Chapter 3
Big Data Challenges for the Internet of Things (IoT) Paradigm

Pornpit Wongthongtham, Jaswinder Kaur, Vidyasagar Potdar, and Abhishek Das

Abstract Millions of devices equipped with sensors are connected together to communicate with each other in order to collect and exchange data. The phenomenon of daily life objects that are interconnected through a worldwide network is known as the Internet of Things (IoT) or Internet of Objects. These sensors from a large number of devices or objects simultaneously and continuingly generate a huge amount of data, often referred to as Big Data. Handling this vast volume, and different varieties, of data imposes significant challenges when time, resources, and processing capabilities are constrained. Hence, Big Data analytics become even more challenging for data collected via the IoT. In this chapter, we discuss the challenges pertaining to Big Data in IoT; these challenges are associated with data management, data processing, unstructured data analytics, data visualization, interoperability, data semantics, scalability, data fusion, data integration, data quality, and data discovery. We present these challenges along with relevant solutions.

3.1 Introduction

The Internet of Things (IoT) paradigm asserts that each individual object in everyday life can be equipped with sensors which can acquire useful information about the objects and will be on the network in one form or another [1]. Over the past decade, an increasing number of objects (e.g., smart devices, cars, intelligent roadways, pacemakers and other personal health monitoring units, refrigerator, cattle, smart billboards, etc.) have been connected to the Internet, collecting and exchanging data without requiring human-to-human or human-to-computer interaction. This network infrastructure enables anything and anyone to be connected anytime and anywhere.

P. Wongthongtham (✉) • J. Kaur • V. Potdar
Curtin University, Perth, Australia
e-mail: p.wongthongtham@curtin.edu.au

A. Das
Tripura University (A Central University), Agartala, India

Z. Mahmood (ed.), *Connected Environments for the Internet of Things*,
Computer Communications and Networks,
https://doi.org/10.1007/978-3-319-70102-8_3

At present, there are about 1.5 billion Internet-enabled PCs and over 1 billion Internet-enabled mobile phones [2], and it is expected that 50 to 100 billion smart devices will be connected to the Internet by 2020 [3]. According to IDC, the worldwide IoT market spend will increase from $592 billion in 2014 to $1.3 trillion in 2019 [4]. Sensors from these devices will simultaneously generate a huge amount of data in an automated way. In the future, 40% of all the data in the world will be generated by machine-to-machine communication [4].

We are all constrained by time, limited resources, and capability, making it impossible to manually handle this vast amount of data. This data continues to increase at a rapid pace because embedded sensor devices have steadily been increasing with advances in technology. It is the greatest force driving Big Data analytics. A comprehensive data analytics model or framework is needed to analyze this enormous amount of sophisticated data.

There are three key IoT elements which enable seamless and ubiquitous computing: (a) hardware, comprising sensors, actuators, and embedded hardware; (b) middleware, on-demand storage and computing tools for data analytics; and (c) presentation, perception of visualization and interpretation tools which can be extensively accessed on different platforms and adapted for different applications. IoT middleware solutions are gaining more attention in the marketplace as they simplify the sensor data by performing data binding, filtering, fusing, reasoning, etc. In addition, the variety of IoT applications that are built on top of this middleware poses further challenges. The IoT middleware consists of a mechanism to combine high-tech infrastructure with a service-oriented architecture and sensor networks to provide access to discordant sensor sources in a disposition-independent manner [5]. The IoT middleware needs to assist users to retrieve the data streams required for their application. It is evident that data analytics will be critical for IoT in this Big Data era. In this chapter, we discuss the challenges associated with Big Data in IoT.

The rest of this chapter is organized as follows. Sections 3.2 and 3.3 introduce IoT and Big Data, respectively. Sections 3.4, 3.5 and 3.6 explain the challenges facing IoT Big Data. The challenges include data management issues presented in Sect. 3.4, data analytics challenges presented in Sect. 3.5, and semantics challenges presented in Sect. 3.6. Section 3.7 concludes the chapter.

3.2 Internet of Things

The Internet is the most widely adopted technology, which has radically changed the way people communicate with each other. The Internet as we know it is a large network of interconnected servers that host a huge amount of valuable information. However, the Internet is changing rapidly, and it now connects machines, equipment, sensors, actuators, home appliances, surveillance cameras, and numerous other objects in our environment. This communication network does not require constant human intervention, and this new phenomenon of an interconnected world where everything is connected is referred to as IoT [6]. According to [7], the number of physical things that are now connected to the Internet is greater than the world's population. It is estimated that 25 billion devices were connected to the Internet in 2015, and this number is rising at an alarming rate. It is estimated that by 2020 we will have at least 50 billion devices feeding data to the Internet via IoT [8].

3.2.1 Definitions of the Internet of Things (IoT)

The IoT is syntactically comprised of two terms. The first term, "Internet," focuses on the vision which is network oriented; the second term, "Things," refers to the "objects" which are generic and are integrated to form a common framework. Hence, IoT is defined as "a worldwide network of interconnected objects uniquely addressable, based on standard communication protocols" [9]. Each and every thing connected to the Internet has a unique identifier such as MAC that addresses and communicates using the TCP/IP protocol. Radio-frequency identifiers (RFID) is a good example of the real power of IoT [6].

These "things" or objects interact with each other in order to accomplish a common goal. For example, smart electric cars such as Tesla have 18 sensors that work together automatically. This car can open the doors of a garage before the person arrives home; it can control the temperature, and it provides a framework whereby the user can design his/her own app and use this app to check the battery status and control the speed of the car from anywhere.

