IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 25, 2016, accepted March 14, 2017, date of publication March 29, 2017, date of current version May 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2689040

Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges

MOHSEN MARJANI[1], FARIZA NASARUDDIN[2], ABDULLAH GANI[1], (Senior Member, IEEE), AHMAD KARIM[3], IBRAHIM ABAKER TARGIO HASHEM[1], AISHA SIDDIQA[1], AND IBRAR YAQOOB[1]

[1]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[3]Department of Information Technology, Bahauddin Zakariya University, Multan, Punjab, Pakistan

Corresponding authors: Fariza Nasaruddin (fariza@um.edu.my); Abdullah Gani (abdullah@um.edu.my)

This work was supported by the Malaysian Ministry of Higher Education under University Malaya Research Grant (UMRG) Project/Program UM.0000168/HRU.RP.IT and RP029D-14AET.

ABSTRACT Voluminous amounts of data have been produced, since the past decade as the miniaturization of Internet of things (IoT) devices increases. However, such data are not useful without analytic power. Numerous big data, IoT, and analytics solutions have enabled people to obtain valuable insight into large data generated by IoT devices. However, these solutions are still in their infancy, and the domain lacks a comprehensive survey. This paper investigates the state-of-the-art research efforts directed toward big IoT data analytics. The relationship between big data analytics and IoT is explained. Moreover, this paper adds value by proposing a new architecture for big IoT data analytics. Furthermore, big IoT data analytic types, methods, and technologies for big data mining are discussed. Numerous notable use cases are also presented. Several opportunities brought by data analytics in IoT paradigm are then discussed. Finally, open research challenges, such as privacy, big data mining, visualization, and integration, are presented as future research directions.

INDEX TERMS Big data, Internet of Things, data analytics, distributed computing, smart city.

I. INTRODUCTION

The development of big data and the Internet of things (IoT) is rapidly accelerating and affecting all areas of technologies and businesses by increasing the benefits for organizations and individuals. The growth of data produced via IoT has played a major role on the big data landscape. Big data can be categorized according to three aspects: (a) volume, (b) variety, and (c) velocity [1]. These categories were first introduced byGartner to describe the elements of big data challenges [2]. Immense opportunities are presented by the capability to analyze and utilize huge amounts of IoT data, including applications in smart cities, smart transport and grid systems, energy smart meters, and remote patient healthcare monitoring devices.

The widespread popularity of IoT has made big data analytics challenging because of the processing and collection of data through different sensors in the IoT environment. The International Data Corporation (IDC) report indicates that the big data market will reach over US$125 billion by 2019 [3]. IoT big data analytics can be defined as the steps in which a variety of IoT data are examined [4] to reveal trends, unseen patterns, hidden correlations, and new information [5]. Companies and individuals can benefit from analyzing large amounts of data and managing huge amounts of information that can affect businesses [6]. Therefore, IoT big data analytics aims to assist business associations and other organizations to achieve improved understanding of data, and thus, make efficient and well-informed decisions. Big data analytics enables data miners and scientists to analyze huge amounts of unstructured data that can be harnessed using traditional tools [5]. Moreover,big data analytics aims to immediately extract knowledgeable information using data mining techniques that help in making predictions, identifying recent trends, finding hidden information, and making decisions [7].

Techniques in data mining are widely deployed for both problem-specific methods and generalized data analytics. Accordingly, statistical and machine learning methods are utilized. IoT data are different from normal big data collected via systems in terms of characteristics because of the various sensors and objects involved during data collection, which include heterogeneity, noise, variety, and rapid growth. Statistics [8] show that the number of sensors will be increased by 1 trillion in 2030. This increase will affect the growth of big data. Introducing data analytics and IoT into big data requires huge resources, and IoT has the capability to offer an excellent solution. Appropriate resources and intensive applications of the platforms are provided by IoT services for effective communication among various deployed applications. Such process is suitable for fulfilling the requirements of IoT applications, and can reduce some challenges in the future of big data analytics. This technological amalgamation increases the possibility of implementing IoT toward a better direction. Moreover, implementing IoT and big data integration solutions can help address issues on storage, processing, data analytics, and visualization tools. It can also assist in improving collaboration and communication among various objects in a smart city [9]. Application areas, such as smart ecological environments, smart traffic, smart grids, intelligent buildings, and logistic intelligent management, can benefit from the aforementioned arrangement. Many studies on big data has focused on big data management; in particular, big data analytics has been surveyed [10], [11]. However, this survey focused on IoT big data in the context of the analytics of a huge amount of data. The contributions of this survey are as follows.

a) State-of-the-art research efforts conducted in terms of big data analytics are investigated.
b) An architecture for big IoT data analytics is proposed.
c) Several unprecedented opportunities brought by data analytics in the IoT domain are introduced.
d) Credible use cases are presented.
e) Research challenges that remain to be addressed are identified and discussed.

These contributions are presented from Sections 3 to 6. The conclusion is provided in Section 7.

II. OVERVIEW OF IoT AND BIG DATA

An overview of IoT technologies and big data is provided before the discussion.

A. IoT

IoT offers a platform for sensors and devices to communicate seamlessly within a smart environment and enables information sharing across platforms in a convenient manner. The recent adaptation of different wireless technologies places IoT as the next revolutionary technology by benefiting from the full opportunities offered by the Internet technology. IoT has witnessed its recent adoption in smart cities with interest in developing intelligent systems, such as smart office, smart retail, smart agriculture, smart water, smart transportation, smart healthcare, and smart energy [12], [13].

IoT has emerged as a new trend in the last few years, where mobile devices, transportation facilities, public facilities, and home appliances can all be used as data acquisition equipment in IoT. All surrounding electronic equipment to facilitate daily life operations, such as wristwatches, vending machines, emergency alarms, and garage doors, as well as home appliances, such as refrigerators, microwave ovens, air conditioners, and water heaters are connected to an IoT network and can be controlled remotely. Ciufo [14] stated that these devices ‘‘talk’’ to one another and to central controlling devices. Such devices deployed in different areas may collect various kinds of data, such as geographical, astronomical, environmental, and logistical data.

A large number of communication devices in the IoT paradigm are embedded into sensor devices in the real world. Data collecting devices sense data and transmit these data using embedded communication devices. The continuum of devices and objects are interconnected through a variety of communication solutions, such as Bluetooth, WiFi, ZigBee, and GSM. These communication devices transmit data and receive commands from remotely controlled devices, which allow direct integration with the physical world through computer-based systems to improve living standards.

Over 50 billion devices ranging from smartphones, laptops, sensors, and game consoles are anticipated to be connected to the Internet through several heterogeneous access networks enabled by technologies, such as radio frequency identification (RFID) and wireless sensor networks. Reference [15] mentioned that IoT could be recognized in three paradigms: Internet-oriented, sensors, and knowledge [16]. The recent adaptation of different wireless technologies places IoT as the next revolutionary technology by benefiting from the full opportunities offered by Internet technology.

B. BIG DATA

The volume of data generated by sensors, devices, social media, health care applications, temperature sensors, and various other software applications and digital devices that continuously generate large amounts of structured, unstructured, or semi-structured data is strongly increasing. This massive data generation results in ‘‘big data’’ [17]. Traditional database systems are inefficient when storing, processing, and analyzing rapidly growing amount of data or big data [18]. The term ‘‘big data’’ has been used in the previous literature but is relatively new in business and IT [19]. An example of big data-related studies is the next frontier for innovation, competition, and productivity; McKinsey Global Institute [20] defined big data as the size of data sets that are a better database system tool than the usual tools for capturing, storing, processing, and analyzing such data [18]. ‘‘The Digital Universe’’ study [21] labels big data technologies as a new generation of technologies and architectures that aim to take out the value from a massive volume of data with various formats by enabling high-velocity capture, discovery, and analysis. This previous study also characterizes big data into three aspects: (a) data sources, (b) data analytics, and (c) the presentation of the results of the analytics.

This definition uses the 3V's (volume, variety, velocity) model proposed by Beyer [2]. The model highlights an e-commerce trend in data management that faces challenges to manage volume or size of data, variety or different sources of data, and velocity or speed of data creation. Some studies declare volume as a main characteristic of big data without providing a pure definition [22]. However, other researchers introduced additional characteristics for big data, such as veracity, value, variability, and complexity [23], [24]. The 3V's model, or its derivations, is the most common descriptions of the term ''big data.''

III. BIG DATA ANALYTICS

Big data analytics involves the processes of searching a database, mining, and analyzing data dedicated to improve company performance [25].

Big data analytics is the process of examining large data sets that contain a variety of data types [4] to reveal unseen patterns, hidden correlations, market trends, customer preferences, and other useful business information [5]. The capability to analyze large amounts of data can help an organization deal with considerable information that can affect the business [6]. Therefore, the main objective of big data analytics is to assist business associations to have improved understanding of data, and thus, make efficient and well-informed decisions. Big data analytics enables data miners and scientists to analyze a large volume of data that may not be harnessed using traditional tools [5].

