Big Data Analytics for Intelligent Internet of Things

Mohiuddin Ahmed, Salimur Choudhury, and Fadi Al-Turjman

1 Introduction

A successful of Internet of Things (IoT) environment requires standardization that contains interoperability, compatibility, reliability, and effectiveness of the operations on a global scale [1]. The rapid growth of cloud computing facility and the IoT causes a sharp growth of data. Enormous amounts of networking sensors are continuously collecting and transmitting data to be stored and processed in the cloud. Such data can be environmental data, geographical data, astronomical data, logistic data, etc. Mobile devices, transportation facilities, public facilities, and home appliances are the primary data acquisition equipment in IoT. The volume of such data will surpass the capacities of the IT architectures and infrastructure of existing enterprises and, due to real-time analysis character, will also greatly impact the computing capacity [2]. Management of these increasingly growing data is a challenge for the community in general. Figure 1 shows a year over year rise on the amount of data.

Due to the characteristics of the data being generated from IoT environment, we can call these data as Big data. The challenges faced by the IoT users compelled to label these data as Big data! Therefore the Big data generated by IoT has

M. Ahmed
College of Technology and Design, Canberra Institute of Technology, Canberra, ACT, Australia
e-mail: m.ahmed.au@ieee.org

S. Choudhury
Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada
e-mail: salimur.choudhury@lakeheadu.ca

F. Al-Turjman (✉)
Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey
e-mail: fadi.alturjman@antalya.edu.tr

F. Al-Turjman (ed.), *Artificial Intelligence in IoT*,
Transactions on Computational Science and Computational Intelligence,
https://doi.org/10.1007/978-3-030-04110-6_6

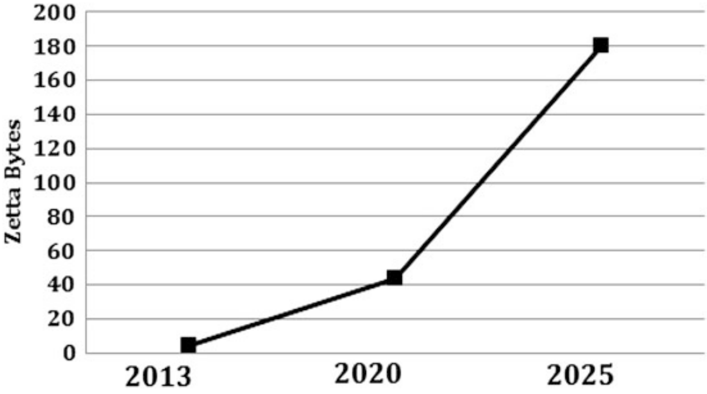

Fig. 1 Year over year rise of data volume. 1 Zetta Byte = 1 Trillion Gigabytes [3]

unique characteristics due to different types of data collected. The most common characteristics of these data reflect heterogeneity, variety, unstructured feature, noise, and high redundancy [2, 4]. It is envisaged that by 2030, the quantity of sensors will reach one trillion, and then the IoT data will be the most important part of Big data, according to the forecast of HP [5]. In fact, various artificial intelligence (AI) techniques have been applied to wireless sensors to improve their performance and achieve specific goals. We can look at AI techniques as a means of introducing an intelligent learning in a key enabling technology in IoT, which is the wireless sensor network (WSN). Learning is an important element in the observe, analyse, decide, and act (OADA) cognition loop [6, 7], used to implement the idea of cognitive wireless networks [8, 9]. We can broadly classify AI techniques as computational intelligence (CI) techniques, reinforcement learning (RL) techniques, cognitive sensor networks and multi-agent systems (MAS), and context-aware computing. Although these techniques are closely related with each other, we can segregate them to show the different goals that learning can achieve for the network.

CI techniques are a set of nature-inspired computation methodologies that help in solving complex problems that are usually difficult to fully formulate using simple mathematical models. Examples of CI techniques include genetic algorithms, neural networks, fuzzy logic, simulated annealing, artificial immune systems, swarm intelligence, and evolutionary computation. In a learning environment, CI techniques are useful when the learning agent cannot accurately sense the state of its environment. In WSNs, CI techniques have been applied to problems such as node deployment planning, task scheduling, data aggregation, energy-aware routing, and QoS management. Authors in [10] have provided an extensive survey of CI techniques applied to WSNs. They elaborate on various CI techniques and associate each with typical problem domains they can solve in WSNs. From their observations, swarm intelligence applied to solving the routing and clustering problem has drawn the most research attention in recent times. However, a major drawback of this methodology is that it can be computationally intense and may require some form of model-based offline learning to deliver to the requirements of the application scenario. Techniques such as ant colony optimization can cause an undesirable increase in communication overhead in WSNs [11] too. Apart from these drawbacks, none of the CI algorithms have been applied to solving problems of data representation, aggregation, and delivery in a distributed, decentralized setup, under dynamic communication constraints, as is the case in data hungry IoT applications.

According to Intel [1, 12], data produced from IoT has three distinguishing features:

- Terminals generating massive amount of data;
- Semi-structured or nonstructured;
- Data of IoT is not useful without analysis.

Due to the generation of Big data by IoT, the existing data processing capacity of IoT is becoming ineffective, and it is imperative to incorporate Big data technologies to promote the development of IoT. It is important to understand that the success of IoT lies upon the effective incorporation of Big data analytics. The widespread deployment of IoT also gives a challenge to Big data community to propose newer techniques as Big data and IoT are interdependent. On one hand, the widespread deployment of IoT produces data both in quantity and category, thus providing the opportunity for the application and development of Big data; on the other hand, the incorporation of Big data analytics to IoT simultaneously accelerates the research advances and business models of IoT.