IoT is also known as the Internet of Objects; these are daily life objects that are interconnected through a network and possess ubiquitous intelligence [10]. IoT increases the Internet's ubiquity, because it integrates the objects so that they can communicate with other devices/objects and with humans. Yoo, Henfridsson et al. [11] define IoT as the combination of components which are both physical and digital. This combination results in the development of new products and creates innovative business models. Wortmann and Fluchter [12] mentioned that in IoT, physical things are combined with IT in the form of hardware and software, thereby improving the physical function of the associated things by means of additional IT-based services. With the combined IT-based services, the functionality of such things can be accessed locally as well as globally via the Internet. For example, home automation can convert a standard home to a smart home by using IoT devices. In a smart home, the homeowner can switch an air conditioner on or off before arriving home or switch off the lights after leaving home. The owner can also receive notification that an unauthorized person has entered the house and police can be called automatically. Moreover, a light bulb can act as a smart security system. The physical function of a bulb is to illuminate a specific area, but this physical function of a bulb can be enhanced with IoT. With IoT capability, this bulb can be used to detect the presence of a human being and can work as a security system which detects the intruder, turns on the flashing mode, and sends a message to the homeowner's smartphone.

According to Internet Telecommunication, IoT is “a global infrastructure for the information society, enabling advanced services by interconnecting things based on existing and evolving, interoperable information and communication technologies” [12]. Another paradigm of IoT is Cyber-Physical Systems (CPSs) as mentioned by [13].

3.2.2 Cyber-Physical Systems

Cyber-Physical Systems (CPSs) are new generation systems which integrate both physical and computational capabilities and can communicate with human beings by using various modalities [14]. These are engineered systems which are developed from the synergy of both physical and cyber components. CPS can be applied in medical services, robotics, avionics, etc. [15]. Future innovative technical developments are possible with CPS because CPSs have the ability to communicate with the physical world by means of computation [14].

Lee [16] mentioned that in CPS computational processes, network processes and physical processes are integrated. Physical processes are controlled and monitored by the embedded computers and networks by using the feedback loops, whereby computations are affected by the physical processes and vice versa.

CPS needs both the computing and networking technologies to capture the physical dynamics as well as the information. CPS requires the interaction between the computing, physical systems, control systems, and network systems in order to establish the interaction among them. CPS requires new design technologies. In CPS, software is embedded in physical devices whose principal goal is not only computation but also to combine computation with physical processes [17]. Autopilots are a good example of CPS. Autopilots were initially used in missiles but were later adopted in aircrafts. Autopilots include sensors and processors that are used to assist the human operator in controlling the aircraft. The airplane has high nonlinear dynamics, so it requires more complex and advanced technologies such as neural network and fuzzy logic, which ensures smooth trajectory navigation [18].

Nowadays, the terms “IoT” and “CPSs” are used interchangeably, although there are several differences between them. According to [17], CPSs and the IoT are almost similar because both use the same architecture. However, a CPS has several characteristics that distinguish it from IoT:

- In a CPS, every physical device has cyber capability. Every device has embedded software and system resources such as network bandwidth, and each device has limited system resources.
- CPSs require a greater integration of computation and physical processes compared to the IoT.
- CPSs are distributed systems which are networked by means of various network types such as wireless network, wired network, Bluetooth, GSM, and others.
- In a CPS, every component has different spatial and time granularity. Spatiality and time capabilities are the strictest constraints of CPS.

- A CPS requires very high degree of automation. For this purpose, feedback technologies are used in these systems. The advanced feedback technologies establish easy interaction between man and machine.
- Because they are complex, large-scale systems, CPSs are reliable and secure and have adaptive capabilities.

3.2.3 IoT Architecture

Said and Masud [6] suggest two main architectures for IoT: a three-layered architecture and a five-layered architecture. Other than these, several special-purpose architectures tailored for specific contexts are also found in the literature.

3.2.3.1 Three-Layered Architecture

The earliest proposed architecture for IoT was a three-layered architecture comprising a perception layer, network layer, and application layer.

The perception layer is used to identify objects in the IoT system [19]. This layer collects information about every object, and for this purpose, the perception layer uses the data gathered from RFID tags, cameras, sensors, etc. Sensors collect information about temperature, motion, acceleration, humidity in the air, etc., and the perception layer passes this information to the network layer [20].

The network layer is the main component of the three-layered IoT architecture [19]. The function of this layer is to securely transmit to the application layer the information collected by the perception layer, using the software and hardware instruments of the Internet. The medium of transmission could be wired or wireless such as Wi-Fi, Bluetooth, 3G, etc. The network layer also contains the information and management centers [20].

The application layer connects the IoT's social needs with industrial technology. It acts as a middle tier linking the industrial technology with the needs of humans. The applications which can be developed by IoT are smart health, smart home, smart farming, intelligent transportation, etc. [6].

3.2.3.2 Five-Layered Architecture

The three-layered architecture became inadequate with the rapid development of IoT; hence, a five-layered architecture was developed [20]. Currently, a TCP/IP protocol stack is used to facilitate communication between network hosts. Billions of devices are connected within the IoT system, creating a huge amount of traffic and requiring larger storage space. Hence, the next-generation architecture must be able to provide security and privacy for such a huge amount of data and should be scalable and interoperable [19]. So, for this purpose, five-layered architecture was proposed.

The first layer is known as the business layer. The main function of this layer is to define the IoT applications and is also responsible for the management of IoT applications and services. The business layer ensures data privacy and creates business models and graphs according to the information acquired from the application layer. Based on these generated models and graphs, one can predict future actions and goals.

The second layer is the application layer, the purpose of which is to determine the types of applications in IoT. This layer develops intelligent, safe, and authenticated applications of IoT. It works similarly to the application layer of the three-layered architecture. IoT can develop many applications such as smart health, smart home, smart farming, intelligent transportation, etc.

The third layer is the processing layer which handles the information collected by the perception layer. This layer is responsible for storing and analyzing the information. Functions of this layer are very critical and difficult, because the perception layer collects huge amounts of data about system objects. So, to handle such a huge amount of information, this layer uses techniques such as database software, intelligent processing, and cloud computing. This layer is linked to the database, and it stores in the database the information received from the transport layer. This layer performs some computations on the information and makes decisions automatically.