Big data analytics require technologies and tools that can transform a large amount of structured, unstructured, and semi-structured data into a more understandable data and metadata format for analytical processes. The algorithms used in these analytical tools must discover patterns, trends, and correlations over a variety of time horizons in the data [26]. After analyzing the data, these tools visualize the findings in tables, graphs, and spatial charts for efficient decision making. Thus, big data analysis is a serious challenge for many applications because of data complexity and the scalability of underlying algorithms that support such processes [27].

Talia (2013) highlighted that obtaining helpful information from big data analysis is a critical matter that requires scalable analytical algorithms and techniques to return well-timed results, whereas current techniques and algorithms are inefficient to handle big data analytics. Therefore, large infrastructure and additional applications are necessary to support data parallelism. Moreover, data sources, such as high-speed data stream received from different data sources, have different formats, which makes integrating multiple sources for analytics solutions critical [28]. Hence, the challenge is focused on the performance of current algorithms used in big data analysis, which is not rising linearly with the rapid increase in computational resources [19].

Big data analytics processes consume considerable time to provide feedback and guidelines to users, whereas only a few tools [29] can process huge data sets within reasonable amount of processing time. By contrast, most of the remaining tools use the complicated trial-and-error method to deal with massive amounts of data sets and data heterogeneity [30]. Big data analytics systems exist. For example, the Exploratory Data Analysis Environment [31] is a big data visual analytics system that is used to analyze complex earth system simulations with large numbers of data sets.

A. EXISTING ANALYTICS SYSTEMS

Different analytic types are used according to the requirements of IoT applications [32]. These analytic types are discussed in this subsection under real-time, off-line, memory-level, business intelligence (BI) level, and massive level analytics categories. Moreover, a comparison based on analytics types and their levels is presented in Table 1.

Real-time analytics is typically performed on data collected from sensors. In this situation, data change constantly, and rapid data analytics techniques are required to obtain an analytical result within a short period. Consequently, two existing architectures have been proposed for real-time analysis: parallel processing clusters using traditional relational databases and memory-based computing platforms [33]. Greenplum [34] and Hana [35] are examples of real-time analytics architecture.

Off-line analytics is used when a quick response is not required [32]. For example, many Internet enterprises use Hadoop-based off-line analytics architecture to reduce the cost of data format conversion [36]. Such analytics improves data acquisition efficiency. SCRIBE [37], Kafka [38], Time-Tunnel [39], and Chukwa [40] are examples of architectures that conduct off-line analytics and can satisfy the demands of data acquisition.

Memory-level analytics is applied when the size of data is smaller than the memory of a cluster [32]. To date, the memory of clusters has reached terabyte (TB) level [41]. Therefore, several internal database technologies are required to improve analytical efficiency. Memory-level analytics is suitable for conducting real-time analysis. MongoDB [42] is an example of this architecture.

BI analytics is adopted when the size of data is larger than the memory level, but in this case, data may be imported to the BI analysis environment [43]. BI analytic currently supports TB-level data [32]. Moreover, BI can help discover strategic business opportunities from the flood of data. In addition, BI analytics allows easy interpretation of data volumes. Identifying new opportunities and implementing an effective strategy provide competitive market advantage and long-term stability.

Massive analytics is applied when the size of data is greater than the entire capacity of the BI analysis product and traditional databases [44]. Massive analytics uses the Hadoop distributed file system for data storage and map/reduce for data analysis. Massive analytics helps create the business foundation and increases market competitiveness by extracting meaningful values from data. Moreover, massive analytics

TABLE 1. Comparison of different analytics types and their levels.

Analytic Types/Level	Specified Use	Existing Architectures/Tools	Advantages/Category
Real time[33]	To analyze the large amounts of data generated by the sensors	+Greenplum +HANA	+Parallel processing clusters using traditional databases memory based computing platforms
Offline [36]	To use for the Applications where there is no high requirements on response time	+Scribe + Kafka +Timetunnel +Chukwa	+Efficient Data acquisition +Reduce the cost of data format conversion
Memory level [41]	To use where the total data volume is smaller than the maximum Memory of the cluster	+MongoDB	+Real time
Business intelligence level [43]	To use when the data scale surpasses the memory level	+Data analysis plans.	+Both offline and Online
Massive level [44]	To use when data scale is totally surpassed the capacity of business intelligence products and traditional databases	+MapReduce	+Mostly belong to Offline

obtains accurate data that leverage the risks involved in making any business decision. In addition, massive analytics provides services effectively.

B. RELATIONSHIP BETWEEN IoT AND BIG DATA ANALYTICS

Big data analytics is rapidly emerging as a key IoT initiative to improve decision making. One of the most prominent features of IoT is its analysis of information about "connected things." Big data analytics in IoT requires processing a large amount of data on the fly and storing the data in various storage technologies. Given that much of the unstructured data are gathered directly from web-enabled "things," big data implementations will necessitate performing lightning-fast analytics with large queries to allow organizations to gain rapid insights, make quick decisions, and interact with people and other devices. The interconnection of sensing and actuating devices provide the capability to share information across platforms through a unified architecture and develop a common operating picture for enabling innovative applications.

The need to adopt big data in IoT applications is compelling. These two technologies have already been recognized in the fields of IT and business. Although, the development of big data is already lagging, these technologies are inter-dependent and should be jointly developed. In general, the deployment of IoT increases the amount of data in quantity and category; hence, offering the opportunity for the application and development of big data analytics. Moreover, the application of big data technologies in IoT accelerates the research advances and business models of IoT. The relationship between IoT and big data, which is shown in Figure 1, can be divided into three steps to enable the management of IoT data. The first step comprises managing IoT data sources, where connected sensors devices use applications to interact with one another. For example, the interaction of devices such as CCTV cameras, smart traffic lights, and smart home devices, generates large amounts of data sources with different formats. This data can be stored in low cost commodity storage on the cloud. In the second step, the generated data are called "big data," which are based on their volume, velocity, and variety. These huge amounts of data are stored in big data files in shared distributed fault-tolerant databases. The last step applies analytics tools such as MapReduce, Spark, Splunk, and Skytree that can analyze the stored big IoT data sets. The four levels of analytics start from training data, then move on to analytics tools, queries, and reports.

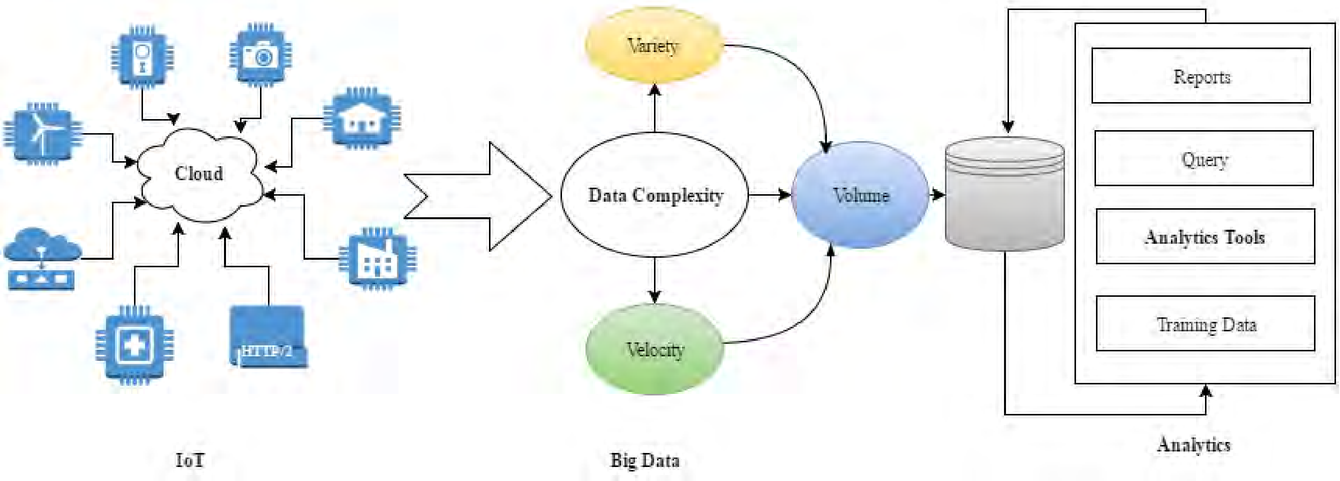

FIGURE 1. Relationship between IoT and big data analytics.

C. BIG DATA ANALYTICS METHODS

Big data analytics aim to immediately extract knowledgeable information that helps in making predictions, identifying recent trends, finding hidden information, and ultimately, making decisions [7]. Data mining techniques are widely deployed for both problem-specific methods and generalized data analytics. Accordingly, statistical and machine learning methods are utilized. The evolution of big data also changes analytics requirements. Although the requirements for efficient mechanisms lie in all aspects of big data management [30], such as capturing, storage, preprocessing, and analysis; for our discussion, big data analytics requires the same or faster processing speed than traditional data analytics with minimum cost for high-volume, high-velocity, and high-variety data [45].