Figure 2 shows a holistic view of the relationship between IoT and Big data. Earlier the amount of data generated by IoT systems could be easily handled by the traditional data analytics techniques. However, as amount of data being generated by the IoT systems are transformed into Big data, the traditional analytics methods

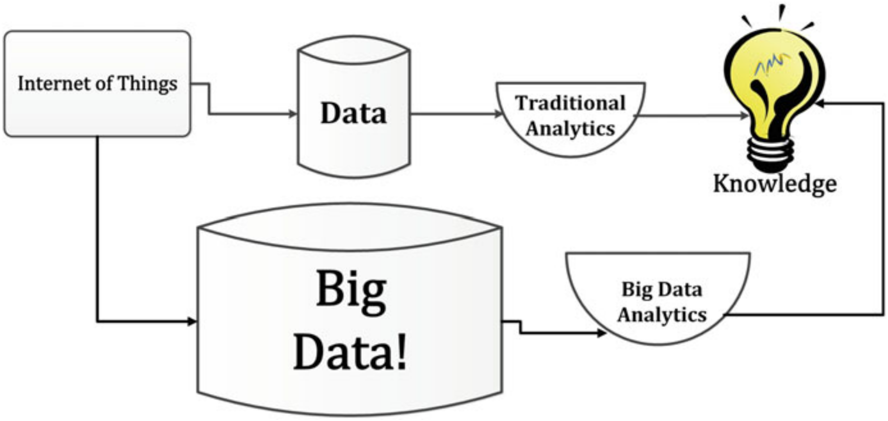

Fig. 2 Association relation between IoT and Big data

are not effective. It is important to understand the fact that, the traditional data analytics methods are not suitable to extract knowledge from Big data. For gaining meaningful insights or knowledge from the Big data generated by IoT, a set of robust data analytics method is required. Therefore, it is imperative to incorporate the analytics techniques suitable for Big data generated by the IoT systems.

In this chapter, the key aspects of Big data analytics are presented in detail. Section 2 starts with the definition of Big data; Sect. 3 reflects the challenges associated with Big data. Section 4 provides the taxonomy of Big data analytics followed by subsections on data acquisition & storage, programming model, benchmarking, analysis, and applications. Section 5 contains the future research directions of IoT data analytics followed by conclusion of the chapter.

2 Definition of Big Data

Big data is an abstract concept [13]. The concept of Big data is rudimentarily dependent on the configuration of a system, i.e. RAM, HDD capacity, etc. [14]. The significance of Big data has been recognized very recently and has different opinions on its definition. In layman's term, Big data reflects the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools in efficient manner [15]. Communities like scientific and technological enterprises, research scholars, data analysts, and technical practitioners have different definitions of Big data [16]. The following set of definitions provides an understanding on the profound social, economic, and technological connotations of Big data:

- Apache Hadoop [17]: "Datasets which could not be captured, managed, and processed by general computers within an acceptable scope".
- McKinsey & Company [18]: "Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software".
- Doug Laney [19]: "Challenges and opportunities brought about by increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety".
- IBM, Microsoft [20, 21]: "In the 3V model, Volume means, with the generation and collection of masses of data, data scale becomes increasingly Big; Velocity means the timeliness of Big data; Variety indicates the various types of data, which include semi-structured and unstructured data such as audio, video, webpage, and text, as well as traditional structured data".
- NIST [22]: "Big data means the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies, which focuses on the technological aspect of Big data".
- Manyika et al. [23]: "Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze".

- Davis and Patterson [24]: "Big data is data too Big to be handled and analyzed by traditional database protocols such as SQL".
- Edd Dumbill [25]: "Explicitly conveys the multi-dimensionality of Big Data when adding that the data is too Big, moves too fast, or doesn't fit the strictures of your database architectures".

To describe Big data, a number of V have been used in the literature. Here we combine all the Vs as below:

- Volume: The term "volume" is related with the amount of data and its dimensionality. The advantage from the ability to process large amounts of information is the main attraction of Big data analytics. The consequence is that it is a trend for many companies to store vast amount of various sorts of data: social networks data, healthcare data, financial data, biochemistry and genetic data, astronomical data, etc.
- Variety: "Variety" refers to the mix of different types of data. These data do not have a fixed structure and rarely present themselves in a perfectly ordered form and ready for processing [26]. Indeed, such data can be highly structured, semi-structured, or unstructured (video, still images, audio, clicks, etc.).
- Variability: It can be added to "variety" to emphasize on semantics, or the variability of meaning in language and communication protocols.
- Velocity: "Velocity" involves streams of data, structured records creation, and availability for access and delivery. Most of the Internet-based applications are streaming in nature and a source of Big data. The importance is reflected in the speed of the feedback loop, taking data from input through to decision.
- Value: This feature is the purpose of Big data technology. This view is well expressed by the International Data Corporation [4] when saying that Big data architectures are "designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis". This value falls into two categories: analytical use and enabling new business models, products, and services [17–25].
- Veracity: Veracity defines the truth or fact, or in short, accuracy, certainty, and precision. Uncertainty can be caused by inconsistencies, model approximations, ambiguities, deception, fraud, duplication, incompleteness, spam, and latency. Due to veracity, results derived from Big data cannot be proven, but they can be assigned a probability [4].

3 Challenges with Big Data

The Big data era brings a set of challenges on data acquisition, storage, management, analysis, and so on. Traditional data management and analysis systems are based on the relational database management system (RDBMS) [4]. However, such systems are usable with structured data only and cannot deal with the semi-structured or unstructured data which are a significant portion of Big data. In fact, Big data are not necessarily structured data and require preprocessing before analysis. It is visible that the traditional RDBMSs cannot handle the huge volume and heterogeneity of Big data. The research community has proposed some solutions from different perspectives. A promising solution is cloud computing which is utilized to meet the requirements on infrastructure for Big data, i.e. cost-efficiency and smooth upgrading/downgrading, etc. In a nutshell, it is a non-trivial task to deploy Big data analysis infrastructure. The key challenges are listed as follows.

3.1 Data Representation

The collected data from different sources are composed of certain levels of heterogeneity in the type, structure, semantics, organization, granularity, and accessibility. Therefore, it is important to properly represent Big data for further analysis. The goal of proper data representation is to make data more meaningful for analysis and user interpretation. An improper data representation will significantly impact the value of the original data and barrier to effective data analysis. An example of efficient data representation contains data structure, class, and type, as well as integrated technologies to enable efficient operations on different datasets [4, 26].

