

V1.5 Analysis of a single log file with three nodes

For one single log file of a single server from the farm of virtual servers, get the number of accesses grouped by hour, splitting the computation into three nodes.

V2.0 Analysis of several one-day log files on multiple nodes

For the set of log files in a single day of multiple servers in the campus, obtain the number of accesses grouped by hour, for each server and for each time slot.

V2.5 Analysis of several one-day log files on multiple nodes with date and time restrictions

For the set of log files from a single day of multiple servers in the campus, obtain the number of accesses grouped by hour, for each server and for each time slot, and filter the restrictions from-day/from-time and to-day/to-time specified in a special file.

V3.0 DHCP log analysis

We must determine the MAC devices that connect to the network at any time. The system will read the log files from the DHCP servers and generate a new output file with the MAC addresses associated to each IP address at each time instant.

V4.0 Linking unstructured data and correlating elements in different time instants

For a certain access log line to the virtual campus server, we have to determine the MAC address associated with the IP address that made that precise access. The output may now include the tuple: Date, Time, IP, MAC and Access Object.

V4.5 Deploying V4.0 for a period specified in the input filters

This task must link all the above sections and get the tuple of V4.0 for all log files for a given period of time in the input filter.

V5.0 Analyzing WiFi connections

Analyze the log files of the different WiFi device controllers to obtain the tuples: Date, Time, MAC, WiFi Access Point.

V5.5 Analyzing WiFi connections and georeferenced access

Analyze the georeferenced files for the WiFi access points, and link to the previous section the coordinates of the detected accesses.

V6.0 Linking all together

Link all previous sections to obtain the following tuple for a specific time period defined in the input filter: Date, Time, Virtual Campus Access, IP, MAC, Access Point, GPS coordinates.

5 Results

The main data sources are the semi-structured access log files that are automatically generated each time a user's device access the university's network. Currently, the amount of data generated daily as a result of network monitoring exceeds 200 MB/day.

This is mainly due to the extensive network, which provides access to approximately 10,000 devices and renders services to more than 26 buildings, with a total of more than 1,000 network devices (wired and WiFi).

```
Mar 16 12:04:32 udvweb1.stic.ull.es apache2_access: 85.59.43.201 - -
[16/Mar/2014:12:04:32 +0000] "GET /1314/theme/image.
php/institucional/core/1387354149/t/hide HTTP/1.1" 200 1496 "http:
//campusvirtual.ull.es/1314/course/view.php?id=637" "Mozilla/5.0 (Windows
NT 6.1; WOW64; rv:27.0) Gecko/20100101 Firefox/27.0"
```

Fig. 8 Sample line in an access log file

Besides, the access log files lack sensitive information and hence they need not be anonymized. Indeed, the files have the characteristic morphology of being a concatenation of an unsorted standard Apache log along with data added by the log management system of the STIC. The fields are separated by spaces, and the symbol "-" denotes the absence of data. An example of a log line in this type of files is shown in Fig. 8.

Before presenting the results obtained over a given set of server logs from 2014, first we must process these log files by running a Hadoop MapReduce program stored in the HDFS system within a cluster of nodes.

5.1 Cluster Configuration

A cluster of 21 data nodes (and 1 node manager) with 4 GB of RAM in each node was deployed with Hadoop YARN version 2.7. When submitting MapReduce procedures in Hadoop, we are interested in measuring the cluster performance and benchmarking the job executions. In this sense, Hadoop provides a high valuable Java API for debugging problems in MapReduce jobs, analyzing the use of memory and CPU time, and tracking the filesystem operations.

Among the most important resources within this Hadoop API are the counters, which are a wide selection of indicators and statistics reported by the individual tasks associated to job submissions. The counters in Hadoop are used to track the job progress of a MapReduce algorithm, and to study the events occurred during the job execution. White [24] describes five categories or groups of counters in Hadoop (see Table 2).

The group of task counters is updated as a task progresses, whereas the group of job counters are updated as a job progresses. Some of the built-in MapReduce counters are detailed in the following Table 3.

Table 2 Counter groups in Hadoop

Group	Java enum
Map Reduce task counters	org.apache.hadoop.mapreduce.TaskCounter
FileSystem counters	org.apache.hadoop.mapreduce.FileSystemCounter
Job counters	org.apache.hadoop.mapreduce.JobCounter
FileInputFormat counters	org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter
FileOutputFormat counters	org.apache.hadoop.mapreduce.lib.output.FileOutputFormatCounter

Table 3 MapReduce task counters

Counter	Description
MAP_INPUT_RECORDS	The number of input records consumed by all the maps in the job
MAP_OUTPUT_RECORDS	The number of map output records produced by all the maps in the job
MAP_OUTPUT_BYTES	The number of bytes of uncompressed output produced by all the maps in the job
PHYSICAL_MEMORY_BYTES	The physical memory being used by a task in bytes, as reported by/proc/meminfo
CPU_MILLISECONDS	The cumulative CPU time for a task in milliseconds, as reported by/proc/cpuinfo
GC_TIME_MILLIS	The elapsed time for garbage collection in tasks in milliseconds (reported by GarbageCollectorMXBean. getCollectionTime())

Table 4 Filesystem counters

Counter	Description
BYTES_READ	The number of bytes read by the filesystem by map and reduce tasks. There is a counter for each filesystem (local filesystem, HDFS, etc.)
BYTES_WRITTEN	The number of bytes written by the filesystem by map and reduce tasks
READ_OPS	The number of read operations (i.e. open, file status) by the filesystem by map and reduce tasks
WRITE_OPS	The number of write operations (i.e. create, append) by the filesystem by map and reduce tasks

Table 5 Job counters

Counter	Description
TOTAL_LAUNCHED_MAPS	The number of map tasks that were launched
TOTAL_LAUNCHED_REDUCES	The number of reduce tasks that were launched
NUM_FAILED_MAPS	The number of map tasks that failed
NUM_FAILED_REDUCES	The number of reduce tasks that failed
NUM_KILLED_MAPS	The number of map tasks that were killed
NUM_KILLED_REDUCES	The number of reduce tasks that were killed
MILLIS_MAPS	The total time taken running map tasks, in milliseconds
MILLIS_REDUCES	The total time taken running reduce tasks, in milliseconds
MB_MILLIS_MAPS	The total time taken running map tasks multiplied by RAM allocated, in miliseconds*megabyte
MB_MILLIS_REDUCES	The total time taken running reduce tasks multiplied by RAM allocated, in miliseconds*megabyte

Likewise, several built-in filesystem counters can be outlined as well (see Table 4).

The job counters measure statistics at the job level, and their values are not changed while a task is running. Some significant counters in this group are shown in Table 5.

5.2 Sample Results for Each Task

Next, we show a sample of the MapReduce output for each stage of the project on the input data, namely, the access log files, the DHCP log files, and a set of filter files or static information such as location coordinates of the IP addresses within the campus.

V0.5 Analysis of a log file folder

Given a set of access log files, we obtain the following Table 6

V1.0 Analysis of a single log file with a single node

The fragment of the generated output is shown in Table 7.

V1.5 Analysis of a single log file with three nodes

The output generated contains the same format as seen in Table 7, but the input now comes from three files, one for each node, in order to verify that the aggregation between node output files works properly.

V2.0 Analysis of several one-day log files on multiple nodes

The output shown in Table 8 counts the number of accesses per IP/date/server, and then computes a total result by date/IP.