Various solutions are available for big data analytics, and advancements in developing and improving these solutions are being continuously achieved to make them suitable for new big data trends. Data mining plays an important role in analytics, and most of the techniques are developed using data mining algorithms according to a particular scenario. Knowledge on available big data analytics options is crucial when evaluating and choosing an appropriate approach for decision making. In this section, we present several methods that can be implemented for several big data case studies. Some of these analytics methods are efficient for big IoT data analytics. Diverse and tremendous size data sets contribute more in big data insights. However, this belief is not always valid because more data may have more ambiguities and abnormalities [7].

We present big data analytics methods under classification, clustering, association rule mining, and prediction categories. Figure 2 depicts and summarizes each of these categories. Each category is a data mining function and involves many methods and algorithms to fulfill information extraction and analysis requirements. For example, Bayesian network, support vector machine (SVM), and *k*-nearest neighbor (KNN) offer classification methods. Similarly, partitioning, hierarchical clustering, and co-occurrence are widespread in clustering. Association rule mining and prediction comprise significant methods.

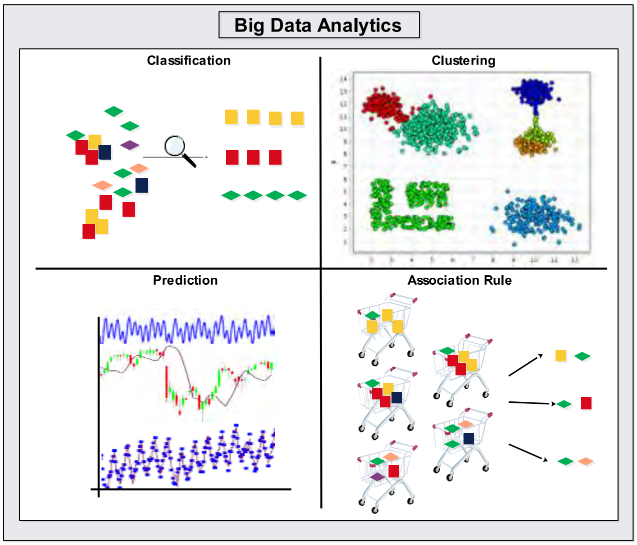

FIGURE 2. Overview of big data analytics methods.

Classification is a supervised learning approach that uses prior knowledge as training data to classify data objects into groups [46]. A predefined category is assigned to an object, and thus, the objective of predicting a group or class for an object is achieved (see Figure 2). Finding unknown or hidden patterns is more challenging for big IoT data. Furthermore, extracting valuable information from large data sets to improve decision making is a critical task. A Bayesian network is a classification method that offers model interpretability. Bayesian networks are efficient for analyzing complex data structures revealed through big data rather than traditional structured data formats. These networks are directed acyclic graphs, where nodes are random variables and edges denote conditional dependency [47]. Naïve, selective naïve, semi-naïve Bayes, and Bayes multi-nets are the proposed categories for classification [48].

Analyzing data patterns and creating groups are efficiently performed using SVM, which is also classification approach for big data analytics. SVM utilizes statistical learning theory to analyze data patterns and create groups. Several applications of SVM classification in big data analytics include text classification [49], pattern matching [50], health diagnostics [51], and commerce. Similarly, KNN is typically designed to provide efficient mechanisms for finding hidden patterns from big data sets, such that retrieved objects are similar to the predefined category [52]. Using cases further improve the KNN algorithm for application in anomaly detection [53], high-dimensional data [54], and scientific experiments [55]. Classification has other extensions while adopting a large number of artificial intelligence and data mining techniques. Consequently, classification is one of the widespread data mining techniques for big data analytics.

Clustering is another data mining technique used as a big data analytics method. Contrary to classification, clustering uses an unsupervised learning approach and creates groups for given objects based on their distinctive meaningful features [56]. As we have presented in Figure 2 that grouping a large number of objects in the form of clusters makes data manipulation simple. The well-known methods used for clustering are hierarchical clustering and partitioning. The hierarchical clustering approach keeps combining small clusters of data objects to form a hierarchical tree and create agglomerative clusters. Divisive clusters are created in the opposite manner by dividing a single cluster that contains all data objects into smaller appropriate clusters [57].

Market analysis and business decision making are the most significant applications of big data analytics. The process of association rule mining involves identifying interesting relationships among different objects, events, or other entities to analyze market trends, consumer buying behavior, and product demand predictions (see Figure 2). Association rule mining [58] focuses on identifying and creating rules based on the frequency of occurrences for numeric and non-numeric data. Data processing is performed in two manners under association rules. First, sequential data processing uses priori-based algorithms, such as MSPS [59] and LAPIN-SPAM [60], to identify interaction associations. Another significant data processing approach under association rule is temporal sequence analysis, which uses algorithms to analyze event patterns in continuous data.

Predictive analytics use historical data, which are known as training data, to determine the results as trends or behavior in data. SVM and fuzzy logic algorithms are used to identify relationships between independent and dependent variables and to obtain regression curves for predictions, such as for natural disasters. Furthermore, customer buying predictions and social media trends are analyzed through predictive analytics [61] (see Table 2). In the case of big data analytics, processing requirements are modified according to the nature and volume of data. Fast data access and mining methods for structured and unstructured data are major concerns related to big data analytics. Furthermore, data representation is a significant requirement in big data analytics. Time series analysis reduces high dimensionality associated with big data and offers representation for improved decision making. Research related to time series representation includes ARMA [62], bitmaps [63], and wavelet functions [64].

The big data analytics methods discussed in this section are widely adopted in many application areas of big data,

TABLE 2. Applications of big data mining for IoT.

Method	Applications										
	Disaster management	Healthcare	Medical Imaging	Human Genetics	Market Analysis	Industry	Speech Recognition	Bioinformatics	NLP	Social Network Analysis	e-governance
Classification [46]	-	-	✓	-	-	✓	✓	-	✓	-	✓
Clustering [57]	-	✓	✓	✓	✓	✓	-	✓	-	✓	✓
Association rule[58, 65]	-	✓	-	-	✓	✓	-	✓	-	-	✓
Prediction [61]	✓	-	-	-	✓	-	-	-	-	✓	-
Time Series [62] [63] [64]	✓	-	✓	-	-	-	✓	-	-	✓	✓

✓ has support
- not obvious

such as disaster management, healthcare, business, industry, and e-governance. In Table 2, we present the application areas of big data mining functionalities that are elaborated in this section, '✓' is used to show the support for an application whereas '-' denotes that it is not obvious whether the method supports to an application or not. In particular, Table 2 shows that classification methods are suitable for medical imaging, industry, speech recognition, natural language processing, and e-governance. Clustering and association rule-based data analytics methods are applicable to industry and e-governance and are well adopted in healthcare, e-commerce, and bioinformatics. Predictive analytics are useful for disaster and market predictions, whereas time series analysis is used in disaster forecasting, medical imaging, speech recognition, social network analysis, and e-governance.

D. IoT ARCHITECTURE FOR BIG DATA ANALYTICS

The architectural concept of IoT has several definitions based on IoT domain abstraction and identification. It offers a reference model that defines relationships among various IoT verticals, such as, smart traffic, smart home, smart transportation, and smart health. The architecture for big data analytics offers a design for data abstraction. Furthermore, this standard provides a reference architecture that builds upon the reference model. Many IoT architectures are found in the literature [13], [66], [67]. For example, [13] offered an IoT architecture with cloud computing at the center and a model of end-to-end interaction among various stakeholders in a cloud-centric IoT framework for better comparison with the proposed IoT architecture. This architecture is achieved by seamless ubiquitous sensing, data analytics, and information representation with IoT as the unifying architecture. However, the current architecture focuses on IoT with regard to communications. To our knowledge, our proposed architecture, which integrates IoT and big data analytics, has not been studied in the current literature. Figure 3 illustrates the IoT architecture and big data analytics. In this figure, the sensor layer contains all the sensor devices and the objects, which are connected through a wireless network. This wireless network communication can be RFID, WiFi, ultra-wideband, ZigBee, and Bluetooth. The IoT gateway allows