3.2 Redundancy

Usually the collected data comes with a high level of redundancy. For effective data analysis, it is important to use redundancy reduction and data compression approaches. For example, a large portion of data generated by sensor networks is highly redundant, which are required to be filtered and compressed for a robust analysis. In Big data environment, newer technologies are required to be incorporated as the redundant Big data will have a significant impact on the analysis [4, 26].

3.3 Privacy and Security

Most Big data service providers or owners outsource their datasets for effective maintenance and analysis due to their limited capacity. Therefore, usage of external bodies or tools increases the potential privacy and safety risks. For example, the transactional dataset contains details of the lowest granularity and sensitive information such as credit card numbers. Therefore, outsourced analysis of Big data is only recommended with proper preventive measures such as data anonymization of sensitive data, to ensure its security and privacy [4, 26].

3.4 Energy Efficiency

The energy consumption of high-end computing facilities is alarming due to their impact on both economic and environmental perspectives. Needless to mention that in the Big data environment, the energy consumption will be much higher than before and is unexpected from both financial and environmental perspectives. The technology industry is looking for green computing; however, the Big data is a main constraint for this venture. Therefore, it is urgent to devise new approaches to control power consumption and management mechanism without affecting the expandability and accessibility are ensured [4, 26].

3.5 Challenges with Big IoT Data

Smart cities are constructed in IoT paradigm. Therefore, Big data originates from a number of sectors such as industry, agriculture, traffic, transportation, healthcare, public departments, and so on [1–4]. According to the data acquisition and transmission approach in IoT, its network architecture may be divided into three layers: the sensing layer, the network layer, and the application layer. The sensing layer is responsible for data acquisition and mainly consists of sensor networks. The network layer is responsible for information transmission and processing, where close transmission may rely on sensor networks, and remote transmission shall depend on the Internet. Finally, the application layer support specific applications of IoT. The challenges associated with the Big IoT data are summarized as below:

- Large-scale data: Plenty of data acquisition sensors are distributed which acquire heterogeneous data.
- Strong time and space correlation: The time and space correlation is an important property of IoT data. During data analysis and processing, time and space are also important dimensions for statistical analysis.
- Effective data: Unexpected and huge amount of noises usually occur during the acquisition and transmission of data in IoT. Among datasets acquired by acquisition devices, only a small amount of abnormal data is valuable.

4 Taxonomy of Big Data Analytics

Figure 3 shows the taxonomy of Big data analytics. It consists of five basic aspects of Big data. The first category, Big data acquisition and storage, covers data acquisition and storage management. The Big data programming model includes research on programming models used in the Big data environment. The benchmark process covers the evaluation of Big data systems. The Big data analysis involves studies which focus on approaches to extract knowledge from Big data. The final category,

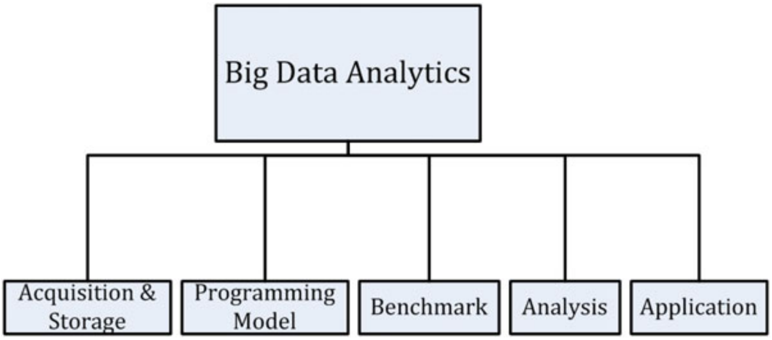

Fig. 3 Big data analytics taxonomy

application, covers studies related to the applications of Big data analytics in social, scientific, and business domains along with IoT which is the main focus of this chapter.

4.1 Acquisition and Storage

Big data comes along a new set of challenges to collect and store data efficiently. Compared to existing methods, Big data analytics need to deal with huge amounts of heterogeneous and unstructured data. It is not a trivial task to collect, integrate, and store these data by traditional data mining techniques and relevant infrastructures. These phenomena led the researchers to open a newer branch called Big data acquisition and storage. Next, the Big data acquisition and storage management is discussed.

4.1.1 Big Data Acquisition

Data acquisition is the process to aggregate information in a well-organized digital form for further storage and analysis. It is a combination of data collection, data transmission, and data preprocessing. Due to the fastest-growing sensor technology such as the Internet of Things (IoT) and radio-frequency identification (RFID), sensor-based data collection has become quite popular data acquisition approach [1–4]. However, due to the high initial investment for installation and maintenance creates a barrier in the Big data environment.

To avoid high expenditure, crowd-driven data collection was suggested by some researchers as an alternative to sensor-based data collection. The incorporation of the crowd workers in the data acquisition process helps reduction of noisy data and the collection of new types of data. There are a number of examples of crowd-driven data acquisition which are summarized below:

- FixMyStreet [27]: Users can specify the spatial location of any given street on a map and report problems associated with the location.
- Ushahidi [28]: Provides real-time data collection by enabling data collection from different channels such as email, social media, etc.
- EcoTop [29]: Reduces the noise during data collection by issuing rewards to the mobile crowd workers.

These platforms provide incentives for peer collaboration among the crowd workers to achieve data availability and quality [4]. After the data collection step, Big data is required to be transmitted to the data centre for cleaning, processing, and integration. The transmission of Big data is posed to a set of challenges such as input/output bottlenecks, network traffic delays, and data replication [4]. To overcome these challenges, researchers adopt various approaches to improve the efficiency of Big data transmission [13]. Another major challenge of Big data analytics is the integration of heterogeneous unstructured data collected from different sources. Data accessibility, common data platform, and consolidated data model were identified as three key levels of data integration [4]. Many researchers have proposed their approaches and platforms based on these levels.