The next layer is the transport layer. It functions like the network layer of the three-layered architecture. This layer is also known as the transmission layer. The transport layer is responsible for receiving the information from the perception layer and transmitting it to the processing layer and vice versa. This layer uses many network technologies such as Wi-Fi, Bluetooth, etc. This layer is responsible for the secure transmission of data between the perception layer and the processing layer.

The last, the fifth, layer is the perception layer. It works similarly to the perception layer of the three-layered architecture. This layer collects information about every object in the IoT system such as the temperature and location of each object. This layer transmits collected data into signals. The layer uses technologies such as RFID, GPRS, etc. for the collection of data.

3.3 Big Data

As mentioned in [21], in the last 20 years or so, there has been a great increase in the volume of data in every field. A report from the International Data Corporation (IDC) in 2011 stated that 1.8ZB data was copied and found in the world, and within 5 years, the amount of data had increased ninefold [22]. For example, if we consider social media as a major source of data, it is anticipated that by mid-2019, there will be nearly 65 million Twitter tweets per day and around 190 million users [23]. So, given the colossal amount of data, “Big Data” is the term used to describe huge datasets. These datasets are very difficult to manage, acquire, perceive, and process by means of traditional tools in real-time environments.

According to [24], Big Data is defined as the data which is so huge that it cannot be captured, processed, and managed using traditional techniques. Big Data includes massive amounts of structured, unstructured, and semi-structured data, which require more real-time analysis than do the traditional datasets. Moreover, Big Data provides the opportunity to explore new values and to acquire an in-depth understanding of data.

Nowadays, because of its high potential, companies and government agencies are becoming more interested in Big Data and have undertaken major research on Big Data and its applications [21]. Big Data is relatively new, although the term has been around for a long time and has appeared in many scientific papers [25]. Big Data is not only about the volume of data; it has many other features apart from size. In the next section, we present various definitions of Big Data.

3.3.1 Definitions and Characteristics of Big Data

According to [26], Big Data consists of three Vs: volume, velocity, and variety. Volume indicates that the data generated by the Internet is very high in volume compared to that of earlier years. Velocity refers to the speed of data generation; i.e., systems generate data at a very high speed compared to the speed of traditional systems. Variety refers to the various forms of data; that is, data is present in many forms on the Internet. These three Vs were originally suggested by Gartner for describing Big Data elements. Gantz and Reinsel [22] added a fourth V to the characteristics of Big Data: value. The fourth characteristic is highly accepted because it defines the actual meaning and requirement of Big Data. Chen, Mao, and Liu [21] added a fifth V: veracity. Hence, Big Data analytics is required to disclose hidden data (or gather actionable insights) from very huge datasets, which are complex, diverse, and very big. The main characteristics of Big Data are described below, in more detail.

3.3.1.1 Volume

Volume indicates the data magnitude and the huge amount of different kinds of data which are generated by various sources, and this data is continuously increasing [27]. The size of Big Data is in terabytes and petabytes. IBM conducted a survey of 1144 respondents in mid-2012 and found that only half of the respondents believed that a Big Data dataset exceeded one terabyte [28]. One terabyte of storage is equivalent to 1500 CDs, which can store around 16 million photographs. According to [29], in one second, Facebook processes one million photographs, and it stores 260 billion photographs in 20 petabytes of storage space. So, one can only imagine the volume of data that is being processed, managed, stored, and analyzed. The volume of data needs to be measured in terabytes or petabytes, because huge amounts of data are generated by different sources such as sensors. Hence, it is difficult, if not impossible, to manage such a huge amount of data using traditional database techniques [30].

As an example, smart traffic management systems are one of the developments of IoT. Nowadays, because of affordable car prices, the number of cars on the road has increased significantly leading to traffic congestion. To manage congestion, traffic management systems are connected to the digital road map of the city, and traffic displays are installed within cities to guide drivers. For traffic management, sensors are connected to the traffic lights, and these sensors send information to a central server about the number of vehicles. The analytical software at central location receives real-time data from sensors, traffic lights, and digital road maps. When the number of vehicles on a road exceeds the total capacity, traffic screens advise drivers to take a detour 1 km before the signal, which reduces both the travel time and the fuel consumption. This is possible because a large amount of sensor data from road sensors is sent to a central management system for real-time analysis. However, such large datasets cannot be managed using traditional database techniques and therefore require Big Data analytics approaches. The analysis of such large datasets can reveal hidden patterns and information which are then used to improve the traffic management systems.

3.3.1.2 Variety

Variety refers to the heterogeneous nature of Big Data such as data that is collected by different types of sources such as sensors, social networks, etc. The collected data could be of any type such as audio, video, text, or data logs, and it could be structured, semi-structured, or unstructured. Structured data is data which is stored in tabular form in spreadsheets or in a relational database. The data which is not organized in a structured way is called unstructured data, such as text in the form of paragraphs on the Internet. Semi-structured data is the data whose formats lie between structured and unstructured data. The format of semi-structured data does not follow strict standards. An extensible markup language, XML, which is used to exchange data on Internet, is an example of semi-structured data. XML documents have data tags, which are readable by machines [27].

The data which are generated by mobile phones, such as game data, text messages, and blogs, are mostly unstructured [31]. For example, in a smart traffic management system, data from different sources such as sensors, traffic lights, and digital road map are analyzed for better traffic management. Every source will produce data in a different form; these different types of data presentations are managed and analyzed in a central location for better decision-making.

3.3.1.3 Velocity

Velocity indicates the speed at which the data are generated and analyzed. With the development of digital devices such as sensors and smartphones, an extraordinary amount of data is created which requires real-time analytics. Data generated through sensors are collected and analyzed in real time [32]. Retailers such as Amazon are also generating data at very high speeds. For example, Wal-mart processes approximately 1 million transactions per hour [33]. Data generated through mobile phones help to produce personalized offers for customers. Another example of the velocity of data is the data generated by traffic sensors. These sensors gather and transfer information in real time, because the data collected by these sensors are useful only if they give information to the driver before she/he reaches the congested area. Hence, data analysis needs to be done at an equally fast speed because data have time value; i.e., after a specific time, the data will no longer be useful.