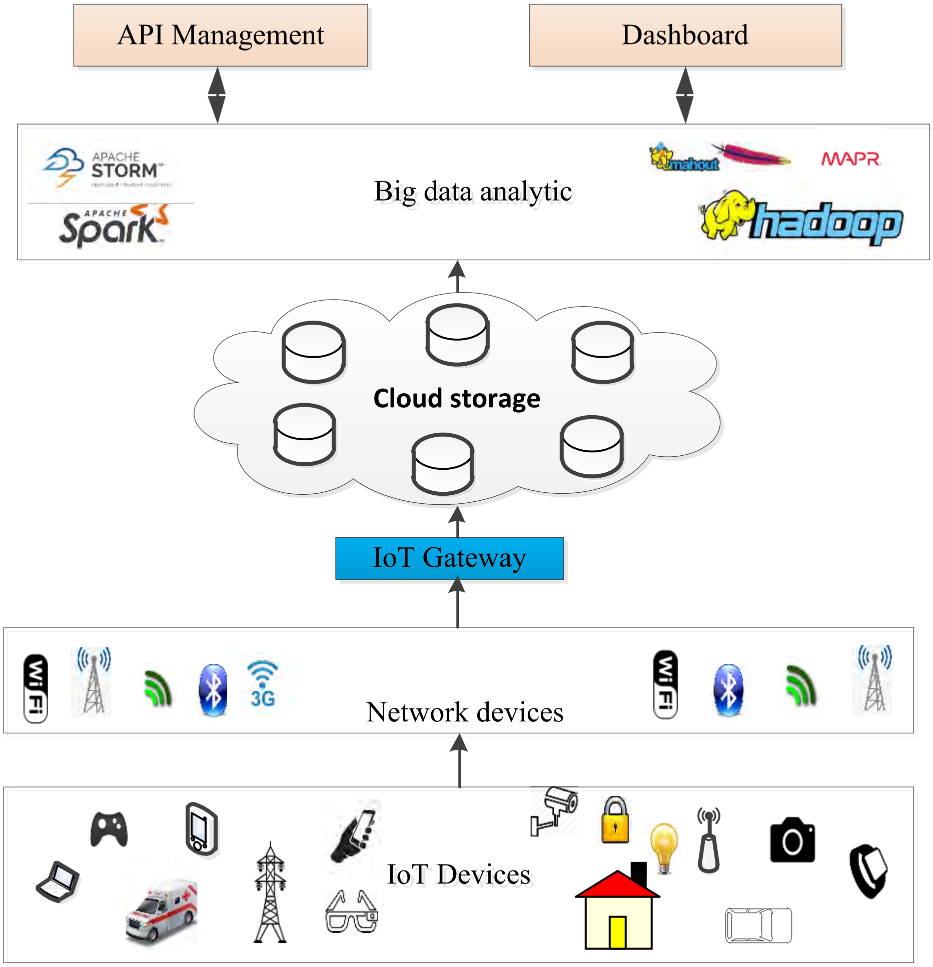

FIGURE 3. IoT architecture and big data analytics.

communication of the Internet and various webs. The upper layer concerns big data analytics, where a large amount of data received from sensors are stored in the cloud and accessed through big data analytics applications. These applications contain API management and a dashboard to help in the interaction with the processing engine.

A novel meta-model-based approach for integrating IoT architecture objects is proposed. The concept is semi-automatically federated into a holistic digital enterprise architecture environment. The main objective is to provide an adequate decision support for complex business, architecture management with the development of assessment systems, and IT environment. Thus, architectural decisions for IoT are closely connected with code implementation to allow users to understand the integration of enterprise architecture management with IoT.

IV. USE CASES

This section presents a number of use cases for big IoT data analytics. Although the use cases are relevant to IoT applications, the choices have been guided for the ones that are most commonly used in IoT applications and for the amount of data that can be generated for analytics.

A. SMART METERING

Smart metering is one of the IoT application use cases that generates a large amount of data from different sources, such as smart grids, tank levels, and water flows, and silos stock calculation, in which processing takes a long time even on a dedicated and powerful machine [68]. A smart meter is a device that electronically records consumption of electric energy data between the meter and the control system. Collecting and analyzing smart meter data in IoT environment assist the decision maker in predicting electricity consumption. Furthermore, the analytics of a smart meter can also be used to forecast demands to prevent crises and satisfy strategic objectives through specific pricing plans. Thus, utility companies must be capable of high-volume data management and advanced analytics designed to transform data into actionable insights.

B. SMART TRANSPORTATION

A smart transportation system is an IoT-based use case that aims to support the smart city concept. A smart transportation system intends to deploy powerful and advanced communication technologies for the management of smart cities. Traditional transportation systems, which are based on image processing, are affected by weather conditions, such as heavy rains and thick fog. Consequently, the captured image may not be clearly visible. The design of an e-plate system [69] using RFID technology provides a good solution for intelligent monitoring, tracking, and identification of vehicles. Moreover, introducing IoT into vehicular technologies will enable traffic congestion management to exhibit significantly better performance than the existing infrastructure. This technology can improve existing traffic systems in which vehicles can effectively communicate with one another in a systematic manner without human intervention.

Satellite navigation systems and sensors can also be applied in trucks, ships, and airplanes in real time. The routing of these vehicles can be optimized by using the bulk of available public data, such as traffic jams, road conditions, delivery addresses, weather conditions, and locations of refilling stations. For example, in case of runtime address change, the updated information (route, cost) can be optimized, recalculated, and passed on to drivers in real time. Sensors incorporated into these vehicles can also provide real-time information to measure engine health, determine whether equipment requires maintenance, and predict errors [70].

C. SMART SUPPLY CHAINS

Embedded sensor technologies can communicate bidirectionally and provide remote accessibility to over 1 million elevators worldwide [71]. The captured data are used by on- and off-site technicians to run diagnostics and repair options to make appropriate decisions, which result in increased machine uptime and enhanced customer service. Ultimately, big IoT data analytics allows a supply chain to execute decisions and control the external environment. IoT-enabled factory equipment will be able to communicate within data parameters (i.e., machine utilization, temperature) and optimize performance by changing equipment settings or process workflow [72]. In-transit visibility is another use case that will play a vital role in future supply chains in the presence of IoT infrastructure. Key technologies used by in-transit visibility are RFIDs and cloud-based Global Positioning System (GPS), which provide location, identity, and other tracking information. These data will be the backbone of supply chains supported by IoT technologies. The information gathered by equipment will provide detailed visibility of an item shipped from a manufacturer to a retailer. Data collected via RFID and GPS technologies will allow supply chain managers to enhance automated shipment and accurate delivery information by predicting time of arrival. Similarly, managers will be able to monitor other information, such as temperature control, which can affect the quality of in-transit products.

D. SMART AGRICULTURE

Smart agriculture is a beneficial use case in big IoT data analytics. Sensors are the actors in the smart agriculture use case. They are installed in fields to obtain data on moisture level of soil, trunk diameter of plants, microclimate condition, and humidity level, as well as to forecast weather. Sensors transmit obtained data using network and communication devices. These data pass through an IoT gateway and the Internet to reach the analytics layer shown in Table 1. The analytics layer processes the data obtained from the sensor network to issue commands. Automatic climate control according to harvesting requirements, timely and controlled irrigation, and humidity control for fungus prevention are

TABLE 3. Comparison of IoT big data analytics use cases.

Use cases	Benefits	IoT devices	Data source	Big data analytics applications
Smart metering [68]	Predict electricity consumption	Sensors	Text	Hadoop
Smart transportation[69] [70]	Improve existing traffic system by which vehicles can effectively communicate with one another in a systematic manner without human intervention	Sensors, cameras	Text, video, audio	Hadoop, Spark, Hive
Smart supply chains [71] [72]	Allow a supply chain to execute decisions and control the external environment	Sensors, mobile devices	Text, image	Hadoop
Smart agriculture [12, 13]	Obtain moisture level of soil, trunk diameter of plants, microclimate condition and humidity level; forecast weather	Sensors	Text, image	Hadoop
Smart grid [73, 74] [75] [76]	Improves reliability, safety, and efficiency, along with real-time control and monitoring	Sensors	Text	Hadoop
Smart traffic [77]	Detect the presence of vehicles, bikers, and pedestrians	Cameras	Video, image	Hadoop, Spark

examples of actions performed based on big data analytics recommendations.

E. SMART GRID

The smart grid is a new generation of power grid in which managing and distributing electricity between suppliers and consumers is upgraded using two-way communication technologies and computing capabilities to improve reliability, safety, efficiency with real-time control, and monitoring [73], [74]. One of the major challenges in a power system is integrating renewable and decentralized energy. Electricity systems require a smart grid to manage the volatile behavior of distributed energy resources (DERs) [75]. However, most energy systems have to follow governmental laws and regulations, as well as consider business analysis and potential legal constraints [76]. Grid sensors and devices continuously and rapidly generate data related to control loops and protection and require real-time processing and analytics along with machine-to-machine (M2M) or human-to-machine (HMI) interactions to issue control commands to the system. However, the system must fulfill visualization and reporting requirements.

F. SMART TRAFFIC LIGHT SYSTEM

The smart traffic light system consists of nodes that locally interact with IoT sensors and devices to detect the presence of vehicles, bikers, and pedestrians. These nodes communicate with neighboring traffic lights to measure the speed and distance of approaching transportation means and manage green traffic signals [77]. IoT data gathered using the system require real-time analytics processing to perform necessary tasks, such as changing the timing cycles according to traffic conditions, sending informative signals to neighboring nodes, and detecting approaching vehicles that use IoT sensors and devices to prevent long queues or accidents. Moreover, smart traffic light systems can send their collected IoT data to cloud storage for further analytics. Table 3 presents the use cases of IoT big data analytics.