4.1.2 Big Data Storage

The next step after Big data acquisition (combination of collection and transmission) is storage. The main functionality of data storage is to store and manage Big datasets with reliability and availability [4]. Infrastructure and data storage management are two basic parts of data storage [17].

- Infrastructure: Traditional infrastructure for data storage includes random access memory (RAM), magnetic disks, and storage class memory [20, 21]. Due to the specific performance of these infrastructures, it is a challenge to combine all these for Big data environment. The transmission of large amounts of data from hard disks to memory often limits the performance of Big data analytics [22]. Considering a network architecture, the data storage infrastructure can be categorized [18] as direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN). These architectures are unable to support Big data analytics. However, storage virtualization proposed by Hasan and Al-Turjman [30] offers a way to accommodate the requirements of Big data analytics. Storage virtualization is the combination of multiple network storage devices that become a single storage unit [31] and allows Big data to be easily searched and linked through a single source. Thus, data can be transferred consistently regardless of the physical infrastructure reducing the cost of storage and easier Big data analysis [32].
- Data storage management: Data storage management focuses on the file systems and database technology [33]. Google designed the Google File System (GFS) for large distributed data-intensive applications [34]. By reducing the cost of hardware, it is able to provide fault tolerance and high performance [4]. GFS

lacks efficiency for small-sized files, and some other systems such as Hadoop Distributed File System [35], Kosmos distributed file system [36], and few others were developed to fulfil the requirements of Big data storage. The variety and volume features of Big data are the important challenges to traditional relational database systems. None Structured Query Language (NoSQL) is a new type of database modelled using means other than the tabular relations [37] where the key characteristics are partition tolerance, high availability [4]. Therefore, NoSQL is a good solution for Big data storage management. Most popular NoSQL databases are SimpleDB, Cassandra, HBase, Bigtable [38], and MongoDB [39]. Other than these solutions for storage management for Big data, many researchers proposed their own solutions based on these NoSQL databases.

4.2 Programming Model

Big data processing is the next challenge after handling the storage issue. According to Pino et al. [40], there are four primary requirements involved in Big data processing as shown in Fig. 4. A number of solutions are available to fulfil these requirements. Specifically, the available programming models are designed to map applications to the parallel environment. Traditional parallel models lack the scalability and fault tolerance required by Big data [26]. These led to the development of new architectures like MapReduce [40], PreGel, GraphLab, Dryad [41], and so on. Most popular model is the MapReduce paradigm due to its robust Big data handling approaches. We briefly discuss the MapReduce below.

MapReduce is a Big data programming model that uses a wide variety of clusters to achieve automatic parallel processing and distribution. The computing

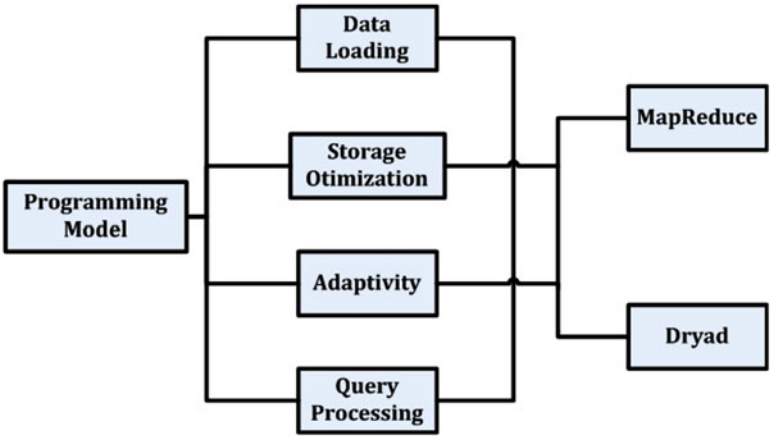

Fig. 4 Big data programming models

model contains a map function and a reduce function. The job of map function is to partition large computational tasks into smaller tasks and assign them to the appropriate key pair. After obtaining the output of the map function, the reduce function merges all values which share the same key value and generates a set of merged output values. The basic idea of MapReduce is to split a Big task into several chunks and execute the chunks in parallel to optimize. In the MapReduce model, the user only needs to focus on these above-mentioned two functions (map and reduce). As a popular and powerful programming model, MapReduce has been widely deployed to solve Big data-related problems such as in distributed computation, online aggregation, database system optimization, high-performance computing, DNA sequencing, text analytics, and many more.

Dryad is another framework for parallel applications [41] similar to MapReduce. In this model, a task is represented as a directed acyclic graph (DAG) which includes vertices and channels. Dryad completes the task by executing the vertices of the graph on a set of high-end computers and communicating through data channels. As a variant of MapReduce, Dryad generalizes MapReduce to arbitrary DAGs. This feature makes Dryad more flexible for the Big data applications with different structures [4].

One negative aspect of the MapReduce and Dryad models is that some agendas like behaviour abstraction, application optimization, and system simulation and migration are not well approached. These lead to a need for a generalized model that can bridge applications and various software frameworks for Big data analytics.

4.3 Benchmark

Advent of Big data in the scientific arena brings forth a newer set of benchmarking techniques among the researchers. The benchmarking techniques developed so far can be classified into two groups (component benchmarks and system benchmarks), and a simple taxonomy is shown in Fig. 5. The component benchmark has a limited scope and evaluates the performance of components in a Big data environment [4]. On the other hand, system benchmark focuses on the performance evaluation of an entire system [4].

Different types of benchmarking systems are discussed and summarized briefly below:

- PigMix [42], GridMix [43], GraySort [44]: The Standard Performance Evaluation Corporation (SPEC)'s central processing unit (CPU) benchmark [4]
- TeraSort [45] TeraSort (benchmark for Apache Hadoop system [17]) measures the amount of time to sort a large amount of distributed data in a given system. It has three parts as generation, sorting, and validation. The generation part creates random data. The sorting part performs the sorting and writes sorted data to Hadoop's distributed file system. The validation function reads sorted data to verify if it is in order.