Table 6 Output of task V0.5

Filename	Number of lines
access.log-20140317	318599
access.log-20140318	527553
access.log-20140319	513896
access.log-20140320	505612
Total number of lines	1865660

Table 7 Output of task V1.0

Month	Day	Hour	IP	Number of accesses
Mar	17	10	10.212.2.230	843
Mar	17	10	10.212.2.94	68
Mar	17	10	10.212.3.189	59
Mar	17	10	10.212.3.212	1357
Mar	17	10	10.212.3.233	277

Table 8 Output of task V2.0

Month	Day	Hour	Ip	udvweb1 log-20140318	udvweb2 log-20140318	udvweb3 log-20140318	udvweb4 log-20140318	Total
Mar	17	8	10.219.3.69	0	0	19	0	19
Mar	17	8	10.219.8.238	6	914	0	0	920
Mar	17	8	10.219.8.239	0	1522	0	10	1532
Mar	17	8	10.219.8.242	176	1	0	0	177
Mar	17	8	10.219.8.243	0	0	0	32	32

V2.5 Analysis of several one-day log files on multiple nodes with date and time restrictions

The output of this task is identical to Table 8, but in this case we use a custom filter file with timeslots to restrict the time of the accesses. This filter file has the following structure:

```
From Mar 17 10:00:00
Until Mar 17 17:00:00
```

V3.0 DHCP log analysis

In this case, we filter according to file containing IP ranges associated with different logical origins within the network management. This filter file has the following contents:

```
192.168.x.x Danger1
172.16.x.x Danger2
10.x.x.x Internal
193.145.96.x PublicInternal
193.145.97.x PublicInternal
193.145.120.x Administration
193.145.125.x PublicInternal
*.ull.es InternalWithDNS
Others External
```

Table 9 shows a sample of the output generated in this task.

Table 9 Output of task V3.0

Month	Day	Hour	Admin.	External	Internal	Internal with DNS	Public internal	Danger1	Danger2	Total
Mar	17	6	0	4686	103	128	0	0	0	4917
Mar	17	7	0	20003	3495	384	0	0	0	23882
Mar	17	8	0	44578	64334	384	0	0	0	109296
Mar	17	9	0	67733	91041	384	0	0	0	159158

Table 10 Output of task V4.0

Month	Day	Hour	Server	Access type	IP address	MAC address
Mar	18	11:34:32	udvweb1.stic.ull.es	apache2_access	10.225.2.207	00:17:31:c4:30:71

Table 11 Output of task V5.0

Date	Hour	MAC	Access point
01/03/2016	8	88:9f:fa:93:35:7c	10.174.3.28

Table 12 Output of task V5.5

Building code	Building name	Network type	Netmask	Building GPS coordinates
BA2	Bellas Artes Nuevo	Red Azul	10.101.0.0/24	28.461210, –16.276267

Table 13 Output of task V6.0

Date	Access point	Device MAC	Building	Building GPS coordinates
01/03/2016 8:38	10.174.3.28	88:9f:fa:93:35:7c	PE Periodismo Red WIFI	28.469314, –16.301921
01/03/2016 8:56	10.158.0.207	20:64:32:4c:44:94	IA IUBO + Agricolas Red WIFI	28.481374, –16.319417

V4.0 Linking unstructured data and correlating elements in different time instants

The goal is to analyze the DHCP log file in order to have a baseline to gather all the data previously analyzed from the V5.5 task on (see Table 10).

V4.5 Deploying V4.0 for a period specified in the input filters

This output is similar to task V2.5 but over another data set, whereas the structure of the output is the one obtained in task V4.0 but considering some time filters.

V5.0 Analyzing WiFi connections

The output is similar to task V4.0 but for WiFi connections (see Table 11).

V5.5 Analyzing WiFi connections and georeferenced access

Given a set of files containing information related to the logical design of the WiFi network, we generate an output with the format shown in Table 12. This output format will help to link all previously developed tasks with the goal described in task V6.0.

V6.0 Linking all together

The output of this task (shown in Table 13) was from the beginning the main goal of the project, that is, to aggregate access data with DHCP information and georeferenced positions in order to track a particular device, and to analyze both the network traffic flow and the network load.

5.3 Dashboards Using R Charts and Data from Counters

The tasks described in Sect. 4 were executed on four instance problems to analyze the server logs of several months in 2014. In particular, the following months

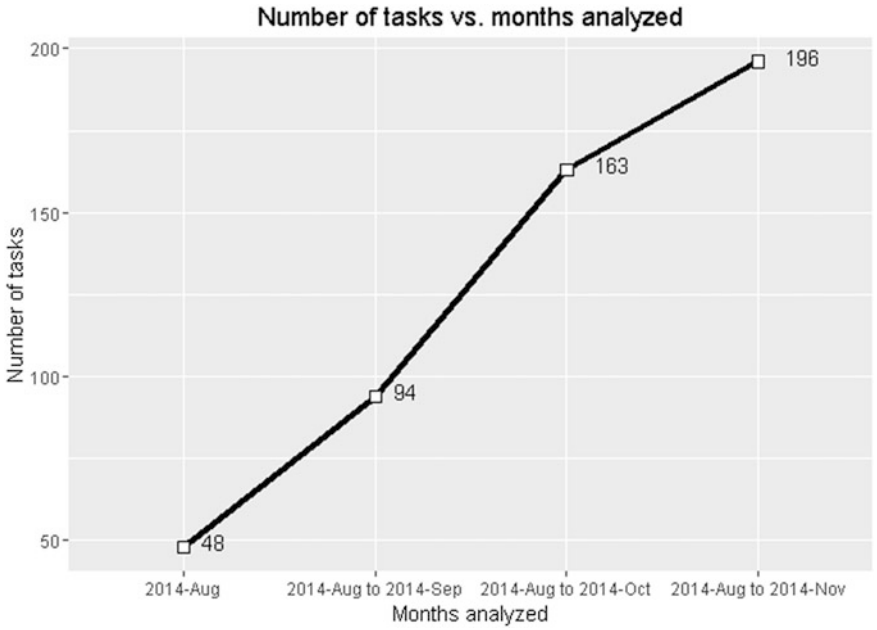


Fig. 9 Number of tasks versus months analyzed

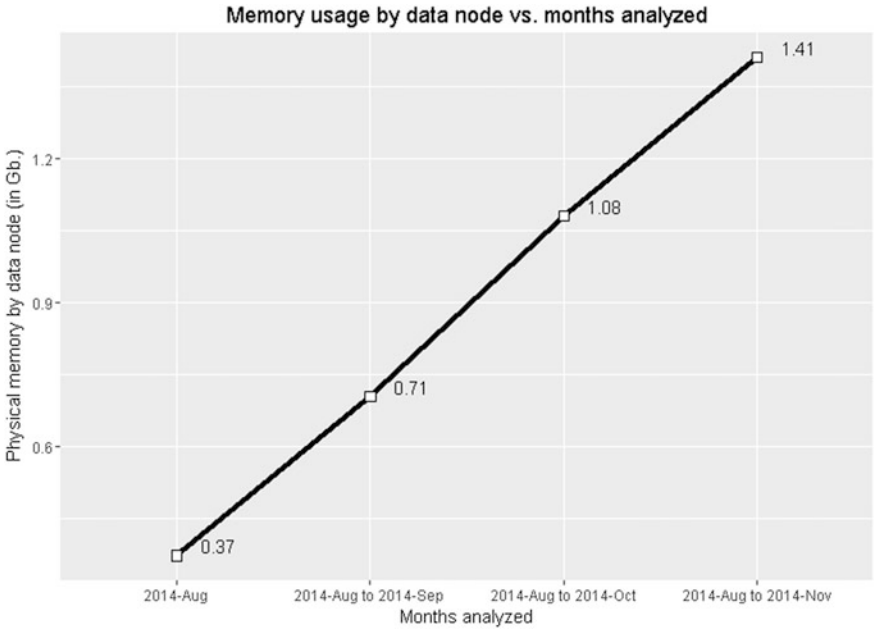


Fig. 10 Memory usage by data node versus months analyzed

periods were processed: 1 month (August), 2 months (August to September), 3 months (August to October) and, finally, 4 months (August to November).