As shown in Table 3, most use cases are related to M2M communication technologies and decrease the role of human interaction. However, the technologies use prediction methods and decision-making techniques to improve real-time control, monitoring, and performance. Textual data are among the common data types generated by IoT devices, which are mostly sensors and cameras. Text-based data are suitable for analysis by distributed file systems, such as Hadoop.

V. OPPORTUNITIES

IoT is currently considered one of the most profound transitions in technology. Current IoT provides several data analytics opportunities for big data analytics. Figure 4 shows the examples of use cases and opportunities discussed in Sections 4 and 5.

A. E-COMMERCE

Big IoT data analytics offers well-designed tools to process real-time big data, which produce timely results for decision making. Big IoT data exhibit heterogeneity, increasing volume, and real-time data processing features. The convergence of big data with IoT brings new challenges and opportunities to build a smart environment. Big IoT data analytics has widespread applications in nearly every industry. However, the main success areas of analytics are in e-commerce, revenue growth, increased customer size, accuracy of sale forecast results, product optimization, risk management, and improved customer segmentation.

B. SMART CITIES

Big data collected from smart cities offer new opportunities in which efficiency gains can be achieved through an appropriate analytics platform/infrastructure to analyze big

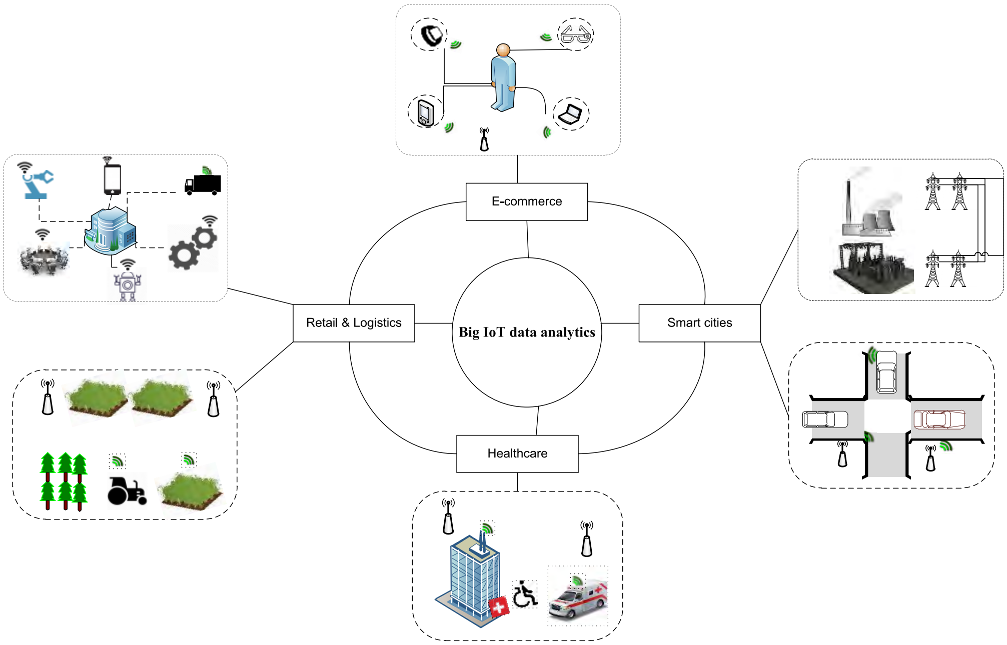

FIGURE 4. Example of use cases and opportunities for big IoT data analytics architecture.

IoT data. Various devices connect to the Internet in a smart environment and share information. Moreover, the cost of storing data has been reduced dramatically after the invention of cloud computing technology. Analysis capabilities have made huge leaps. Thus, the role of big data in a smart city can potentially transform every sector of the economy of a nation. Hadoop with YARN resource manager has offered recent advancement in big data technology to support and handle numerous workloads, real-time processing, and streaming data ingestion.

C. RETAIL AND LOGISTICS

IoT is expected to play a key role as an emerging technology in the area of retail and logistics. In logistics, RFID keeps track of containers, pallets, and crates. In addition, considerable advancements in IoT technologies can facilitate retailers by providing several benefits. However, IoT devices generate large amounts of data on a daily basis. Thus, powerful data analytics enables enterprises to gain insights from the voluminous amounts of data produced through IoT technologies. Applying data analytics to logistic data sets can improve the shipment experience of customers. Moreover, retail companies can earn additional profit by analyzing customer data, which can predict the trends and demands of goods. By looking into customer data, optimizing pricing plans and seasonal promotions can be planned efficiently to maximize profit.

D. HEALTHCARE

Recent years have witnessed tremendous growth in smart health monitoring devices. These devices generate enormous amounts of data. Thus, applying data analytics to data collected from fetal monitors, electrocardiograms, temperature monitors, or blood glucose level monitors can help healthcare specialists efficiently assess the physical conditions of patients. Moreover, data analytics enables healthcare professionals to diagnose serious diseases in their early stages to help save lives. Furthermore, data analytics improves the clinical quality of care and ensures the safety of patients. In addition, physician profile can be reviewed by looking into the history of treatment of patients, which can improve customer satisfaction, acquisition, and retention.

VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

IoT and big data analytics have been extensively accepted by many organizations. However, these technologies are still in their early stages. Several existing research challenges have not yet been addressed. This section presents several challenges in the field of big IoT data analytics.

A. PRIVACY

Privacy issues arise when a system is compromised to infer or restore personal information using big data analytics tools, although data are generated from anonymous users. With the proliferation of big data analytics technologies used in big IoT data, the privacy issue has become a core problem in the data mining domain. Consequently, most people are reluctant to rely on these systems, which do not provide solid service-level agreement (SLA) conditions regarding user personal information theft or misuse. In fact, the sensitive information of users has to be secured and protected from external interference. Although temporary identification, anonymity, and encryptions provide several ways to enforce data privacy, decisions have to be made with regard to ethical factors, such as what to use, how to use, and why use generated big IoT data [7].

Another security risk associated with IoT data is the heterogeneity of the types of devices used and the nature of generated data, such as raw devices, data types, and communication protocols. These devices can have different sizes and shapes outside the network and are designed to communicate with cooperative applications. Thus, to authenticate these devices, an IoT system should assign a non-repudiable identification system to each device. Moreover, enterprises should maintain a meta-repository of these connected devices for auditing purposes. This heterogeneous IoT architecture is new to security professionals, and thus, results in increased security risks. Consequently, any attack in this scenario compromises system security and disconnects interconnected devices.

In the context of big IoT data, security and privacy are the key challenges in processing and storing huge amounts of data. Moreover, to perform critical operations and host private data, these systems highly rely on third party services and infrastructure. Therefore, an exponential growth in data rate causes difficulty in securing each and every portion of critical data. As previously discussed, existing security solutions (Karim, 2016 #86) are no long applicable to providing complete security in big IoT data scenarios. Existing algorithms are not designed for the dynamic observation of data, and thus, are not effectively applied. Legacy data security solutions are specifically designed for static data sets, whereas current data requirements are changing dynamically (Lafuente, 2015). Thus, deploying these security solutions is difficult for dynamically increasing data. In addition, legislative and regulatory issues should be considered while signing SLAs.

With regard to data generated through IoT, the following security problems can emerge [78]: (a) timely updates - difficulty in keeping systems up to date, (b) incident management - identifying suspicious traffic patterns among legitimate ones and possible failure to capture unidentifiable incidents, (c) interoperability - proprietary and vendor-specific procedures will pose difficulties in finding hidden or zero day attacks, (d) and protocol convergence - although IPv6 is currently compatible with the latest specifications, this protocol has yet to be fully deployed. Therefore, the application of security rules over IPv4 may not be applicable to protecting IPv6.

At present, no answer can address these challenges and manage the security and privacy of interconnected devices. However, the following guidelines can overcome these adversities. (a) First, a true open ecosystem with standard APIs is necessary to avoid interoperability and reliability problems. (b) Second, devices must be well protected while communicating with peers. (c) Third, devices should be hardcoded with the best security practices to protect against common security and privacy threats.

B. DATA MINING

Data mining methods provide efficient and best-fitting predictive or descriptive solutions for big data that can also be generalized for new data [45]. The evolution of big IoT data and cloud computing platforms has brought the challenges of data exploration and information extraction [79]. However, for the overall big IoT data architecture, Figure 5 presents the primary challenges related to processing and data mining.