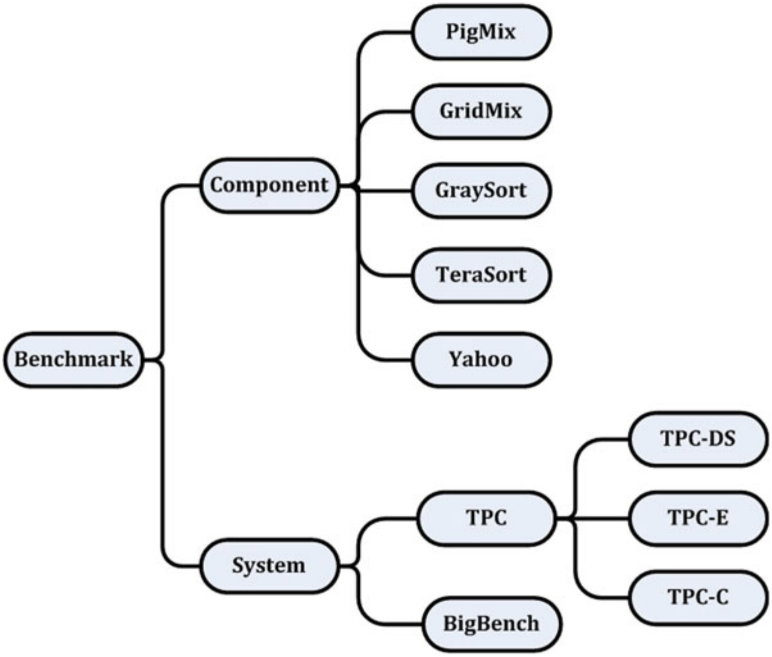

Fig. 5 Big data benchmarking

- Yahoo! [46] Yahoo! Cloud Serving Benchmark (YCSB) tries to bridge the gap of the existing benchmarking systems which were only designed for databases such as SQL. YCSB contains an extensible workload generator and several core workloads. The open-source YCSB workload generator can be used to load and execute various datasets and workloads. System Transaction Processing Performance Council [47] TPC-C, TPC-E, TPC-H, and TPC-DS are the perfect tools to measure a Big data system's ability for database loading and query executing.
- BigBench [48] It has two main components: (i) a data generator and (ii) query workload. The data generator helps to provide upgradeable volumes of structured, semi-structured, and unstructured data. The workload is devised with a set of queries that can cross different dimensions based on data.

4.4 Big Data Analytics

Data analytics is an important aspect of a successful IoT and in the dawn of Big data era; its importance has been being observed as never before. Organizations are interested in information from massive data to bring values. In this chapter, taxonomy is shown as in Fig. 6 based on the type of analysis being practised.

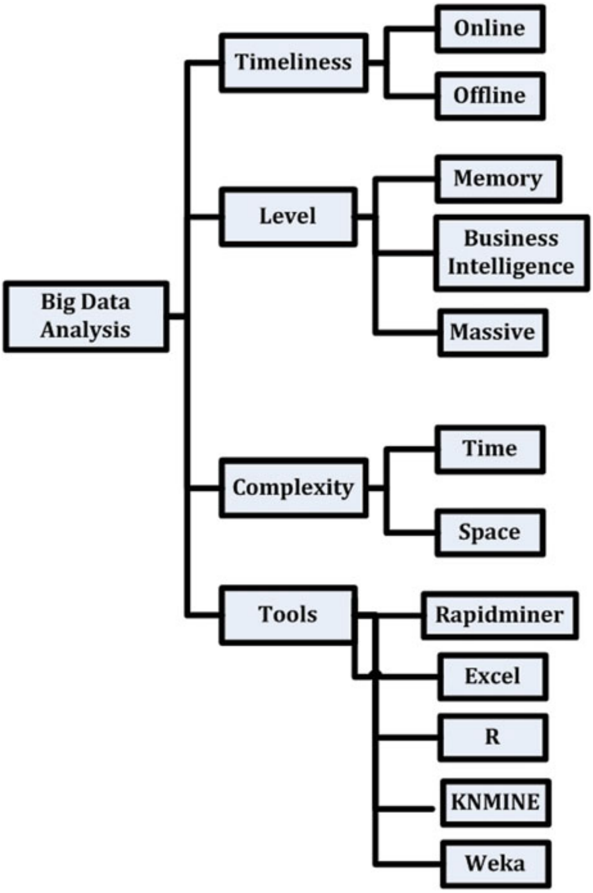

Fig. 6 Types of analytics

Because of the characteristics of Big data discussed in Sect. 3, different analytical architectures shall be considered for different application requirements.

4.4.1 Timeliness of Analysis

According to timeliness requirements, Big data analysis can be classified into online and offline analysis. These are discussed as follows:

- Online Analysis: From the IoT perspective, the online analysis is extremely important as the deployed sensors are constantly collecting data. Therefore from a particular time frame, it is necessary to analyse the data collected to reach out to any decision on anomalous event detection, cybersecurity assurance, etc. Also, in other applications such as e-commerce, financial trading data constantly changes, and rapid data analysis is needed. The widely accepted architectures for

online analysis include (i) parallel processing clusters using traditional relational databases and (ii) memory-based computing platforms. For example, Greenplum from EMC and HANA from SAP are both real-time analysis architectures [4, 16].
- Offline Analysis: Offline analysis is usually required for applications without high requirements on response time, e.g. machine learning, statistical analysis, and recommendation algorithms [4, 16, 26]. Offline analysis carried out by retrieving data into a special platform through Big data acquisition tools. In the Big data environment, it is important to have specialized platforms to reduce the cost of data processing and improve the efficiency of data acquisition. Such platforms include the open-source tool Scribe from Facebook [4, 16], LinkedIn's open-source tool Kafka [4], Hadoop [17], and so on. These tools are capable of meeting the demands of offline data analysis with hundreds of MB per second.