3.3.1.4 Value

This is the most important characteristic of Big Data. It refers to the exploration of data to discover hidden patterns and values of large datasets of different types by using different techniques [21]. Very valuable data can be acquired by analyzing a huge amount of Big Data. It also has the potential to provide cost-beneficial criteria. For example, sensors in the IoT system of a smart traffic management system send huge amounts of data to a central control system, where the data are processed and analyzed. Data have value only if they can assist in predicting the future and current traffic conditions of traffic lights.

3.3.1.5 Veracity

Veracity refers to the accuracy, reliability, and truthfulness of data, which means that the data are noise-free and nonredundant and can therefore be confidently used for decision-making and for future predictions [34]. Achieving veracity of data is very difficult because data are produced by different sources.

3.3.2 Big Data Analytics

Big Data has demonstrated its great potential to transform decision-making in the business realm. Efficient and effective processes are needed to turn the high volume of rapidly generated and diverse data into significant information that can inform decision-making. Big Data analytics are the techniques used to procure and analyze an intelligence acquired from Big Data. There are four types of analytics which are presented here.

3.3.2.1 Descriptive Analytics

Descriptive analytics are used to diagnose what has happened or is happening [35]. These analytics are applied to categorize, classify, and consolidate massive amounts of historical data in order to understand what the data imply. They include the presentation of raw data in summarized or query form to manage otherwise elusive information. This sort of analysis is mainly concerned with processing the very diverse collected data by monitoring data from device sensors and databases to detect patterns and trends in such data [35]. Descriptive analytics can produce data visualization in the form of tables, drawings, maps, interactive dashboards, charts (fever, pie, bar, etc.), etc. to summarize and report the trends.

3.3.2.2 Diagnostic Analytics

Diagnostic analytics are applied in order to determine why a phenomenon is occurring or has occurred and to analyze the factors leading to this occurrence which may include the inputs and operational policies [27]. Diagnostic analytics can benefit from sensitivity analysis using a simulation model of the system that mimics the current operation.

3.3.2.3 Predictive Analytics

Predictive analytics harnesses sophisticated machine learning and data mining techniques to examine the historical data in an effort to predict the upcoming future. Predictive analytics is capable to detect hidden patterns from data in large scale and cluster these data into segments which share common characteristics. Predictive analytics are used to estimate efficiency based on planned inputs. They can be applied to all domains ranging from weather forecasting and market volatility predictions to predictions of customers' next moves based on their spending and even on what they tweet [27]. It also has applications in other domains such as healthcare, education, marketing, supply chain logistics, etc. In essence, predictive analytics explore and interpret patterns in order to find relationships among the data. Predictive analytics use simulation models to predict a future occurrence using a set of inputs and "what-if" scenarios.

3.3.2.4 Prescriptive Analytics

Prescriptive analytics are concerned with how we can make it happen and what the consequences will be [28]. Prescriptive analytics are used to identify the policies and inputs that will lead to a desired outcome and may include identifying changes in input parameters and policies that will reduce the cycle time and increase throughput in order to reach the desired levels. Prescriptive analytics are intended to provide the optimal solution(s) to an existing problem through the use of optimization and simulation techniques. This significantly helps decision-makers to select the best option.

3.4 Management Challenges of Internet of Things Big Data

In this section, we discuss the challenges of managing IoT data, including data and process challenges.

3.4.1 Data Challenges

Challenges associated with Big Data characteristics are discussed below.

3.4.1.1 Massive Amount of Data Collected

According to [6], the main problem is related to the huge amount of information which is collected through RFID. IoT systems may have millions of devices. Every object in the IoT generates information about itself. This generated information must be gathered and amounts to a massive quantity of data, producing problems of transmission, storage, and processing.

The transmission issue relates to the necessity of transferring all the gathered information in real time, which is very difficult because the bandwidth which is required to transfer that information might not be available at that time. Another problem is related to the storage of information because a large amount of space is required for storage and backup. The last issue is the processing problem. In order to determine the actions that must be taken, the information about things must be handled by web applications, and information must be handled in real time [36].

The volume of data is increasing day by day. As mentioned in [30], 80,000 petabytes of data were stored across the world in 2000, and this is predicted to rise to 35 zettabytes by 2020. In today's world, many objects and/or activities are tracked and recorded, such as environmental data, medical data, industrial data, etc. Information is even recorded for every event; for example, speed cameras store information about speed limit breaches, etc. What we observe nowadays is that massive amounts of data are being stored, but the processing of such huge datasets is becoming difficult; hence, the percentage of processed data is decreasing, resulting in blind zones [37].

3.4.1.2 Various Forms of Data Collected

The data which comes from sensors are sometimes combined with other unstructured data, so there is a strong relationship between sensor and other unstructured data. So different forms of data such as structured, semi-structured, and unstructured are collected and stored by Big Data. Of the massive amounts of data that are collected, only 20% is processed; the remaining 80% cannot be processed and analyzed using traditional techniques. Hence, most of the collected data are not useful for decision-making [30]. In addition, there needs to be a technique which can effectively combine structured data with unstructured images, text, or data [38].

3.4.1.3 Data Transmission Speed

The transmission speed of data on the Internet is also known as velocity. In order to explore and acquire some insight about the data, this high-speed data needs to be analyzed in real time. The current software applications can generate data streams at very high speeds which can be very difficult to analyze in real time [39]. This is still a challenge for Big Data. For example, in 1999, the data warehouse of Walmart could store data up to 1000 terabytes, but in 2012 it had increased to 2.5 petabytes of data [40]. This shows a rapid increase in data accessed through the sensors and presents new challenges regarding the storage processing and analysis of such high-speed data in real time.