Exhaustive data reads/writes: The high-volume, high-velocity, and high-variety qualities of big IoT data challenge exploration, integration, heterogeneous communication, and extraction processes. The size and heterogeneity of data impose new data mining requirements, and diversity in data sources also poses a challenge [80]–[82]. Furthermore, compared with small data sets, large data sets comprise more abnormalities and ambiguities that require additional preprocessing steps, such as cleansing, reduction, and transmission [23], [83]. Another issue lies in the extraction of exact and knowledgeable information from the large volumes of diverse data. Consequently, obtaining accurate information from complex data requires analyzing data properties and finding association among different data points.

Researchers have introduced parallel and sequential programming models and proposed different algorithms to minimize query response time while dealing with big data.

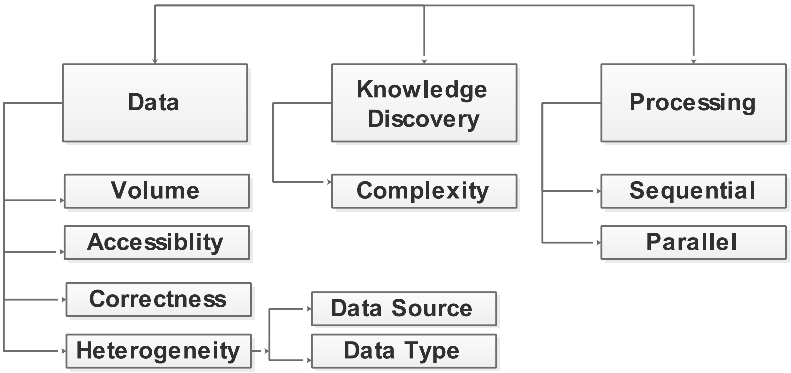

FIGURE 5. Big data mining issues in IoT.

Moreover, researchers have selected existing data mining algorithms in different manners to (a) improve single source knowledge discovery, (b) implement data mining methods for multi-source platforms, and (c) study and analyze dynamic data mining methods and stream data [84]. Hence, parallel k-means algorithm [85] and parallel association rule mining methods [65] are introduced. However, the need to devise algorithms remains to provide compatibility with the latest parallel architectures. Moreover, synchronization issues may occur in parallel computing, while information is exchanged within different data mining methods. This bottleneck of data mining methods has become an open issue in big IoT data analytics that should be addressed.

C. VISUALIZATION

Visualization is an important entity in big data analytics, particularly when dealing with IoT systems where data are generated enormously. Furthermore, conducting data visualization is difficult because of the large size and high dimension of big data. This situation shows underlying trends and a complete picture of parsed data. Therefore, big data analytics and visualization should work seamlessly to obtain the best results from IoT applications in big data. However, visualization in the case of heterogeneous and diverse data (unstructured, structured, and semi-structured) is a challenging task. Designing visualization solution that is compatible with advanced big data indexing frameworks is a difficult task. Similarly, response time is a desirable factor in big IoT data analytics. Consequently, cloud computing architectures supported with rich GUI facilities can be deployed to obtain better insights into big IoT data trends [86].

Different dimensionality reduction methods have been introduced as a result of complex and high-dimensional big IoT data [87], [88]. However, these methods are unsuitable for all types of presented data. Similarly, when fine-grained dimensions are visualized effectively, the probability to identify observable correlations, patterns, and outliners is high [89]. Moreover, data should be kept locally to obtain usable information efficiently because of power and bandwidth constraints. In addition, visualization software should run with the concept of reference locality to achieve efficient outcome in an IoT environment. Given that the amount of big IoT data is increasing rapidly, the requirement of enormous parallelization is a challenging task in visualization. Thus, to decompose a problem into manageable independent tasks to enforce concurrent execution of queries is a challenge for parallel visualization algorithms [90].

At present, most big data visualization tools used for IoT exhibit poor performance results in terms of functionality, scalability, and response time. To provide effective uncertainty-aware visualization during the visual analytics process, avoiding uncertainty imposes a considerable challenge [32]. Furthermore, several important issues are addressed [91], such as (a) visual noise - most data set objects are closely related to one another, and thus, users may perceive different results of the same type; (b) information loss - applying reduction methods to visible data sets can cause information loss; (c) large image observation - data visualization tools have inherent problems with respect to aspect ratio, devise resolution, and physical perception limits; (d) frequently changing image - users will not notice rapid data changes in an output; and (e) high performance requirements - high performance requirements are imposed because data are generated dynamically in an IoT environment. Moreover, methods supported by advanced analytics enable interactive graphics on laptops, desktops, or mobile devices, such as smartphones and tablets [92].

Real-time analytics is another consideration highlighted in IoT architectures. Several guidelines on visualization in big data are presented [93], such as (a) data awareness, i.e., appropriate domain expertise, (b) data quality - cleaning data using information management or data governance policies, (c) meaningful results - data clustering is used to provide high-level abstraction such that the visibility of smaller groups of data is possible, and (d) outliers should be removed from the data or treated as a separate entity. Reference [94] suggested that visualization should adhere to the following guidelines: (a) the system should provide special attention to metadata, (b) visualization software should be interactive and should require maximum user involvement, and (c) tools should be built based on the dynamic nature of the generated data.

D. INTEGRATION

Integration refers to having a uniform view of different formats. Data integration provides a single view of the data arriving from different sources and combines the view of data [95]. Data integration includes all processes involved in collecting data from different sources, as well as in storing and providing data with a unified view. For each moment, different forms of data are continuously generated by social media, IoT, and other communication and telecommunication approaches. The produced data can be categorized into three groups: (a) structured data, such as data stored in traditional database systems, including tables with rows and columns; (b) semi-structured, such as HTML, XML, and Json files; and (c) unstructured data, such as videos, audios, and images. Good data offer good information; however, this relationship is only achieved through data integration [96]. Integrating diverse data types is a complex task in merging different systems or applications [97]. Overlapping the same data, increasing performance and scalability, and enabling real-time data access are among the challenges associated with data integration that should be addressed in the future.

Another challenge is to adjust structures in semi-structured and unstructured data before integrating and analyzing these types of data [98]. Information, such as entities and relationships, can be extracted from textual data by using available technologies in the eras of text mining, machine learning, natural processing, and information extraction. However, new technologies should be developed to extract images, videos, and other information from other non-text formats of unstructured data [98]. Text mining is expected to be conducted by applying several specialized extractors on the same text. Hence, managing and integrating different extraction results from a certain data source require other techniques [99].

VII. CONCLUSION

The growth rate of data production has increased drastically over the past years with the proliferation of smart and sensor devices. The interaction between IoT and big data is currently at a stage where processing, transforming, and analyzing large amounts of data at a high frequency are necessary. We conducted this survey in the context of big IoT data analytics. First, we explored recent analytics solutions. The relationship between big data analytics and IoT was also discussed. Moreover, we proposed an architecture for big IoT data analytics. Furthermore, big data analytics types, methods, and technologies for big data mining were presented. Some credible use cases were also provided. In addition, we explored the domain by discussing various opportunities brought about by data analytics in the IoT paradigm. Several open research challenges were discussed as future research directions. Finally, we concluded that existing big IoT data analytics solutions remained in their early stages of development. In the future, real-time analytics solution that can provide quick insights will be required.

REFERENCES

[1] P. Tiainen, ''New opportunities in electrical engineering as a result of the emergence of the Internet of Things,'' Tech. Rep., AaltoDoc, Aalto Univ., 2016.

[2] M. Beyer, ''Gartner says solving 'Big Data' challenge involves more than just managing volumes of data,'' Tech. Rep., AaltoDoc, Aalto Univ., 2011.

[3] J. Gantz and D. Reinsel, ''Extracting value from chaos,'' *IDC Iview*, vol. 1142, pp. 1–12, Jun. 2011.

[4] R. Mital, J. Coughlin, and M. Canaday, ''Using big data technologies and analytics to predict sensor anomalies,'' in *Proc. Adv. Maui Opt. Space Surveill. Technol. Conf.*, Sep. 2014, p. 84.

[5] N. Golchha, ''Big data-the information revolution,'' *Int. J. Adv. Res.*, vol. 1, no. 12, pp. 791–794, 2015.

[6] P. Russom, *Big Data Analytics*. TDWI, 4th Quart., 2011.

[7] C.-W. Tsai, ''Big data analytics: A survey,'' *J. Big Data*, vol. 2, no. 1, pp. 1–32, 2015.

[8] M. Chen, *Related Technologies in Big Data*. Heidelberg, Germany: Springer, 2014, pp. 11–18.

[9] Z. Khan, A. Anjum, and S. L. Kiani, ''Cloud based big data analytics for smart future cities,'' in *Proc. IEEE/ACM 6th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2013, pp. 381–386.

[10] P. Russom, *Big Data Analytics*. TDWI, 4th Quart., 2011, pp. 1–35.

[11] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, ''Big data, analytics and the path from insights to value,'' *MIT Sloan Manag. Rev.*, vol. 52, no. 2, p. 21, 2011.

[12] E. Al Nuaimi *et al.*, ''Applications of big data to smart cities,'' *J. Internet Services Appl.*, vol. 6, p. 25, Dec. 2015.