4.4.2 Analysis at Different Levels

Big data analysis can also be classified into memory-level analysis, business intelligence (BI) level analysis, and massive level analysis, which are briefly discussed below:

- Memory-level analysis: When the data volume is smaller than the maximum memory of a cluster, this type of analysis is required. In recent times, the memory of server cluster surpasses hundreds of gigabytes. As a result, an internal database technology is advisable to use to improve the analytical efficiency. For online analysis, memory-level analysis is extremely suitable. A representative memory-level analytical architecture is MongoDB [39]. In the age of SSD (solid-state drive), the capacity and performance of memory-level data analysis have been further improved and widely applied.
- Business intelligence (BI) analysis: This analysis is required when the data scale surpasses the memory level; however, it can be imported into the BI analysis environment. The currently mainstream BI products are provided with data analysis plans to support the level over TB [4, 16].
- Massive analysis: When the BI analysis and traditional analysis are overwhelmed by the Big data, it is required to introduce the technologies like Hadoop and MapReduce to store and analyse the data. In recent times, most massive analysis utilizes HDFS of Hadoop to store data and use MapReduce for data analysis. Most massive analysis belongs to the offline analysis category.

4.4.3 Analysis with Different Complexity

The time and space complexity of Big data analysis algorithms varies from each other due to Big data characteristics (variety) and also application demands. For example, for the applications that require parallel processing, a distributed algorithm may be designed, and a parallel processing model may be used for data analysis [16, 26].

4.5 Tools for Big Data Mining and Analysis

A wide range of tools for Big data analytics are available which includes professional, expensive commercial software and also open-source software [4]. This section covers top five Big data analytics tool widely used in the community according to a survey of "What Analytics, Data mining, Big Data software that you used in the past 12 months for a real project?" of 798 professionals made by KDnuggets in 2012 [49].

4.5.1 R

R is a language and environment for statistical computing and graphics [50]. It is a predecessor of the S language and environment which was developed at Bell Laboratories. R can be considered as a different implementation of S. R provides a wide variety of statistical techniques including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc. and is highly extensible. Furthermore, in a survey of "Design languages you have used for data mining/analysis in the past year" in 2012, R was also in the first place, defeating SQL and Java. Due to the popularity of R, database manufacturers, such as Teradata and Oracle, have released products supporting R. Programming with Big data in R (pbdR) is a series of R packages and an environment for statistical computing with Big data by using high-performance statistical computation [4, 16]. The pbdR uses the same programming language as R with S3/S4 classes and methods which is used among statisticians and data miners for developing statistical software.

4.5.2 Excel

MS Excel is the most commonly used and powerful data processing and statistical analysis tool. Excel becomes a resourceful Big data analytics tool when some advanced plug-ins, such as Analysis ToolPak and Solver Add-in, with powerful functions for data analysis are integrated. This plug-ins can be used only if users enable them. Excel is also the only commercial software among the top five [49].

4.5.3 RapidMiner

RapidMiner [51] is an open-source software used for data mining, machine learning, and predictive analysis. However, to explore the Big data analytics, it is required to have licensed version which is not free. In an investigation of KDnuggets in 2011, it was more frequently used than R (ranked Top 1). Data mining and machine learning programs provided by RapidMiner include extract, transform, and load (ETL), data preprocessing and visualization, modelling, evaluation, and deployment. The data mining flow is described in XML and displayed through a graphic user interface (GUI). RapidMiner is written in Java. It integrates the learner and evaluation method of Weka and works with R. A handsome amount of regularly used data mining and machine learning algorithms can be implemented with connection of processes including various operators.

4.5.4 KNMINE

KNIME (Konstanz Information Miner) is a user-friendly and open-source data integration, processing, and analysis platform [52]. It provides the users to create data flows in a visualized manner and to selectively run analytical processes along with the analytical results, models, and interactive views. KNIME was written in Java and contains a large number of plug-ins. Through plug-in files, users can insert processing modules for files, pictures, and time series and integrate them into various open-source projects, e.g. R and Weka. In addition, it is easy to expand KNIME. Developers can effortlessly expand various nodes and views of KNIME.

4.5.5 Weka

Weka, developed by University of Waikato researchers, which comes from Waikato Environment for Knowledge Analysis, is an open-source machine learning and data mining software written in Java [53]. Weka provides almost all the fundamental functions such as data processing, feature selection, classification, regression, clustering, association rule, visualization, etc.

4.6 Applications of Big Data

The incorporation of Big data analytics makes it easier for organizations to gain meaningful information for being successful in their venture. Big data from various sources such as websites, emails, mobile devices, and social media are all important for knowledge extraction. In recent times, Big data has attracted a lot of attention from academia and industry which is easily understood by the amount of research paper published and job advertisements for data analyst/scientists. In [26], it is mentioned that the most common applications of Big data are in business, social, and scientific applications. A brief discussion on these areas along with IoT is given below.

4.6.1 Business Applications

There are huge amounts of operational and financial data stored at millions of different data sources in the business organizations. Big data analytics provides more agility to firms and makes it easier for firms to collect and analyse operational and financial data. Valuable information from Big data analytics allows managers to make more successful decision and identification of market conditions. Big data analytics have been already incorporated in various areas such as customer behaviour analysis, purchasing patterns, supply chain management, market forecasting, risk management, and fraud detection. In summary, Big data analytics empowers the business organizations to create new products and business processes, expand customer intelligence, and increase revenue.

4.6.2 Social Applications

Big data analytics facilitate information sharing in society, detect correlations among social events, and aggregate and analyse information to assist decision-making [26]. Out of many applications of Big data analytics for social good, a timely example of Big data analytics has been observed in election campaigns. With the help of Big data analytics, political data analytics advanced from simple micro targeting to true predictive data science, and the track record is good according to Jon Markman who is an investment adviser, trader, columnist, and author. In education sector, a number of research projects are ongoing to develop new algorithms for student behaviour analysis and effective mode of learning and teaching. Hidden patterns and trends identified using Big data analytics techniques can provide educators with valuable insights for the evaluation of the learning process. Additionally, human behaviours can be analysed through gathered Big data from a variety of sources. The modelling of human behaviours will allow governing body to interpret and predict social events such as traffic distribution, civil unrest incidents, and disease outbreaks. In the healthcare sector, Big data analytics provides the intelligence for electronic health records (EHRs) by connecting operational and clinical analytic systems and supports evidence-based healthcare. Evidence-based healthcare encompasses the systematic reviewing of previous clinical data in order to provide decision-makers with information as well as predictive analytics.