3.4.1.4 Time Series for Data Analysis

Generally, in the case of sensors, some events are captured at a specific point in time. The data captured by specific events or at specific times are sometimes useless. However, if something serious happens, it must be recorded and addressed. As a starting point, it is good to use a static threshold to analyze the datasets, gathered at particular time intervals. Most technical companies find this difficult to handle [38].

3.4.1.5 Security and Privacy

In the IoT, data are transferred between objects using a wireless medium; therefore, it becomes critical to ensure the privacy and security of information. There could be a number of attacks such as physical attacks or wireless information attacks, which can affect the security and authenticity of the transmitted information. The attacker can attack the IoT devices physically or steal the information during transmission. Most of the IoT devices do not accept security packages, which leads to low self-defense.

Privacy means to ensure three things: firstly who collects the personal information, secondly how this information is collected, and lastly the time when the information is gathered. Moreover, the acquired personal information must be used by an authorized person and should be stored on an authorized server, and only an authorized client should be able to access the information [41].

3.4.2 Process Challenges

Challenges relating to the processing of Big Data are discussed below.

3.4.2.1 Selective Data Acquisition

In today's world, data acquired using sensors and other devices are in petabytes. However, not all collected data are important, so data must be filtered and compressed. These filters decide the data that should be collected and those that should be discarded. For example, if all the sensors except one are giving readings within an acceptable range, then it is possible that that sensor is either faulty or something has gone wrong in that sensing area, which should be investigated. Therefore, the task of designing a smart filter to make such decisions in real time presents a significant challenge [41].

3.4.2.2 Data Extraction

The gathered information is mostly in different formats. For instance, a health record can comprise MRI data, prescriptions, medical reports, x-ray images, etc., all of which information is in different formats. In order for this information to be used effectively, the data must be transformed into a single structured format. Therefore, a new extraction process is needed that can extract the required data from the source and transform it into a structured format suitable for analysis. The correct design and maintenance of this extraction process is a big challenge [41].

3.4.2.3 Data Heterogeneity

Data gathered from diverse sources are heterogeneous in nature; hence, data processing is not a straightforward process because finding, identifying, and understanding information are difficult when the data sources cannot be integrated seamlessly. When data are heterogeneous, analysis becomes difficult because the data have different structures and different semantics. Thus, the integration of heterogeneous data for processing in real time presents a major challenge. New data mapping and data integration systems need to be designed to ensure seamless integration of data from heterogeneous sources.

3.4.2.4 Nature of Big Data

Big Data is unreliable, dynamic, heterogeneous, noisy, and interconnected [42]. Sometimes, noisy data is more useful than small datasets because repeated patterns can be extracted from general statistics. Hidden information can also be revealed through interrelation analysis [30]. Redundant data can sometimes be useful in finding missing data and can also be analyzed to find unreliable relationships and to discover hidden models [43].

3.5 Analytics Challenges of the IoT Big Data

In this section, we discuss several challenges associated with IoT data analytics. These challenges are related to unstructured data analytics (i.e., text analytics, audio analytics, video analytics, and social media analytics) and visualization challenges.

3.5.1 Analytics Challenges over Unstructured Data

The analysis of unstructured data such as text, audio, video, and social media is difficult. Text analytics are those procedures that extract information from textual data. Some examples of textual data are feeds from social networks like Facebook, Twitter, etc. and online forums, blogs, emails, white papers and other documents, etc. It involves statistical analysis, natural language processing, and deep learning. Transforming large volumes of randomly generated text into meaningful abstracts, which support cue-based decision-making, is challenging. Apple’s Siri and IBM’s Watson are examples of commercial question answering systems which have been implemented in various domains like healthcare, education, finance, marketing, and banking, and these systems rely on complex natural language processing, information retrieval, and knowledge-based approaches [28].

Audio analytics refer to processes that analyze and extract information from raw audio data. It is also known as speech analytics. Business process outsourcing (BPO) uses audio analytics for the effective analysis of recorded calls, which in turn helps to improve customer experience, appraise agent performance, elevate sales turnover rates, cue into customer behavior, identify service problems, and monitor compliance with security and privacy policies, among other tasks [27]. Audio analytics systems are designed to scrutinize a live call, forecast recommendations based on customers’ past interactions, and provide feedback to BPO agents in real time.

Video analytics are those procedures that monitor, analyze, and extract meaningful information from raw video streams. The increased ubiquity of CCTV cameras and video sharing websites is leading to the proliferation of computerized video analysis. However, a key challenge is the enormity of the video data. Big Data analytics overcomes the need for manual processing to automatically scrutinize and derive intelligence from millions of hours of streaming video. In modern times, video analytics have been applied in automated surveillance systems, in order to detect trespassing in restricted zones, identify unknown objects, and recognize spying or suspicious activities. On detection of a threat, an automated alarm goes off to notify the security personnel in real time. In retail outlets, data generated by CCTV cameras may provide business intelligence to discover the demographics, choices, behaviors, buying patterns, etc. of consumers [27].

Social media analytics are the processes that analyze and extract meaningful information from social media channels such as Facebook, Twitter, LinkedIn, Instagram, Wikipedia, wikiHow, YouTube, ResearchGate, Ask.com, etc. Social media analytics is a relatively new area. The challenges of the modern social analytics are its data-centric nature and its research which is interdisciplinary and may include the domains of psychology, sociology, computer science, mathematics, economics, and statistics. The primary application of social media analytics has been in marketing and business management. Content generated by users (e.g., photos, videos, emotions, thoughts, etc.) and the relationships and synergy between the network entities (e.g., people, businesses, and merchandise) are the different sources of information in social media.