The job counters of the tasks execution were collected and the information was gathered into several JSON files, in order to properly show it in a dashboard developed using R [20]. This dashboard will allow us to study the efficiency of our MapReduce scripts. For example, the charts shown in Figs. 9 and 10 represent the number of tasks and the memory usage by data node (simply averaging

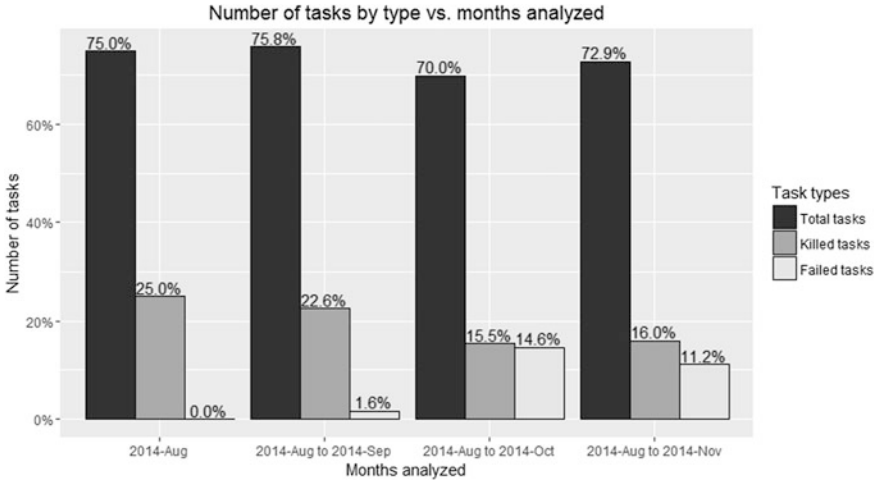


Fig. 11 Number of tasks by type versus months analyzed

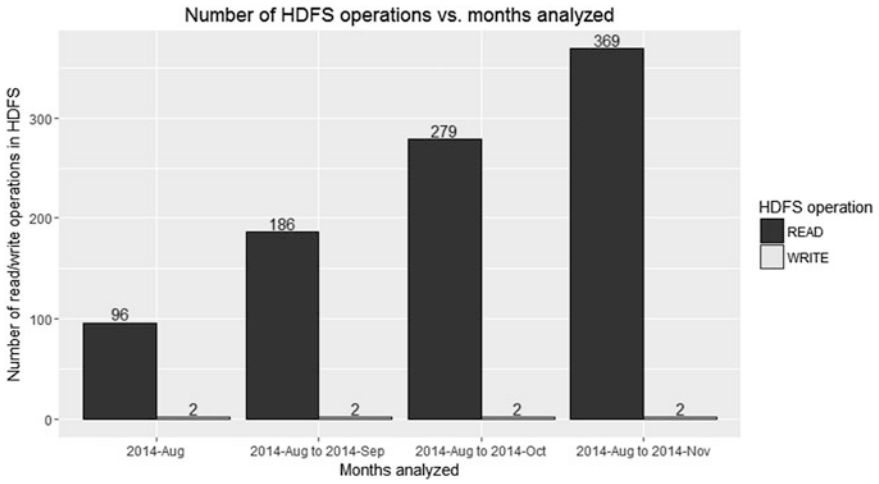


Fig. 12 Number of HDFS operations versus months analyzed

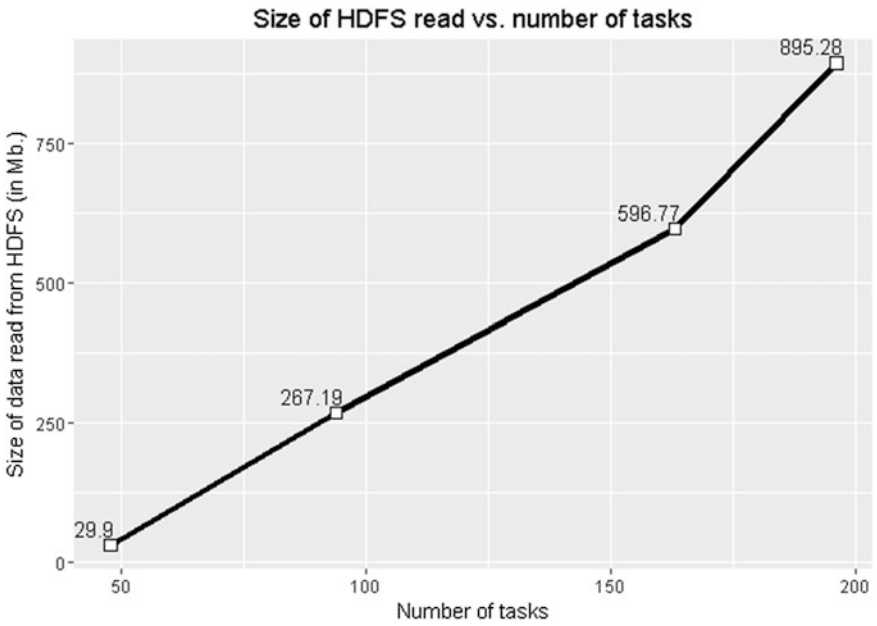


Fig. 13 Size of HDFS reads versus number of tasks

PHYSICAL_MEMORY_BYTES by the number of data nodes) for each problem instance, respectively.

The tasks (sum of TOTAL_LAUNCHED_MAPS and TOTAL_LAUNCHED_REDUCES) in each job execution can be studied taking into account the success or failure states (tasks that were killed or failed). This is shown in Fig. 11.

The number of HDFS read and write operations can also be represented (see Fig. 12) in order to study the data workflow in each problem. This issue is very important since it could reduce the number of failures.

The comparison of counters can be very useful to detect and debug other problems in job executions. The following charts represent the size of data read operation in HDFS (computed from BYTES_READ, see Fig. 13) and the memory usage by data node with respect the number of total tasks (Fig. 14).

In order to compute the CPU time, we multiply the counter CPU_MILLISECONDS by the data size read from the HDFS (that differs considerably between each problem instance). The chart in Fig. 15 illustrates the time of computation with respect to the amount of data that is processed.

The counters can be used also to define more complex indicators as a measure of memory consumed in a job execution. In this case, the memory allocated and occupied by tasks is given by:

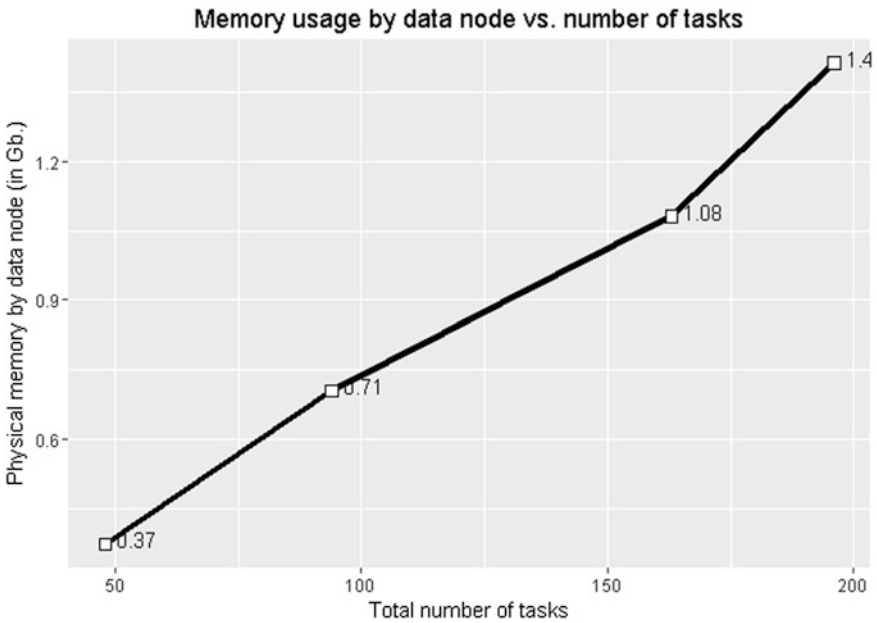


Fig. 14 Memory usage by data node versus number of tasks

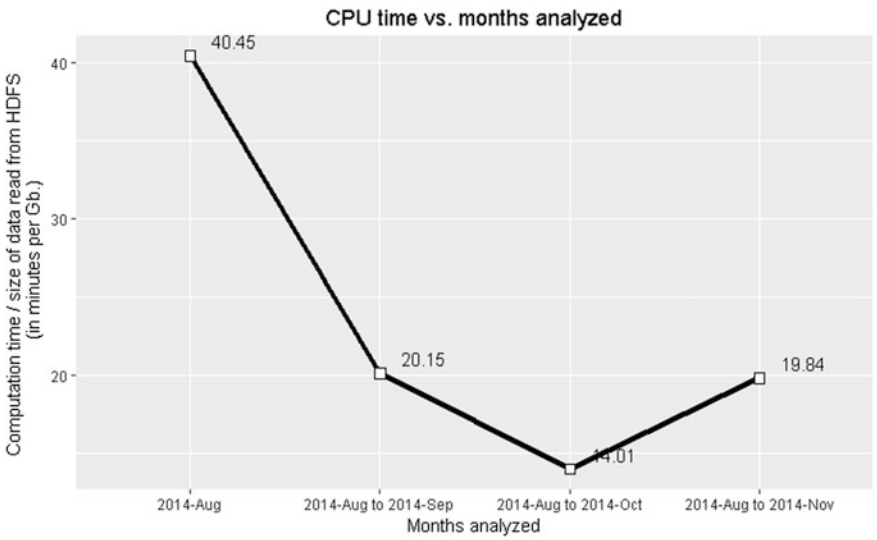


Fig. 15 CPU time versus months analyzed

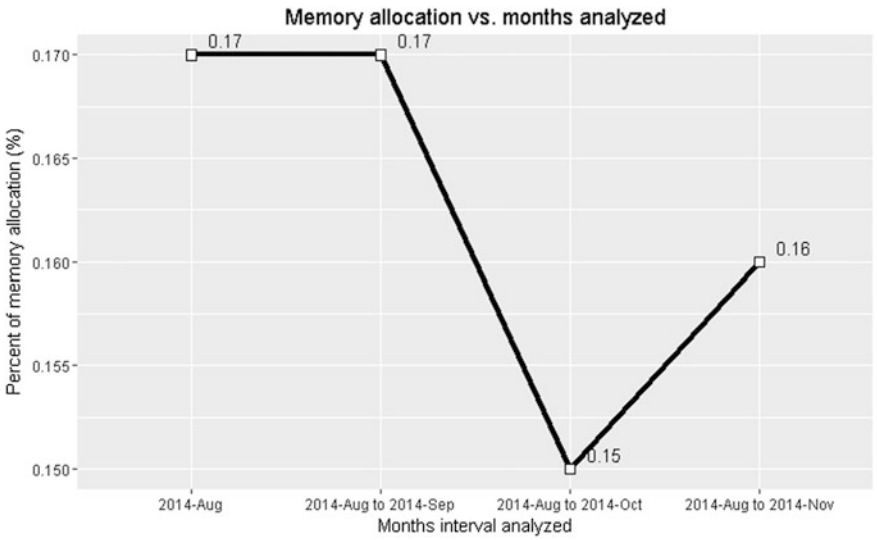


Fig. 16 Memory allocation versus months analyzed

- Allocated_Memory = (MB_MILLIS_MAPS + MB_MILLIS_REDUCES)
- Occupied_Memory = millis_tot * memory_mb/total_tasks

where:

- millis_tot = MILLI_MAPS + MILLI_REDUCES
- memory_mb = PHYSICAL_MEMORY_BYTES/(1024 * 1024)
- total_tasks = TOTAL_LAUNCHED_MAPS + TOTAL_LAUNCHED_REDUCES

Therefore, the percentage of memory allocation used can be computed as PERCENT_MEM_ALLOC = 100 * Occupied_Memory/Allocated_Memory, and the corresponding chart shows in Fig. 16 the memory usage for each problem instance.

Another indicator of interest could be the percentage of Garbage Collector (GC) time. When a Java application has excessive heap utilization, the corresponding JVM can run a full garbage collection that blocks other works and uses large amounts of CPU. The percentage of time spent in GC is computed as GC_TIME_MILLIS/(MILLIS_MAPS + MILLIS_REDUCES). The following chart in Fig. 17 represents this time for each problem instance, and can be considered as a control indicator to prevent this problem.

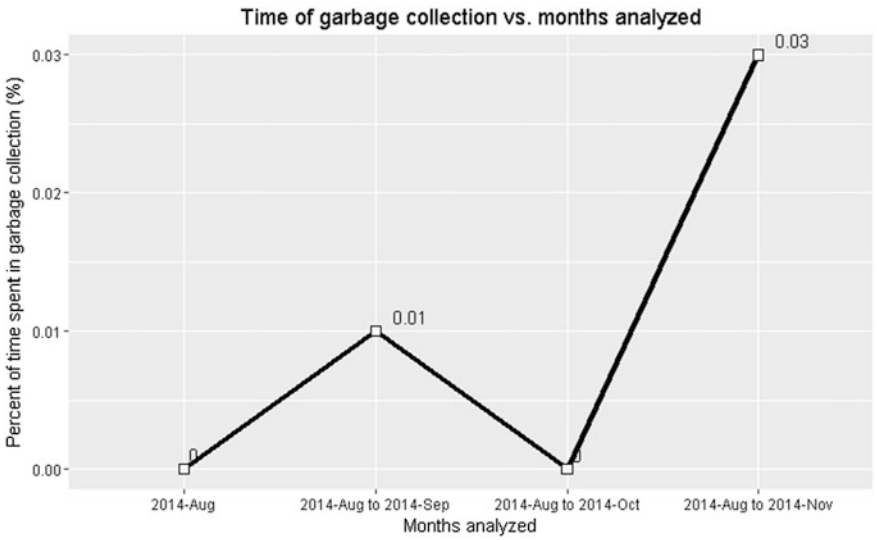


Fig. 17 Time of garbage collection versus months analyzed

5.4 Graphical Results

The summarized data obtained after processing can be represented by grouping the different servers where the logs come from, as shown in the next Fig. 18.