4.6.3 Scientific Applications

The continuous development of Big data analytics supports scientists to access large amounts of data quickly, facilitate data collection and sharing, and discover hidden patterns in Big datasets. Currently, Big data analytics has been applied in a lot of research areas. Especially in astronomy disciplines, Big data analytics has proven to be an efficient tool to address multiwavelength, multi-messenger, and huge amounts of astronomical data. NASA uses Big data analytics for real-time data processing on the flight operations. There are thousands of scientific applications where Big data analytics is being used and new inventions are in place.

4.6.4 Application of Intelligent IoT-Based Big Data

IoT is undoubtedly an important source of Big data and, simultaneously, one of the major market shares of Big data applications. As sensors are being used across almost every industry, the IoT is going to trigger a massive influx of Big data. IoT is going to have the biggest impact in the future of Big data analytics. A simple example may be given by logistic enterprises that may have profoundly experienced with the application of IoT Big data. In this scenario, the delivery vehicles of DHL may be equipped with sensors, wireless adapters, and GPS, so the headquarter is able to pinpoint the location. In addition to that application, supervision and management of employees can be executed with optimized performances.

5 Conclusion and Research Directions

In this chapter, the relationship between IoT and Big data is explored as the technological advances are inevitable. The Big data generated from IoT requires proper management and analysis to make IoT successful. Therefore, the importance of Big data analytics in IoT is a challenge. This chapter covers the Big data terminologies in the light of IoT and discusses the taxonomy of Big data analytics. Finally the chapter finishes with a set of research directions for the collaborative Big IoT data analytics. This chapter is going to be useful resource for anyone who is interested in IoT analytics and can be used as reference for graduate research students. In the next few subsections, the research direction of Big IoT data analytics is provided.

The Big data analytics research is in its early stage as the era just started few years ago and is confronting many challenges in different areas. Significant research efforts are required to improve the efficiency of Big data analytics. It is indeed an interesting research area with great potential, and there are many important problems to be solved by the collaboration of both academia and industry. There is a universal definition required for Big data. As observed in Sect. 2 of this chapter, the researchers/technology organizations are yet to reach to an agreed definition of Big data. An accepted formal definition is urgent for a variety of application domains to correlate Big data.

The presence of Big data challenges the traditional data management approaches. Currently, a plethora of research contributions on Big data technologies including data acquisition, storage, programming, benchmarking, and analytics are made. However, in the way the Big data is growing forth, it is imperative to continue research and development of relevant technologies.

As highlighted in the IoT perspective, the value gained from Big data is far higher than the value of non-Big dataset! As a result, the integration of different data sources (as in Big data) is a prerequisite for a successful venture nowadays. Moreover, the integration is posed too many challenges, such as different data patterns, redundant data, etc.

In Big data environment, the traditional security and privacy providers are insufficient. Since the data volume is fast growing, safety risks are more than ever before, and the existing security measures are not suitable for Big data. The Big data privacy is concerned with the protection of data acquisition patterns, i.e. personal interests, properties, etc. of users, and the privacy of data which may be leaked during storage, transmission, and usage, even if acquired with the permission of users. In a nutshell, as the Big data emerges, it is vital to ensure its security and privacy. A lack of data security and privacy can cause detrimental effects such as great financial losses and reputational damage for any organization.

References

1. Giusto, D., Iera, A., Morabito, G., & Atzori, L. (2010). *The internet of things: 20th Tyrrhenian workshop on digital communications*. New York: Springer Science & Business Media.
2. Li, S., Da Xu, L., & Zhao, S. (2015). The internet of things: A survey. *Information Systems Frontiers, 17*(2), 243–259.
3. Big Data: 20 Mind-Boggling Facts Everyone Must Read. (2015). https://www.forbes.com. [Online; accessed 29-August-2017].
4. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171–209.
5. HP: Big Data Platform. (2017). http://www8.hp.com/us/en/software-solutions/Big-data-platform-haven/index.html. [Online; accessed 29-August-2017].
6. Haykin, S., et al. (2005). Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications, 23*(2), 201–220.
7. Mitola, J., & Maguire, G. Q. (1999). Cognitive radio: Making software radios more personal. *IEEE Personal Communications, 6*(4), 13–18.
8. Zaki Hasan, M., & Al-Turjman, F. (2018). Swarm-based data delivery in social internet of things. In F. Al-Turjman (Ed.), *Smart things and femtocells* (pp. 179–218). Boca Raton: CRC Press.
9. Friend, D. H., Thomas, R. W., MacKenzie, A. B., & Silva, L. A. (2007). Distributed learning and reasoning in cognitive networks: Methods and design decisions. In Q. H. Mahmoud (Ed.), *Cognitive networks: Towards self-aware networks* (pp. 223–246). Hoboken: Wiley.
10. Al-Turjman, F. (2018). Fog-based caching in software-defined information-centric networks. *Computers & Electrical Engineering, 69*(1), 54–67.