3.5.2 Visualization Challenges

Visualization helps to improve the human cognitive process by quickly identifying interesting and significant events and patterns in collected data [44–47]. Some other benefits of visualization include better understanding of large datasets, quick recognition of errors and outliers in datasets, facilitation of hypothesis formation from data, etc. [48]. A wide range of studies on visualization have been carried out, proposing techniques and methods to facilitate the process in order to obtain insights from data; some of these techniques include visualization of unstructured temporal data with a parallel rendering algorithm [49], taxonomies of interaction techniques [50], the focus-on-context technique [51], tree maps for visualizing hierarchical data structure while making use of all of the available space [52], and artificial reality in visualization [53].

The total amount of data generated is expected to experience a significant growth. However, approximately 3% of the collected data was tagged, and approximately 0.5% of the world’s digital data was analyzed [54]. Approaches are needed to represent data in a more intuitive way to improve the understanding of data and provide adequate support for decision-making. Visualization is expected to assist in tackling some of these challenges. Visualization challenges include its applicability for a large volume of data, the possibility of visualization of data being presented in different data formats, speed, and effectiveness of data presentation.

3.6 Semantics Challenges of the IoT Big Data

In this section, we present challenges related to IoT data semantics. These challenges are associated with data interoperability, data semantics, data scalability, data fusion, data integration, data quality and trustworthiness, and data discovery.

3.6.1 Data Interoperability Challenges

To make the data interoperable, semantic description of and an ontology for the data are required. Ontologies describe formally shared conceptualizations of a domain of interest [55]. Solodovnik [56] described the concept of ontology from its philosophical origins to its adoption within the IT field as follows: *Philosophically, ontology is a systematic explanation of being that describes the features of Reality. Nowadays Ontology is proliferating in organizing Knowledge of different domains managed by advanced computer tools. Ontology qualifies and relates semantic categories, dragging, however, the idea of what, since the seventeenth century, was a way to organize and classify objects in the world. Ontology maximizes the reusability and interoperability of concepts, capturing new Knowledge within the most granular levels of information representation. Ontology is subjected to a continuous process of exploration, formation of hypothesis, testing and review.*

Data will be interoperable for users, who use the same ontology. In most cases, ontology and semantic description are defined only for a specific project, but for achieving global semantic interoperability, a common definition of ontology and semantic description framework must be adopted. For this reason, the ontologies must be reusable by a large number of applications. The sharing of the ontologies of current and previous applications is an effective means of achieving semantic interoperability on a global level.

There are millions of heterogeneous devices in our environment. These heterogeneous devices must be connected in such a way that they can communicate easily. We need semantic interoperability which enables all the stakeholders to interpret and access the data from these heterogeneous devices without any issue. Within the IoT, objects/things are required to exchange data with other things and users on the Internet. This data must be processed and interpreted by machines in such a way that information communication can be automated in the IoT. Data semantic annotation provides information that is machine interoperable, and this information can reveal the source of data, relationship of data with surroundings, provider of data, quality of data, and description of technical and nontechnical terms [57]. Therefore, the accessing and processing of data from a number of heterogeneous devices are going to become increasingly challenging in the years to come.

3.6.2 Data Semantics Challenges

Millions of heterogeneous devices are connected to different types of sensors in order to collect real-world data and to communicate with other devices. Interoperable service-oriented technologies are intended to share the real-world data among these heterogeneous devices to integrate and fuse these semantic data [58]. Data semantics is one of the major elements of data analysis. It is a challenging task to deal with different data structures and information types and to analyze the data as the structure of information is very complex. Also, the system does not have adequate knowledge enabling it to describe fully the semantic meaning of the analyzed information. Computer cognitive resonance techniques have been proposed by [59], which can solve the problem by using a cognitive information system that uses features extracted from records and knowledge in the database. It is quite conducive to the analysis of the semantic data of different information records. The integration of various heterogeneous collections of data has become a colossal issue as the existing data sources are very sparse and incomplete which makes it an onerous task to find a logical connection between the data.

3.6.3 Data Scalability Challenges

It is challenging for data engineers to create domain knowledge models and semantic annotation frameworks which can describe a huge number of devices in the IoT. Domain knowledge must be associated with semantic descriptions of data because IoT data can refer to separate phenomena. In many applications, to define IoT data's spatial aspects, linked open data (an approach that interconnects different resources of IoT) are used as domain knowledge. However, linked sensor data is mostly inconsistent and contains numerous errors. As a solution for this problem, most of the applications design and maintain their own domain knowledge. However, this limits their interoperability. Another big challenge concerns granularity description; if the terms and concepts are very specific, then the domain knowledge is very extensive. The semantic web community has done a great deal of work in developing an efficient technique for storing and querying large semantic data in a distributed environment. However, the challenges in handling semantic data are the scale of data developed by IoT resources, the changing status of resources and data, and the volatility of the IoT environment. Research should address these issues and develop solutions to define linked IoT data which can analyze the links between the resources, and semantic repositories must be developed which can access and query the sensory data [57].

3.6.4 Data Fusion Challenges

Data fusion is used as a means of improving the quality of the data. Data fusion focuses on the computation of structured and comparable semantic data in order to obtain appropriate decisions. Semantic data fusion is challenging as data are acquired from multiple sensors, and different types of algorithms are used to improve the quality and accuracy of the data. Data fusion in the IoT, based on such multi-sensor data, produces new information. Information fusion is the major part of the information and comprises of several theories, techniques, and algorithms. It can improve the accuracy and produce more accurate results as the data is produced from multiple sensors and cognate information which is obtained from the affiliated databases. The major function of information fusion is to integrate diverse types of semantic data, without which the related data and information cannot be integrated, because it is impossible to process information fusion computation using a variety of algorithms as heterogeneous data cannot be correlated.

3.6.5 Data Integration Challenges

Mostly, IoT data are generated from sensor devices, humans, or a physical entity. To create multiple environment abstraction, this data can be merged with other data. This data can be combined with the processing chain in an application which already exists, and this data can support situation awareness. It is necessary that different types of data be combined seamlessly [60]. Semantic description assists this combination process by facilitating interoperability among different sources of data. However, to enable IoT data integration, the mapping and analysis of different semantic description models are required.