The information can also be completed with georeferenced data to show the spatial distribution of web accesses through the different WiFi access points in

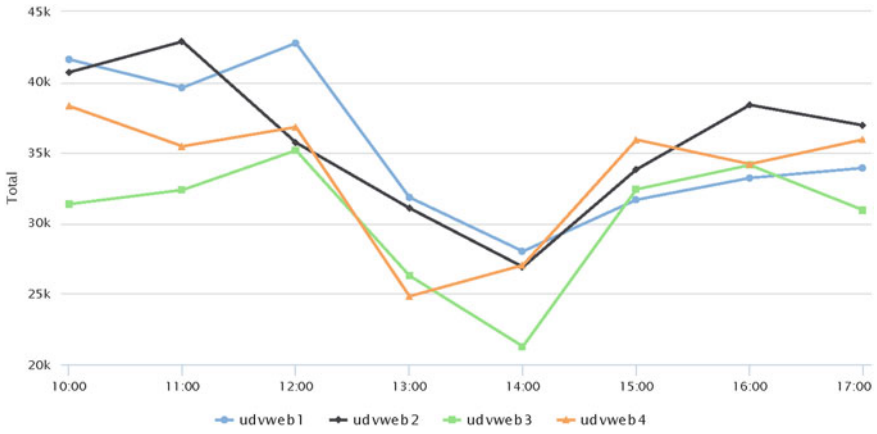


Fig. 18 Total number of accesses for different servers along a time slot

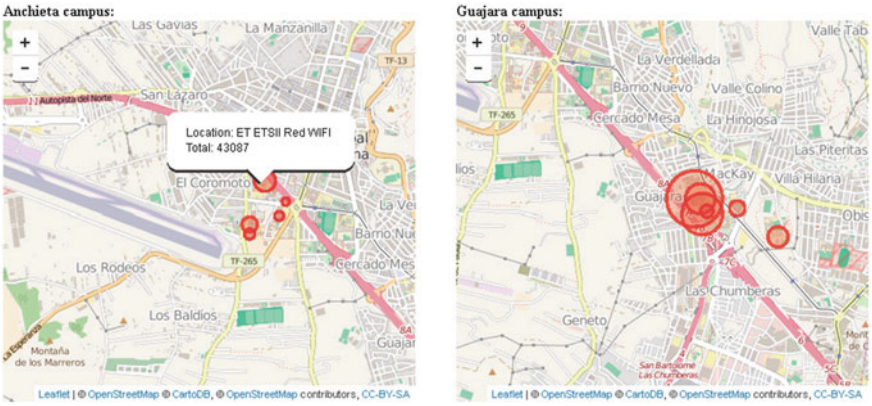


Fig. 19 Spatial distribution of the web accesses through the WiFi access points

several centers belonging to different campus of the university, as depicted in Fig. 19.

6 Conclusions

The IT Department of the Universidad de Laguna (STIC) provides service to 26 buildings with more than 1,000 network devices, and renders access to more than 10,000 user devices, which generate around 200 MB/day of log data. With such a huge infrastructure, it is highly desirable to provide new tools to explore this semi-structured data to get insights for the decision making.

In this chapter we have addressed the design and development of an application that uses Big Data techniques to analyze those log files in order to track information on user devices, as well as the number and type of network accesses for each building. Indeed, we have obtained several interesting statistical measures regarding the frequency and type of accesses.

Besides, the collaboration with STIC has tested an iterative and incremental working methodology that has been very useful to obtain quite interesting results to improve both network indicators and analysis metrics.

Accordingly, TOGAF and Archimate become necessary tools when we want to analyze, communicate and maintain complex systems. In this work we present two of the most important viewpoints used in Archimate. The Layered viewpoint gives the developer team a graphical view of the complete WiFi logs system, from the business functions to the infrastructure technology used. An intermediate layer, the Application layer, shows the software components needed to achieve the actors' goals. The second viewpoint, the Application Behavior viewpoint describes in detail the internal behavior of our MapReduce application.

The final result of processing massive log files has proved to be extremely useful to provide very valuable information in a short time. The charts shown in the last sections of this chapter enable to make a clear analysis of our cluster's performance when the different jobs are submitted. In particular, the information depicted in the different figures eases the detection of errors and the control of the success level of all the associated tasks of the different jobs.

Furthermore, we can obtain the total number of accesses for different servers along a time slot and the spatial distribution of the web accesses through the WiFi access points. This is particularly important to audit the service quality in order to define new policies for the system design and the way the users connect to the network.

All these features, along with some more improvements that could allow the analysis of log files in real time, will be studied and developed in future research.

Acknowledgments This work is partially supported by the European Commission, Agreement no. 621012, "Share PSI 2.0: Shared Standards for Open Data and Public Sector Information", ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme. By the Spanish Ministry of Education and Science, Research Project MTM2013-43396-P, National Plan of Scientific Research, Technological Development and Innovation. And by the Cabildo de Tenerife, through the Open-Big-Smart Data Project. The authors wish to thank Adrián Muñoz-Barrera, Luis A. Rubio-Rodríguez and Pedro González-Yanes for their support and assistance both in the configuration and deployment of the Hadoop cluster and in the development of the solution.

References

1. The Zettabyte Era. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html. Accessed May 2016
2. Intel (2014) What Happens in an Internet Minute? http://www.intel.es/content/www/es/es/communications/internet-minute-infographic.html. Accessed May 2016
3. Vaarandi R, Niziński P (2013) A comparative analysis of open-source log management solutions for security monitoring and network forensics. CCDCOE—NATO Cooperative Cyber Defence. http://ccdcoe.org/multimedia/comparative-analysis-open-source-log-management-solutions-security-monitoring-and-network.html. Accessed May 2016
4. What is Big Data? (in Spanish). http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data. Accessed May 2016
5. Nair R, Narayanan A (2012) Benefitting from big data: leveraging unstructured data capabilities for competitive advantage. Booz & Company. http://www.strategyand.pwc.com/media/file/Strategyand_Benefiting-from-Big-Data.pdf. Accessed May 2016
6. Bloem J, van Doorn M, Duivestein S, van Manen T, van Ommeren E (2012) Creating clarity with Big Data. SOGETI. http://blog.vint.sogeti.com/wp-content/uploads/2012/07/VINT-Sogeti-on-Big-Data-1-of-4-Creating-Clarity.pdf. Accessed May 2016
7. Laney D (2012) Deja VVVu: Others Claiming Gartner's Construct for Big Data. http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data. Accessed May 2016
8. Soubra D (2012) The 3 Vs that define BigData. Data Science Central. http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data. Accessed May 2016

9. Yiu C (2012) The big data opportunity. policy exchange. http://www.policyexchange.org.uk/images/publications/the%20big%20data%20opportunity.pdf. Accesssed May 2016
10. TechAmerica Foundation (2012) Demystifying big data: a practical guide to transforming the business of government. https://www-304.ibm.com/industries/publicsector/fileserve?contentid=239170. Accessed May 2016
11. NIST (2015) Big data interoperability framework: volume 1, Definitions. http://dx.doi.org/10.6028/NIST.SP.1500–1. Accessed May 2016
12. Davenport T, Harris J (2007) Competing on analytics. Harvard Business School Press, Boston, MA
13. SAP (2011) Making business run better with in-memory computing & predictive analytics. http://scn.sap.com/docs/DOC-5024. Accessed May 2016
14. ITU-T (2015) Big data—cloud computing based requirements and capabilities. http://handle.itu.int/11.1002/1000/12584. Accessed May 2016
15. Apache Hadoop. http://hadoop.apache.org. Accessed May 2016
16. Hu H, Wen Y, Chua TS, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687
17. Pentaho. http://www.pentaho.com. Accessed May 2016
18. The Open Group Architecture Framework (TOGAF) Version 9.1. The Open Group. http://www.opengroup.org/togaf. Accessed May 2016
19. Lankhorst MM (2004) Enterprise architecture modelling—the issue of integration. Adv Eng Inform 18(4):205–216
20. The R Project for Statistical Computing. https://www.r-project.org. Accessed May 2016
21. RStudio. https://www.rstudio.com. Accessed May 2016
22. Shiny. http://shiny.rstudio.com. Accessed May 2016
23. Laserson U (2013) A guide to python frameworks for hadoop. http://blog.cloudera.com/blog/2013/01/a-guide-to-python-frameworks-for-hadoop. Accessed May 2016
24. White T (2015) Hadoop: the definitive guide. O'Reilly media

Big Data and Earned Value Management in Airspace Industry

Juan Carlos Meléndez Rodríguez, Joaquín López Pascual, Pedro Cañamero Molina and Fausto Pedro García Márquez

Abstract Earned Value Management (EVM) is one of the most effective methods for the project management. Actual Cost and earned value are the parameters used for monitoring projects. These parameters are compared with planned value to analyze the project status. EVM covers scope, cost and time, and unifies them in a common framework that allows evaluation of project health. This chapter aims to integrate the project management and the Big Data. It is proposed an EVM approach, developed from a real case study in aerospace industry, to manage simultaneously a large number of projects.