[13] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, ''Internet of Things (IoT): A vision, architectural elements, and future directions,'' *Future Generat. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.

[14] C. A. Ciufo. (2014). *Industrial Equipment Talking on the IoT? Better get a Gateway (Device)*. [Online]. Available: http://eecatalog.com/caciufo/2014/07/15/iot-gateway-adlink/

[15] L. Atzori, A. Iera, and G. Morabito, ''The Internet of Things: A survey,'' *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, 2010.

[16] H.-C. Hsieh and C.-H. Lai, ''Internet of Things architecture based on integrated PLC and 3G communication networks,'' in *Proc. IEEE 17th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2011, pp. 853–856.

[17] K. Kambatla, ''Trends in big data analytics,'' *J. Parallel Distrib. Comput.*, vol. 74, no. 7, pp. 2561–2573, 2014.

[18] J. Manyika, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Inst. Rep., 2011.

[19] I. A. T. Hashem *et al.*, ''The rise of 'big data' on cloud computing: Review and open research issues,'' *Inf. Syst.*, vol. 47, pp. 98–115, Jan. 2015.

[20] W. B. Ali, ''Big data-driven smart policing: big data-based patrol car dispatching,'' *J. Geotech. Transp. Eng.*, vol. 1, no. 2, pp. 1–16, 2016.

[21] J. Gantz and D. Reinsel, ''The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,'' IDC iView: IDC Analyze the Future, Tech. Rep., 2012.

[22] V. Borkar, M. J. Carey, and C. Li, ''Inside Big Data management: Ogres, onions, or parfaits?'' in *Proc. 15th Int. Conf. Extending Database Technol.*, 2012, pp. 3–14.

[23] A. Gani, ''A survey on indexing techniques for big data: Taxonomy and performance evaluation,'' *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 241–284, 2016.

[24] A. Paul, ''Video search and indexing with reinforcement agent for interactive multimedia services,'' *ACM Trans. Embed. Comput. Syst.*, vol. 12, no. 2, pp. 1–16, 2013.

[25] O. Kwon and N. B. L. Shin, ''Data quality management, data usage experience and acquisition intention of big data analytics,'' *Int. J. Inf. Manage.*, vol. 34, no. 3, pp. 387–394, 2014.

[26] S. Oswal and S. Koul, ''Big data analytic and visualization on mobile devices,'' in *Proc. Nat. Conf. New Horizons IT-NCNHIT*, 2013, p. 223.

[27] L. Candela and D. P. C. Pagano, ''Managing big data through hybrid data infrastructures,'' *ERCIM News*, vol. 89, pp. 37–38, Jun. 2012.

[28] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya. (2013). ''Big data computing and clouds: Challenges, solutions, and future directions.'' [Online]. Available: https://arxiv.org/abs/1312.4722

[29] D. Singh and C. K. Reddy, ''A survey on platforms for big data analytics,'' *J. Big Data*, vol. 2, no. 1, p. 1, 2014.

[30] A. Siddiqa, ''A survey of big data management: Taxonomy and state-of-the-art,'' *J. Netw. Comput. Appl.*, vol. 71, pp. 151–166, Aug. 2016.

[31] C. A. Steed, ''Big data visual analytics for exploratory earth system simulation analysis,'' *Comput. Geosci.*, vol. 61, pp. 71–82, Mar. 2013.

[32] C. L. P. Chen and C.-Y. Zhang, ''Data-intensive applications, challenges, techniques and technologies: A survey on big data,'' *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[33] M. W. Pfaffl, ''A new mathematical model for relative quantification in real-time RT-PCR,'' *Nucl. Acids Res.*, vol. 29, no. 9, p. e45, 2001.

[34] F. M. Waas, ''Beyond conventional data warehousing-massively parallel data processing with Greenplum database,'' in *International Workshop on Business Intelligence for the Real-Time Enterprise*. Berlin, Germany: Springer, 2008.

[35] F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner, ''SAP HANA database: Data management for modern business applications,'' *ACM SIGMOD Rec.*, vol. 40, no. 4, pp. 45–51, Dec. 2011.

[36] M. Cheng, ''Mu rhythm-based cursor control: An offline analysis,'' *Clin. Neurophys.*, vol. 115, no. 4, pp. 745–751, 2004.

[37] M. Castro, P. Druschel, A. M. Kermarrec, and A. I. T. Rowstron, ''Scribe: A large-scale and decentralized application-level multicast infrastructure,'' *IEEE J. Sel. Areas Commun.*, vol. 20, no. 8, pp. 1489–1499, Oct. 2002.

[38] J. Kreps, N. Narkhede, and J. Rao, ''KAFKA: A distributed messaging system for log processing,'' in *Proc. NetDB*. 2011, pp. 1–7.

[39] H. Notsu, Y. Okada, M. Akaishi, and K. Niijima, ''Time-tunnel: Visual analysis tool for time-series numerical data and its extension toward parallel coordinates,'' in *Proc. Int. Conf. Comput. Graph., Imag. Vis. (CGIV)*, Jul. 2005, pp. 167–172.

[40] A. Rabkin and R. H. Katz, ''Chukwa: A system for reliable large-scale log collection,'' in *Proc. LISA*, Nov. 2010, pp. 1–15.

[41] S. Hong and H. Kim, ''An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness,'' *ACM SIGARCH Comput. Archit.*, vol. 37, no. 3, pp. 152–163, Jun. 2009.

[42] K. Chodorow, *MongoDB: The Definitive Guide*. Newton, MA, USA: O'Reilly Media, Inc, 2014.

[43] Z. Jourdan and R. K. T. E. Rainer Marshall, ''Business intelligence: An analysis of the literature 1,'' *Inf. Syst. Manage.*, vol. 25, no. 2, pp. 121–131, 2008.

[44] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, May 2010.

[45] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "A survey of multiobjective evolutionary algorithms for data mining: Part I," *IEEE Trans. Evol. Comput.*, vol. 18, no. 1, pp. 4–19, Feb. 2014.

[46] V. Estivill-Castro, "Why so many clustering algorithms: A position paper," *ACM SIGKDD Explorations Newslett.*, vol. 4, no. 1, pp. 65–75, 2002.

[47] C. Bielza and P. Larrañaga, "Discrete Bayesian network classifiers: A survey," *ACM Comput. Surveys*, vol. 47, no. 1, p. 5, Jul. 2014.

[48] F. Chen *et al.*, "Data mining for the Internet of Things: Literature review and challenges," *Int. J. Distrib. Sensor Netw.*, vol. 12, Aug. 2015.

[49] R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification," *Quant. Finance*, vol. 15, no. 6, pp. 999–1012, 2015.

[50] P. Melin and O. Castillo, "A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition," *Appl. Soft comput.*, vol. 21, pp. 568–577, Aug. 2014.

[51] A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing health monitoring based on Hilbert-Huang transform, support vector machine, and regression," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 52–62, Jan. 2015.

[52] D. T. Larose, *K-Nearest Neighbor Algorithm*. 2005, pp. 90–106.

[53] M.-Y. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted K-nearest-neighbor classifiers," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3492–3498, 2011.

[54] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, Nov. 2014.

[55] C. Hu, "Data-driven method based on particle swarm optimization and K-nearest neighbor regression for estimating capacity of lithium-ion battery," *Appl. Energy*, vol. 129, pp. 49–55, Sep. 2014.

[56] K. Srivastava, "Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment," *Int. J. Comput. Theory Eng.*, vol. 5, no. 3, p. 520, 2013.

[57] P. Berkhin, *A Survey of Clustering Data Mining Techniques, in Grouping Multidimensional Data*. Berlin, Germany: Springer, 2006, pp. 25–71.

[58] A. Gosain and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining," in *Proc. IEEE Conf. Inf. Commun. Technol.*, Apr. 2013, pp. 1003–1008.

[59] M. Fitzwater, "Efficient mining of maximal sequential patterns using multiple samples," in *Proc. SIAM Int. Conf. Data Mining*, 2005, p. 1.

[60] Z. Yang and M. Kitsuregawa, "LAPIN-SPAM: An improved algorithm for mining sequential pattern," in *Proc. 21st Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2005, p. 1222.

[61] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[62] K. Kalpakis and D. V. Gada Puttagunta, "Distance measures for effective clustering of ARIMA time-series," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2001, pp. 273–280.

[63] N. Kumar, *Time-Series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases*. Philadelphia, PA, USA: SIAM, 2005.

[64] D. Ryan, *High Performance Discovery in Time Series: Techniques and Case Studies*. Berlin, Germany: Springer, 2013.

[65] X. Wu and S. Zhang, "Synthesizing high-frequency rules from different data sources," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 353–367, Mar. 2003.

[66] R. Duan, X. Chen, and T. Xing, "A QoS architecture for IOT," in *Proc. Int. Conf. 4th Int. Conf. Cyber, Phys. Soc. Comput. (iThings/CPSCom)*, Oct. 2011, pp. 717–720.