The combination of appropriate data that reside in a huge number of data sources which are heterogeneous in nature may conflict in terms of value and structure. This type of data integration allows the user to have a unique view of the data. Semantic technology is the fundamental technology of data integration. Data integration systems are commonly defined as a triple GSM, where G is the global schema, S is the discordant set of source schemas, and M is a mapping that maps queries between the source and the global schemas. For each of G and S, their respective relations are defined in languages which consist of symbols. In this way, huge amounts of linked data are transformed from the raw IoT data. Using the basic idea of data integration, different models at schema level are merged together when users need an integration of the relevant heterogeneous data. As a result, the data at the instance level are presented in a unified view to achieve data integration. By means of mapping, different models at schema level can be merged. These mappings are obtained in several ways. Predefined mapping is the first method of mapping which may produce highly accurate data, but is not efficient. The second method is based on mapping which is determined with the help of computation by following several principles such as the linked open data cloud. Schema level mapping is one of the main functions of integrating data.

3.6.6 Data Quality and Trustworthiness Challenges

Sensor devices generate IoT data which have errors and quality issues. Quality means that data must be complete and accurate and must be available when required. The quality of data collected through sensors can change over time. For example, this occurs if there is any environmental change, due to any faulty device or due to any error in the settings of device. It is not possible to avoid inaccuracy in IoT data. To retrieve and process quality data, readings from IoT devices need to be detected and filtered, in addition to having semantic descriptions of the attributes of quality. This could also assist with error detection. Another main issue is trust, especially when data are generated by many different sources. Trustworthiness of data and sources can be achieved by identifying the data provider and verifying data accuracy and reliability, along with the semantics which describe the quality and trust attributes of sources and providers. Although semantics can be used to define trust and reliability attributes, several major issues still need to be addressed such as the development of a trust model, feedback, and the development of a verification mechanism [57].

3.6.7 Data Discovery Challenges

The efficient handling of data and storage is becoming more difficult with time as the volume of data and semantic description is increasing day by day. Sensor data must be stored with semantic descriptions, and this data can be stored temporarily or for a lengthy period. The main challenges include designing and developing repositories, publishing the semantic data, accessing the semantic data in distributed environments, and developing effective indexing and discovery mechanisms. To address these issues, an effective mechanism for information indexing, search, access, and query is required. Such mechanism could be used for the discovery of relevant data from many sources, real-time query and aggregation of multiple data streams, description of various events and data which are generated by many sources, and data discovery when semantic data is distributed among multiple repositories. Cloud computing is a good technical approach which can overcome some of these issues, but in order to handle, process, and maintain data, the solution must be scalable and efficient; it is not sufficient to simply develop a centralized and non-scalable solution and put it in the cloud [57].

3.7 Conclusion

New properties are emerging in IoT with every passing day. Inter-conceivable service-oriented technologies are imperative for sharing real-world data among discordant devices to integrate and fuse multisource IoT data. The IoT can offer only trivial and insignificant benefits if it cannot integrate and incorporate useful information from the data generated by multiple interconnected devices. This is where Big Data analytics plays a critical role and bring out the value from the information and data gathered by IoT devices. Hence, research in the field of Big Data analytics and IoT is becoming important as it has diverse application areas, especially in the context of smart cities. This chapter introduced and described number of challenges at the intersection of IoT and Big Data to provide a holistic view on how to manage these challenges effectively. Managing such large datasets poses substantial difficulties under computing and time constraints. We elaborated the challenges associated with data management (such as size and forms of data, time series analysis, security, and privacy), data processing (such as data acquisition, extraction, and heterogeneity), unstructured data analytics, data visualization, and data semantics (such as interoperability, data fusion, data integration, data quality, and data discovery). We then described the latest solutions to address these upcoming challenges to provide guidance for future research in this field. Overall, this chapter will guide researchers by providing the most up-to-date information on challenges and solutions at the intersection of IoT and Big Data.

References

1. Aggarwal CC, Ashish N, Sheth A (2013) The internet of things: a survey from the data-centric perspective. In: Managing and mining sensor data. Springer, Boston, pp 383–428
2. Perera C, Vasilakos AV (2016) A knowledge-based resource discovery for internet of things. Knowl-Based Syst 109:122–136
3. Sundmaeker H, Guillemin P, Friess P, Woelfflé S (2010) Vision and challenges for realising the internet of things. The Cluster of European Research projects on the Internet of Things, European Commission
4. Verizon (2016) State of the market: internet of things 2016
5. Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst 29:1645–1660
6. Said O, Masud M (2013) Towards internet of things: survey and future vision. Int J Comput Netw IJCN 5:1–17
7. Said O, Tolba A (2012) SEAIoT: scalable e-health architecture based on internet of things. Int J Comput Appl 59
8. Evans D (2012) The internet of things how the next evolution of the internet is changing everything (April 2011). White Paper. Cisco Internet Business Solutions Group (IBSG)
9. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54:2787–2805
10. Xia F, Yang LT, Wang L, Vinel A (2012) Internet of things. Int J Commun Syst 25:1101
11. Yoo Y, Henfridsson O, Lyytinen K (2010) Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. Inf Syst Res 21:724–735