1 Introduction

The Earned Value Management (EVM) has its origin in the concept Earned Value (EV) used by engineers of the first American factories. A primitive version of EVM was part of PERT/COST (Program Evaluation Review Technique/COST) 1962 of the project the U.S. ballistic missile Minuteman in 1962 and in 1967 became the core of the C/SCSC (Cost/Schedule Control System Criteria) with a set of 35 criteria [1]. The Undersecretary of Defense for Acquisition of the United States

J.C.M. Rodríguez (✉) · J.L. Pascual
Juan Carlos I University, Madrid, Spain
e-mail: juan.melendez@airbus.com

J.L. Pascual
e-mail: joaquinlopezpascual@gmail.com

P.C. Molina · F.P.G. Márquez
Ingenium Research Group, Castilla-La Mancha University, Ciudad Real, Spain
e-mail: pedrocanameromolina@gmail.com

F.P.G. Márquez
e-mail: FaustoPedro.Garcia@uclm.es

F.P. García Márquez and B. Lev (eds.), *Big Data Management*,
DOI 10.1007/978-3-319-45498-6_11

published the first standard EVM in the standard ANSI/EIA 748 (American National Standard Institute/Electronic Industry Association) with 32 rules [2].

The Project Management Institute (PMI) published in 2005 the PMBOK® Guide (Project Management Body of Knowledge Guide) with the Practice Standard for Earned Value Management. PMI had incorporated EVM concepts in previous PMBOK® publications [3].

The literature about EVM is wide. Ambari defined the basic principles of the method in Earned Value Project Management Method and Extensions (2003, 2004) [4]. Lipke published the "Schedule is different" where he declared that "from the time of the development of the EVM indicators, it has been known that the schedule indicators are flawed and exhibit strange behavior over the final third of the project when performance is poor" and he proposed the Earned Schedule to remedy this deficiency [5–7]. Khamooshi and Golafshani published the EDM: Earned Duration Management, a new approach to schedule performance management and measurement in 2014, as an EVM extension more effective to manage the schedule than Earned Schedule.

Lipke published the "Schedule is different" where he declared that "from the time of the development of the EVM indicators, it has been known that the schedule indicators are flawed and exhibit strange behavior over the final third of the project when performance is poor" and he proposed the Earned Schedule to remedy this deficiency. Khamooshi and Golafshani published the EDM: Earned Duration Management, a new approach to schedule performance management and measurement in 2014, as an EVM extension more effective to manage the schedule than Earned Schedule [8].

EVM had been used in the aerospace industry from its origins, where it can be summarized as follow:

- Swedish Industry Group JAS (IG JAS) use EVM for the Gripen project (Saab JAS 39 Gripen, light single-engine multirole fighter aircraft manufactured). The Swedish Government decided the implementation of the EVM system for this project in 1982 [9].
- The EVM was applied by the National Aeronautics and Space Administration (NASA) to obtain a more efficient projects management, mainly for the project with budget constraints. An example of this was the STARDUST mission project. STARDUST was the Discovery Program's fourth mission [10]. With EVM and other management tools, the project and mission managers, Lockheed Martin Astronautics (LMA) and Jet Propulsion Laboratory (JPL), managed to complete on time with nearly $2M under budget. The project managers implanted efficiently the EVM and accomplish a reduction in time of evaluation of earned-value of a month to a week, allowing them to react quickly against deviations from the plan.
- In 2004, Exploration Systems Mission Directorate of the NASA (ESMD) decided to implement a highly specific monthly EVM report in relatively small projects (budget from 1 to 10 million dollars) [11]. This implementation was not easy by the difficulty highly specialized language and the absence of adequate tool. ESMD used the NASA Program Management Tool (PMT) to implement a

full set of EVM reporting capabilities. This tool was used by project managers for project planning and reporting and the data input templates was modified to generate the EVM reporting. This new tool (PMT EVM module) reduced the time to collect the cost and schedule data to 1 or 2 days, the project managers had more time for variance analysis and creating actions.

- Department of Defense (DoD) is other US department which uses EVM as a management tool. The US Air Force F-22 fighter program is an example [12].
- This project management methodology is also used by private companies as Airbus and Boeing. The Boeing Company published in 1999 a manual Integrated Performance Management Practice, which had as scope to be used to implement EVM in all Boeing organizations [13]. Boeing also collaborated with the National Defense Industrial Association to write the industry EVMS standard [14]. Bombardier also used EVM concepts for projects management [15].

2 Big Data

In this paper it is we proposed to use Big data technology to manage AIRBUS projects in an airspace firm that it is not mention for confidential reasons. Airbus The firm organize its projects in familie programs. These families are divided into Operative Plans and organized by cost center. This way of managing projects need technology Big Data for simultaneous management.

Big Data is the technology to manage large amounts of data, which traditional technology is not prepared to analyze or manage. Jules J. Berman (author of Principles of Big Data) defined Big Data with the three V's: Volume, Variety and Velocity [16].

- Volume: This V is used to quantify the data size. The volume of Big Data measure in scale of Petabytes (PB, 1015 bytes), Exabytes (EB, 1018 bytes) or Zettabytes (ZB, 1021 bytes). The Volume is the main characteristic of Big Data.
- Velocity: The data is constantly changing, the velocity measures the rapidity of data creation. The data is received, processed and analyzed at a rate that traditional technology can't support.
- Variety: the data comes in different forms, including traditional databases, images, documents, and complex records. The data can be structured and unstructured.

An element more was added by Ohlhorst: Veracity, this characteristic make reference to need quality data. Where purity of the information is critical for its value due to the fact that the massive amounts of data collected for big data purposes can lead to statistical errors and misinterpretation of the collected information [17].

Value is other element added to this definition. It is necessary to separate important data from irrelevant data. The aim is to identify important data to eliminate unimportant and irrelevant data and to acquire insight and domain-specific interpretation [18].

3 Earned Value Management

EVM uses a cost and schedule planning as baseline and two parameters to carry out the monitoring of project: The Planned Value (PV) is the baseline, defined as the time-phased budged package; Actual Cost (AC) and Earned Value (EV), that are the monitoring parameters, being AC the actual cost spent on time and EV is the work that was accomplished [4]. The EV is given by Eq. 1.

$$\mathrm{EV} = \%\mathrm{Progress} \times \mathrm{PV\ total} \tag{1}$$

The technique of EV analysis requires evaluating variance between the parameters EV, AC and PV. The Cost Variance (CV) is utilized to identify if the project is more or less the planned value.

$$\mathrm{CV} = \mathrm{EV} - \mathrm{AC} \tag{2}$$

When the CV is negative, then the project cost is more than the budget, and if it is positive then the project cost is under budget. The Schedule Variance (SV) is the indicator to represent how advanced the project on schedule.

$$\mathrm{SV} = \mathrm{EV} - \mathrm{PV} \tag{3}$$

The SV analysis is similar to CV, a positive values means that the project is ahead to the planned schedule, and when it is negative means that a project is delayed from planned schedule. The Cost Performance Index (CPI) and the Schedule Performance Index (SPI) evaluate the project efficiency.

$$\mathrm{CPI} = \mathrm{EV}/\mathrm{AC} \tag{4}$$

$$\mathrm{SPI} = \mathrm{EV}/\mathrm{PV} \tag{5}$$

CPI and SPI indices indicate the efficiency of the project in cost and schedule.