11. Al-Turjman, F. (2017). Information-centric sensor networks for cognitive IoT: An overview. *Annals of Telecommunications, 72*(1), 3–18.
12. Alabady, S., & Al-Turjman, F. (2018). Low complexity parity check code for futuristic wireless networks applications. *IEEE Access, 6*(1), 18398–18407.
13. Liu, X., Iftikhar, N., & Xie, X. (2014). Survey of real-time processing systems for big data. In *Proceedings of the 18th international database engineering &#38; applications symposium*, IDEAS'14 (pp. 356–361). New York: ACM.
14. Reed, D. A., & Dongarra, J. (2015). Exascale computing and big data. *Communications of the ACM, 58*(7), 56–68.
15. Fang, H., Zhang, Z., Wang, C. J., Daneshmand, M., Wang, C., & Wang, H. (2015). A survey of big data research. *IEEE Network, 29*(5), 6–9.
16. Chong, D., & Shi, H. (2015). Big data analytics: A literature review. *Journal of Management Analytics, 2*(3), 175–201.
17. Apache Hadoop. (2017). http://hadoop.apache.org/. [Online; Accessed 29-Aug-2017].
18. McKinsey & Company. (2017). http://www.mckinsey.com/. [Online; Accessed 29-Aug-2017].
19. Doug Laney. (2017). https://www.gartner.com/analyst/40872/Douglas-Laney. [Online; Accessed 29-Aug-2017].
20. What is Big Data. (2017). https://www.ibm.com/Big-data/us/en/. [Online; Accessed 29-Aug-2017].
21. Understanding Microsoft Big data solutions. (2017). https://msdn.microsoft.com/en-us/library/dn749804.aspx. [Online; Accessed 29-Aug-2017].
22. Big Data Information. (2017). https://www.nist.gov/el/cyber-physical-systems/Big-data-pwg. [Online; Accessed 29-Aug-2017].
23. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, June 2011.
24. Big ethics for Big data. (2017). https://www.oreilly.com/ideas/ethics-Big-data-business-decisions. [Online; Accessed 29-Aug-2017].
25. Planning for Big Data. (2017). http://www.oreilly.com/data/free/planning-for-Big-data.csp. [Online; Accessed 29-Aug-2017].
26. Ahmed, M., Anwar, A., Mahmood, A. N., Shah, Z., & Maher, M. J. (2015). An investigation of performance analysis of anomaly detection techniques for big data in scada systems. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 15*(3), 5.
27. FixMyStreet. (2017). https://www.fixmystreet.com/. [Online; Accessed 29-Aug-2017].
28. Ushahidi. (2017). https://www.ushahidi.com/. [Online; Accessed 29-Aug-2017].
29. Padhariya, N., Mondal, A., Goyal, V., Shankar, R., Madria, S. K. (2011). *EcoTop: An economic model for dynamic processing of top-k queries in mobile-P2P networks* (pp. 251–265). Berlin/Heidelberg: Springer.
30. Hasan, M. Z., & Al-Turjman, F. (2018). Analysis of cross-layer design of quality-of-service forward geographic wireless sensor network routing strategies in green internet of things. *IEEE Access, 6*(1), 20371–20389.
31. U.S. Patent No. 6,948,044. (2017). https://www.uspto.gov/. [Online; Accessed 29-Aug-2017].
32. Huber, N., Becker, S., Rathfelder, C., Schweflinghaus, J., & Reussner, R. H. (2010). Performance modeling in industry: A case study on storage virtualization. In *Proceedings of the 32Nd ACM/IEEE international conference on software engineering – volume 2*, ICSE'10 (pp. 1–10). New York: ACM.
33. Chen, X., Wang, S., Dong, Y., & Wang, X. (2016). *Big data storage architecture design in cloud computing* (pp. 7–14). Singapore: Springer.
34. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access, 2*, 652–687.
35. The Hadoop Distributed File System. (2017). http://www.aosabook.org/en/hdfs.html. [Online; Accessed 29-Aug-2017].
36. Kosmos distributed file system (KFS). (2017). http://kosmosfs.sourceforge.net/. [Online; Accessed 29-Aug-2017].

37. NoSQL. (2017). http://nosql-database.org/. [Online; Accessed 29-Aug-2017].
38. BigTable. (2017). https://cloud.google.com/Bigtable/. [Online; Accessed 29-Aug-2017].
39. MongoDB. (2017). https://www.mongodb.com/. [Online; Accessed 29-Aug-2017].
40. Pino, T., Choudhury, S., & Al-Turjman, F. (2018). Dominating set algorithms for wireless sensor networks survivability. *IEEE Access, 6*(1), 17527–17532.
41. Dryad. (2017). https://www.microsoft.com/en-us/research/project/dryad/. [Online; Accessed 29-Aug-2017].
42. Zhang, Z., Cherkasova, L., Verma, A., & Loo, B. T. (2012). Automated profiling and resource management of pig programs for meeting service level objectives. In *Proceedings of the 9th international conference on autonomic computing*, ICAC'12 (pp. 53–62), New York. ACM.
43. Sandholm, T., & Lai, K. (2009). Mapreduce optimization using regulated dynamic prioritization. In *Proceedings of the eleventh international joint conference on measurement and modeling of computer systems*, SIGMETRICS'09 (pp. 299–310), New York. ACM.
44. Graysort benchmark. (2017). http://sortbenchmark.org. [Online; Accessed 29-Aug-2017].
45. Terabyte sort on Apache Hadoop. (2017). http://sortbenchmark.org/Yahoo-Hadoop.pdf. [Online; Accessed 29-Aug-2017].
46. Baru, C., Bhandarkar, M., Nambiar, R., Poess, M., & Rabl, T. (2013). *Setting the direction for big data benchmark standards* (pp. 197–208). Berlin/Heidelberg: Springer.
47. Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., & Sears, R. (2010). Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on cloud computing*, SoCC'10 (pp. 143–154), New York. ACM.
48. Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., & Jacobsen, H.-A. (2013). Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, SIGMOD'13 (pp. 1197–1208), New York. ACM.
49. Big Data Software. (2017). http://www.kdnuggets.com. [Online; Accessed 29-Aug-2017].
50. The R Project for Statistical Computing. (2017). https://www.r-project.org/. [Online; Accessed 29-Aug-2017].
51. RapidMiner. (2017). https://RapidMiner.com/. [Online; Accessed 29-Aug-2017].
52. KNMINE. (2017). https://www.knime.org/. [Online; Accessed 29-Aug-2017].
53. WEKA. (2017). http://www.cs.waikato.ac.nz/ml/weka/. [Online; Accessed 29-Aug-2017].