12. Wortmann F, Flüchter K (2015) Internet of things. Bus Inf Syst Eng 57:221–224
13. Salim F, Haque U (2015) Urban computing in the wild: a survey on large scale participation and citizen engagement with ubiquitous computing, cyber physical systems, and internet of things. Int J Hum-Comput Stud 81:31–48. https://doi.org/10.1016/j.ijhcs.2015.03.003
14. Baheti R, Gill H (2011) Cyber-physical systems. Impact Control Technol 12:161–166
15. ZHANG Y, XIE F, DONG Y et al (2013) High fidelity virtualization of cyber-physical systems. Int J Model Simul Sci Comput 4:1340005
16. Lee EA (2006) Cyber-physical systems-are computing foundations adequate. 2
17. Wan J, Yan H, Suo H, Li F (2011) Advances in cyber-physical systems research. TIIS 5:1891–1908
18. Chao H, Cao Y, Chen Y (2010) Autopilots for small unmanned aerial vehicles: a survey. Int J Control Autom Syst 8:36–44
19. Khan R, Khan SU, Zaheer R, Khan S (2012) Future internet: the internet of things architecture, possible applications and key challenges. IEEE:257–260
20. Wu M, Lu T-J, Ling F-Y et al (2010) Research on the architecture of internet of things. IEEE:V5-484–V5-487
21. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19:171–209
22. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC Iview 1142:1–12
23. Schonfeld E (2010) Costolo: twitter now has 190 million users tweeting 65 million times a day. Techcrunch June 8
24. Manyika J, Chui M, Brown B et al (2011) Big data: the next frontier for innovation, competition, and productivity
25. Hashem IAT, Yaqoob I, Anuar NB et al (2015) The rise of "big data" on cloud computing: review and open research issues. Inf Syst 47:98–115
26. Zikopoulos P, Eaton C (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, New York
27. Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag 35:137–144
28. Schroeck M, Shockley R, Smart J et al (2012) Analytics: the real-world use of big data: how innovative enterprises extract value from uncertain data, executive report. IBM Institute for Business Value Saïd Business School, University of Oxford
29. Beaver D, Kumar S, Li HC et al (2010) Finding a needle in haystack: facebook's photo storage, pp 1–8
30. Nasser T, Tariq RS (2015) Big data challenges. J Comput Eng Inf Technol 4:3
31. Russom P (2011) Big data analytics. TDWI Best Pract Rep Fourth Quart:1–35
32. Cukier K (2010) Data, data everywhere: a special report on managing information. Economist Newspaper, London
33. Ragothaman B, Prabha MS, Jose E, Sarojini B (2016) A survey on big data and internet of things. World Sci News 41:174
34. Shao G, Shin S-J, Jain S (2014) Data analytics using simulation for smart manufacturing. In: Proceedings 2014 winter simulation conference. IEEE Press, pp 2192–2203
35. Lakshman TV, Madhow U (1997) The performance of TCP/IP for networks with high bandwidth-delay products and random loss. IEEEACM Trans Netw ToN 5:336–350
36. Vilamovska A-M, Hatziandreu E, Schindler HR et al (2009) Study on the requirements and options for RFID application in healthcare
37. Deshpande B (2016) 3 challenges unique to IoT analytics. https://www.owler.com/reports/simafore/3-challenges-unique-to-iot-analytics/1476315363392
38. Yassin AT (2014) Analyzing 6Vs of big data using system dynamics. In: 2nd scientific conference of the College of Science 2014
39. McNulty E (2014) Understanding Big Data: The Seven Vs. http://dataconomy.com/2014/05/seven-vs-big-data/
40. Chan H, Perrig A (2003) Security and privacy in sensor networks. Computer 36:103–105

41. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endow 5:2032–2033
42. Katal A, Wazid M, Goudar RH (2013) Big data: issues, challenges, tools and good practices. IEEE:404–409
43. Pradeepa A, Thanamani A (2013) Significant trends of big data analytics in social network. NGM Coll, India
44. Bauer MI, Johnson-Laird PN (1993) How diagrams can improve reasoning. Psychol Sci 4:372–378
45. Larkin JH, Simon HA (1987) Why a diagram is (sometimes) worth ten thousand words. Cogn Sci 11:65–100
46. Mayer RE, Gallini JK (1990) When is an illustration worth ten thousand words? J Educ Psychol 82:715
47. Card SK, Mackinlay JD, Shneiderman B (1999) Readings in information visualization: using vision to think. Morgan Kaufmann, San Francisco
48. Ware C (2012) Information visualization: perception for design. Elsevier, Amsterdam
49. Ma K-L, Stompel A, Bielak J et al (2003) Visualizing very large-scale earthquake simulations. In: Supercomput. 2003 ACMIEEE conference IEEE, pp 48–48
50. Yi JS, ah Kang Y, Stasko J (2007) Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans Vis Comput Graph 13:1224–1231
51. Lamping J, Rao R, Pirolli P (1995) A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM Press/Addison-Wesley Publishing Co, pp 401–408
52. Johnson B, Shneiderman B (1991) Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In: Proceedings of 2nd conference on visualization. IEEE Computer Society Press, pp 284–291
53. Erickson T (1986) Artificial realities as data visualization environments: problems and prospects. Virtual Real-Appl Explor:3–22
54. Tam NT, Song I (2016) Big data visualization. In: Information science and applications ICISA 2016. Springer, pp 399–408
55. Gruber TR (1993) Toward principles for the design of ontologies used for knowledge sharing
56. Solodovnik I (2010) ONTOLOGY: from philosophy to ICT and related areas
57. Payam B, Wei W, Cory H, Kerry T (2012) Semantics for the internet of things: early progress and back to the future. Int J Semantic Web Inf Syst IJSWIS 1:1–21. https://doi.org/10.4018/jswis.2012010101
58. Nugraheni E, Akbar S, Saptawati GAP (2016) Framework of semantic data warehouse for heterogeneous and incomplete data. In: Region 10 symposium. TENSYMP 2016 IEEE. IEEE, pp 161–166
59. Ogiela L, Ogiela MR (2015) Semantic data analysis algorithms supporting decision-making processes. In: Broadband Wireless Computing and Communication Applications. BWCCA 2015 10th international conference on IEEE, pp 494–496
60. Sheth AP (2011) Computing for human experience: semantics empowered cyber-physical, social and ubiquitous computing beyond the Web