The main acronyms and definitions of EVM are given in Practice Standard for Earned Value Management of PMI (2011) [3].

The approach does a projection of the Estimate at Completion (EAC) that may differ from the Budget at Completion (BAC) was developed. It allows analyzing PV with an estimate of cost (from AC) and schedule estimation (from EC) from the current time. This analysis results the following parameters and index:

- BAC: Budget at completion. This is the total budget baseline of project.
- EAC: Estimate at completion.

Projected cost (EAC) according the initial budget: Whether it is below or above the initial budget, the cost of the remaining work will be carried out as originally budgeted.

$$EAC = AC + (BAC - EV) \tag{6}$$

Projected cost (EAC) according to current CPI: Regardless of the efficiency or inefficiency in resource use, costs of the remaining work will maintain the same level of efficiency or inefficiency, it is expected that the project has experienced the date continue in the future.

$$EAC = AC + (BAC - EV/CPI) \tag{7}$$

Projected cost (EAC) according CPI and SPI: The corresponding to the ETC work will be done according to a ratio of efficiency that takes into account both the rate of cost performance (CPI) and the index of schedule performance (SPI), schedule delays also affect costs.

$$EAC = AC + (BAC - EV)/(CPI \times SPI) \tag{8}$$

Variations of this method measures the CPI and SPI according to different weight values, which are in the opinion of the project manager, for example, you can take 70 % of CPI and 30 % of SPI.

$$EAC = AC + (BAC - EV)/(CPI \times 0.7 + SPI \times 0.3) \tag{9}$$

- ETC: Estimate to complete.

$$ETC = EAC - AC \tag{10}$$

- VAC: Variance at completion.

$$\begin{aligned} VAC &= BAC - EAC \\ &VAC > 0 \rightarrow \text{Cost underrun} \\ &VAC < 0 \rightarrow \text{Cost overrun} \end{aligned} \tag{11}$$

$$VAC\,\% = (VAC/BAC) \times 100 \tag{12}$$

4 EVM Extensions

ES and EDM are EVM extensions more effective in terms of schedule.

Lipke proposed Earned Schedule to improve EVM to add time unit, this term is analogous to EV [5].

$$ES = t + (EV_{t+1} - PV_t)/(PV_{t+1} - PV_t) \quad (13)$$

where *t* is the period before the Actual Time (AT). Lipke recalculated SV and SPI with ES, AT and Planned Duration (PD):

$$SV = ES - AT \quad (14)$$

$$SPI = ES/AT \quad (15)$$

$$EAC(t) = AT + (PD - ES)/SPI \quad (16)$$

Khamooshi and Golafshani proposed Earned Duration Management as a method that improves the EVM with ES [8]. This method is a reformulation of EVM in time units. Table 1 shows a comparison of the 3 methods.

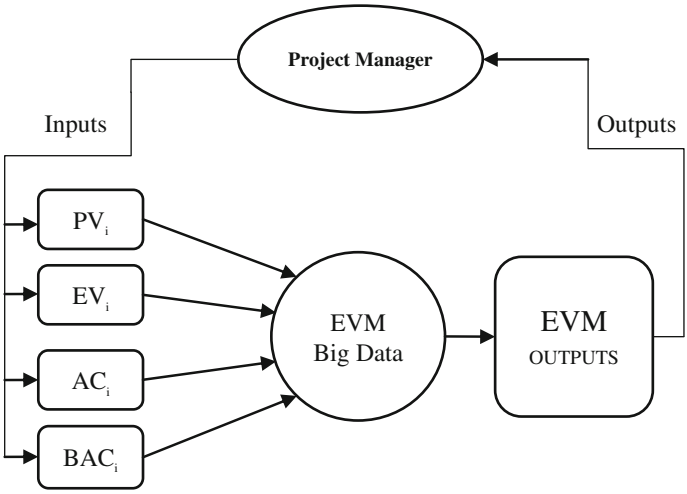
5 Big Data and Earned Value Management

The system analyzes projects integrating EVM methodology and Big Data. The system gives the project manager a tool to evaluate the project progress. The analysis is done with EVM traditional methodology.

Figure 1 shows a scheme of the information flux proposed in this chapter for projects management, where the inputs are the parameters planned and monitored (PV, EV, AC and BAC).

Table 1 EVM extensions comparison

EVM	EVM/ES	EDM	EDM equation
EV	EV	Total earned duration	TED
	ES	Earned duration	ED = t + ((TED$_{t+1}$ − TPD$_t$)/(TPD$_{t+1}$ − TPD$_t$))
PV	PV	Planned duration	PD
		Total planned duration	TPD = PD
	PD	Baseline planned duration	BPD
AC	AC	Total actual duration	TAD = ∑ AD
	AT	Actual duration	AD
SPI	SPI	Duration performance index	DPI = ED/AD
		Earned duration index	EDI = ED/PD
SV	SV	Duration variance	DV = ED − PD
EAC	EAC	Estimated duration at completion	EDAC = AD + ((max (PD, AD) − ED)/SPI)
ETC	ETC	Estimate duration to complete	EDTC = EDAC − AD



The outputs are the variables calculated with the following equations:
Cost Variance

$$\mathrm{CV(i)} = \mathrm{EV(i)} - \mathrm{AC(i)} \tag{17}$$

Cost performance Index

$$\mathrm{CPI(i)} = \mathrm{EV(i)}/\mathrm{AC(i)} \tag{18}$$

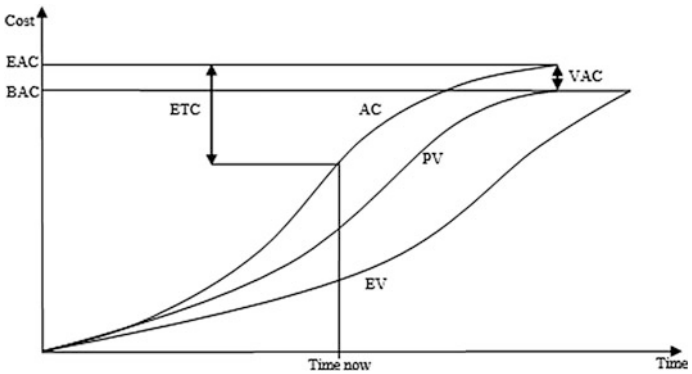


Fig. 1 Budget at completion, estimate at completion, estimate to complete and variance at completion

Schedule Variance

$$SV(i) = EV(i) - PV(i) \tag{19}$$

Schedule Performance Index

$$SPI(i) = EV(i)/PV(i) \tag{20}$$

Estimate at Complete

$$EAC(i) = AC(i) + (BAC(i) - EV(i)) \tag{21}$$

Estimate to Complete

$$ETC(i) = EAC(i) - AC(i) \tag{22}$$

Variance at Completion

$$VAC(i) = BAC(i) - EAC(i) \tag{23}$$

To develop the system has been used a case study of aerospace engineering. The system evaluates the 3600 case study projects using the EVM methodology. The projects are measured in hours as unit cost, this allows to know the times without using ES or EDM.

The system allows the project manager has an overview of the status of projects with the graphs shown in Figs. 2 and 3, to address the problems of the projects have worse outcomes. These graphs allow us to know the difference between actual cost and planned cost and between actual schedule and planned schedule in cost unit.

The Fig. 2 is the representation of the Cost Variance; this graph identifies projects with more over cost.

The Fig. 3 is the representation of the Schedule Variance; this graph identifies the projects with worst schedule to support the project manager.

The project manager also has available the variables calculated (CV, CPI, SV, SPI, EAC, ETC and VAC) for an analysis of each project more precise. The Table 2 shows the parameters of the twenty projects higher Cost Performance Index ordered by the Cost Variance.

The system is designed for visual identification of projects with larger deviations and the study of these projects from the calculated parameters.

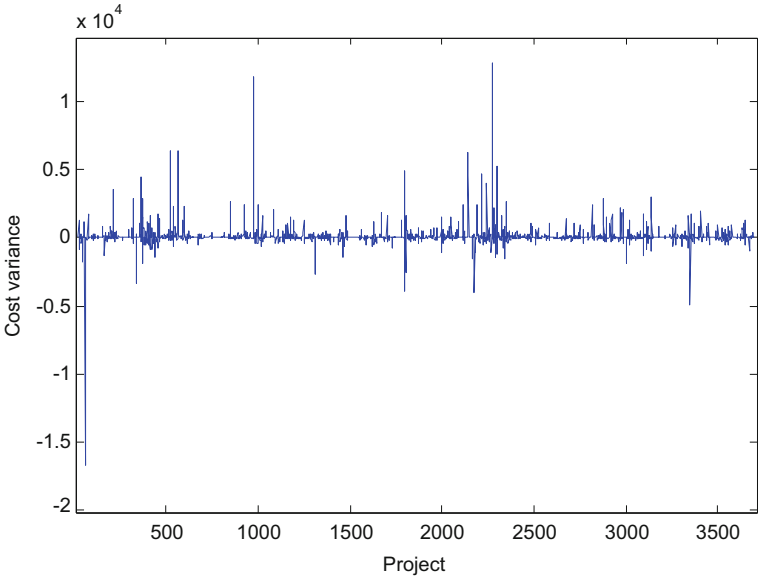


Fig. 2 Cost variance

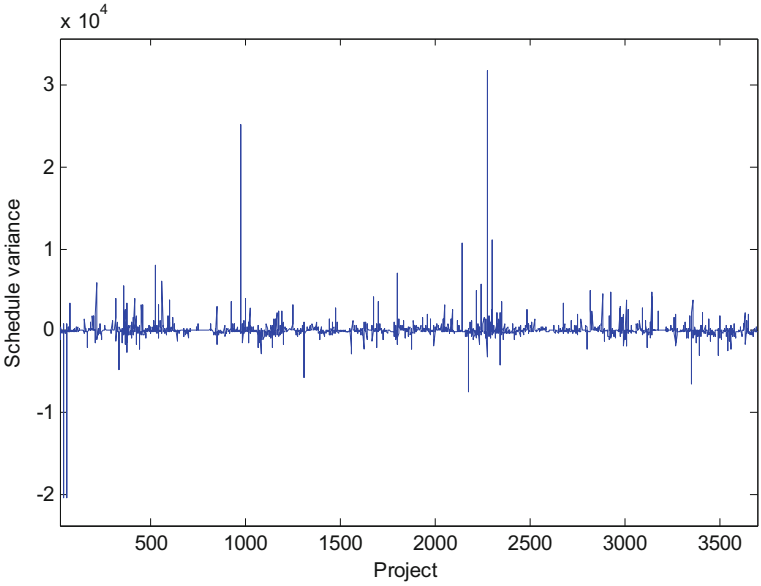


Fig. 3 Schedule variance

Table 2 Parameters EVM

Project	AC	PV	EV	BAC	CV	SV	CPI	SPI	EAC	ETC	VAC
2275	20870	1812	33696	30820	12826	31884	1.6	18.6	17994	−2876	12826
975	15079	1474	26806	22295	11727	25332	1.8	18.2	10568	−4511	11727
563	8229	8630	14643	23504	6414	6013	1.8	1.7	17090	8861	6414
523	11898	10150	18237	17860	6339	8087	1.5	1.8	11521	−377	6339
2140	6744	2159	12977	10000	6233	10818	1.9	6.0	3767	−2977	6233
2297	9328	3418	14574	15267	5246	11156	1.6	4.3	10021	693	5246
1799	8223	6099	13107	16914	4884	7008	1.6	2.1	12030	3807	4884
2214	4655	4400	9266	10158	4611	4866	2.0	2.1	5547	892	4611
2143	5753	4318	10315	12000	4562	5997	1.8	2.4	7438	1685	4562
362	6918	5811	11323	19633	4405	5512	1.6	1.9	15228	8310	4405
2243	6686	5023	10686	19341	4000	5663	1.6	2.1	15341	8655	4000
2276	6918	1619	10575	9748	3657	8956	1.5	6.5	6091	−827	3657
214	4186	1914	7755	5320	3569	5841	1.9	4.1	1751	−2435	3569
3137	3864	2591	6845	9000	2981	4254	1.8	2.6	6019	2155	2981
2879	2992	1389	5871	5260	2879	4482	2.0	4.2	2381	−611	2879
374	2943	2303	5779	9000	2836	3476	2.0	2.5	6164	3221	2836
319	3728	2540	6534	18260	2806	3994	1.8	2.6	15454	11726	2806
365	3377	4102	6147	12200	2770	2045	1.8	1.5	9430	6053	2770
3138	4678	2663	7336	8800	2658	4673	1.6	2.8	6142	1464	2658
847	3132	2802	5755	6357	2623	2953	1.8	2.1	3734	602	2623

6 Conclusions

This chapter proposes a projects management approach based on EVM. The main characteristic is the simultaneous evaluation of projects, the project manager benefits.

The case study used to develop the tool uses hours as the unit cost (engineering hours), therefore it provides an idea of what they are delayed or advanced projects and applying the rate of conversion of hours to euros the cost in monetary unit.

Matlab has been the platform used to develop the tool, which allows to edit the system to analyze other kinds of projects.

References

1. Roman DD (1962) The PERT system: an appraisal of program evaluation review technique. Acad Manag J 5(1):57–65
2. PMI E (2008) A guide to the project management body of knowledge (PMBOK Guide): an American National Standard ANSI. PMI 99-001-2008
3. 5th Edition PMBOK® Guide-Chapter 7: Earned Value Management (Part 1)
4. Anbari FT (2003) Earned value project management method and extensions. Proj Manag J 34 (4):12–23
5. Lipke W (2003) Schedule is different. Meas News 10–15

6. Lipke W (2009) Earned schedule. An extension to earned value management… for managing schedule performance. Lulu® Publishing
7. Lipke W (2013) Earned schedule-ten years after. Meas News 3:15–21
8. Khamooshi H, Golafshani H (2014) EDM: earned duration management, a new approach to schedule performance management and measurement. Int J Project Manage 32(6):1019–1041
9. Antvik S Why was earned value management important to the Swedish Government in the Gripen project?
10. Atkins L, Martin BD, Vellinga JM, Price RA (2003) STARDUST: implementing a new manage-to-budget paradigm. Acta Astronaut 52:87–97
11. Putz P, Maluf D, Bell DG, Gurram MM, Hsu J, Patel HN, Swanson KJ (2007) Earned value management at NASA: an integrated, lightweight solution. In: Aerospace conference, 2007 IEEE. IEEE, pp 1–8
12. Dibert JC, Velez JC, Dibert JC, Velez JC (2006) An analysis of earned value management implementation within the F-22 system program office's software development
13. Robinson R (2001) Earned value management. The Boeing Company Single EVMS. 13th annual international conference (2001)
14. Abba W (2000) How earned value got to primetime: a short look back and a glance ahead. In: Project management institute seminars and symposium in Houston, TX
15. Laporte CY, Doucet M, Roy D, Drolet M (2007) Improvement of software engineering performances an experience report at bombardier transportation–total transit systems signalling group
16. Berman J (2013) Principles of big data. Morgan Kaufmann Elsevier, Waltham
17. Ohlhorst FJ (2012) Big data analytics: turning big data into big money. Wiley
18. Lycett M (2013) Datafication: making sense of (big) data in a complex world