[67] Y. Zhang, *ICN Based Architecture for IoT*. IRTF Contribution, Oct. 2013.

[68] S. Darby, "Smart metering: What potential for householder engagement?" *Building Res. Inf.*, vol. 38, no. 5, pp. 442–457, 2010.

[69] T. A. Rahman and S. K. A. Rahim, "RFID vehicle plate number (e-plate) for tracking and management system" in *Proc. Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2013, pp. 611–616.

[70] J. Sherly and D. Somasundareswari, "Internet of Things based smart transportation systems," *Int. Res. J. Eng. Technol.*, vol. 2, no. 7, 2015.

[71] N. Tohamy, "What you need to know about the Internet of Things," *MHD Supply Chain Solutions*, vol. 45, no. 3, p. 32, 2015.

[72] C. Pettey. (2015). *Five Ways the Internet of Things Will Benefit the Supply Chain*. [Online]. Available: http://www.gartner.com/smarterwithgartner/five-ways-the-internet-of-things-will-benefit-the-supply-chain-2/

[73] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 5–20, Feb. 2013.

[74] S. Bera, S. Misra, and J. J. P. C. Rodrigues, "Cloud computing applications for smart grid: A survey," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 5, pp. 1477–1494, May 2015.

[75] T. Dethlefs, *Energy Service Description for Capabilities of Distributed Energy Resources*. Berlin, Germany: Springer, 2015.

[76] C. Neureiter, "A standards-based approach for domain specific modelling of smart grid system architectures," in *Proc. 11th Int. Conf. Syst. Syst. Eng. (SoSE)*, Kongsberg, Norway, Jun. 2016, pp. 1–6.

[77] F. Bonomi, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile Cloud Comput.*, 2012, pp. 1–6.

[78] K. Steinklauber. (2014). *Data Protection in the Internet of Things*. [Online]. Available: https://securityintelligence.com/data-protection-in-the-internet-of-things

[79] T. Hu, H. Chen, L. Huang, and X. Zhu, "A survey of mass data mining based on cloud-computing," in *Proc. Anti-Counterfeiting, Secur. Identificat.*, Aug. 2012, pp. 1–4.

[80] Y. Sun, "Mining knowledge from interconnected data: A heterogeneous information network analysis approach," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 2022–2023, 2012.

[81] M. Chen, L. T. Yang, T. Kwon, L. Zhou, and M. Jo, "Itinerary planning for energy-efficient agent communications in wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3290–3299, Sep. 2011.

[82] D. Zhang, "A taxonomy of agent technologies for ubiquitous computing environments," *Trans. Internet Inf. Syst.*, vol. 6, no. 2, pp. 547–565, 2012.

[83] M. Chen, V. C. M. Leung, and S. Mao, "Directional controlled fusion in wireless sensor networks," *Mobile Netw. Appl.*, vol. 14, no. 2, pp. 220–229, Apr. 2009.

[84] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[85] K. Su, "A logical framework for identifying quality knowledge from different data sources," *Decision Support Syst.*, vol. 42, no. 3, pp. 1673–1683, 2006.

[86] L. Wang, G. Wang, and C. A. Alexander, "Big data and visualization: Methods, challenges and technology progress," *Digit. Technol.*, vol. 1, no. 1, pp. 33–38, 2015.

[87] A. T. Azar and A. E. Hassanien, "Dimensionality reduction of medical big data using neural-fuzzy classifier," *Soft Comput.*, vol. 19, no. 4, pp. 1115–1127, 2015.

[88] V. L. Popov and M. Heß, *Method of Dimensionality Reduction in Contact Mechanics and Friction*. Springer, 2015.

[89] C. Donalek *et al.*, "Immersive and collaborative data visualization using virtual reality platforms," in *Proc. IEEE Int. Conf. Big Data (USA Big Data)*, Oct. 2014, pp. 609–614.

[90] H. Childs *et al.*, "Research challenges for visualization software," *Computer*, vol. 46, no. 5, pp. 34–42, May 2013.

[91] E. Y. Gorodov and V. V. Gubarev, "Analytical review of data visualization methods in application to big data," *J. Elect. Comput. Eng.*, vol. 2013, Oct. 2013, Art. no. 969458.

[92] Center. (2013). *Big Data Visualization: Turning Big Data Into Big Insights [While Paper]*. [Online]. Available: https://future.transport.nsw.gov.au/wp-content/uploads/2016/02/big-data-visualization-turning-big-data-into-big-insights.pdf

[93] (2013). *Five Big Data Challenges and How to Overcome Them With Visual Analytics*. [Online]. Available: https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[94] P. Simon, *The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions*. Hoboken, NJ, USA: Wiley, 2014.

[95] B. B. Ahamed, T. Ramkumar, and S. Hariharan, "Data integration progression in large data source using mapping affinity," in *Proc. 7th Int. Conf. Adv. Softw. Eng. Appl. (ASEA)*, Dec. 2014, pp. 16–21.

[96] J. Liu and X. Zhang, "Data integration in fuzzy XML documents," *Inf. Sci.*, vol. 280, pp. 82–97, Mar. 2014.

[97] A. Ma'ayan *et al.*, "Lean big data integration in systems biology and systems pharmacology," *Trends Pharmacol. Sci.*, vol. 35, no. 9, pp. 450–460, 2014.

[98] P. B. D. Agrawal *et al.*, "Challenges and opportunities with big data: A community white paper developed by leading researchers across the United States," Computing Community Consortium, White paper, 2012.

[99] R. Agrawal *et al.*, "The claremont report on database research," *Communications*, vol. 52, no. 6, pp. 56–65, 2009.

MOHSEN MARJANI received the B.E. degree in mathematics in Iran, in 2003, and the Master of Information degree in multimedia computing from Multimedia University, Malaysia, in 2011. He is currently pursuing the Ph.D. degree with the Department of Computer Systems and Technology, University of Malaya, and a member of the High Impact Research Project (Mobile Cloud Computing: Data center architecture) which is fully funded by the Malaysian Ministry of Higher Education. He has been teaching for over ten years in different institutions in Iran.

FARIZA NASARUDDIN received the B.Sc. degree in computer science and the M.Sc. degree in MIS from Northern Illinois University, USA, and the Ph.D. degree from the University of Malaya. She was with the industry as a Systems Analyst for ten years before joining the academia in 1997. She is currently a Senior Lecturer with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya. She became involved in multi-disciplined research but her main focuses are in databases, information systems, and data sciences.

ABDULLAH GANI (SM'12) received the bachelor's and master's degrees from the University of Hull, Hull, U.K., and the Ph.D. degree from The University of Sheffield, U.K. He has vast teaching experience with various educational institutions locally and abroad—schools, teaching college, ministry of education, and universities. He is currently a Professor with the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He is the Director of the Centre for Mobile Cloud Computing Research, which focuses on high impact research, a Visiting Professor with King Saud University, Saudi Arabia, and also an Adjunct Professor with the COMSATS Institute of Information Technology, Islamabad, Pakistan. His interest in research started in 1983 when he was chosen to attend the Scientific Research Course in RECSAM by the Ministry of Education, Malaysia. He has published more than 100 academic papers in conferences and respectable journals. He actively supervises many students at all level of study—bachelor's, master's, and Ph.D. His interest of research includes self-organized system, reinforcement learning, and wireless-related networks. He was involved in mobile cloud computing with High Impact Research Grant of USD 500 000 (RM 1.5M) for the period of 2011 through 2016.

AHMAD KARIM received the M.S. degree (Hons.) in CS from Bahauddin Zakariya University, Multan, Pakistan, and the Ph.D. degree in computer security from the University of Malaya, Malaysia. He is currently a Lecturer with the Department of Information Technology, Bahauddin Zakariya University. His areas of research include mobile botnet detection, computer security, computer networks, software-defined networks, big-data analytics, and Internet of Things. He received Cisco International Certifications (CCNA, CCNP, and CCAI).

IBRAHIM ABAKER TARGIO HASHEM received the B.E. degree in computer science in Sudan in 2007, and the M.S. degree in computing in Malaysia in 2012. He received the Ph.D. degree in computer science from the University of Malaya, Kuala Lumpur, Malaysia. He was a Tutor with CISCO Academy, University of Malaya. His main research interests include big data, cloud computing, and distributed computing and network. He received professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and the APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation).

AISHA SIDDIQA received the B.Sc. degree in information technology from Bahauddin Zakariya University, Multan, Pakistan, and the M.Sc. degree in computer science from the COMSATS Institute of Information Technology, Islamabad, Pakistan. She is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, where she is also a Researcher. Her area of research is algorithms and data structures, big data storage, big data analytics, and indexing.

IBRAR YAQOOB received the B.S. degree (Hons.) in information technology from the University of the Punjab, Gujranwala, Pakistan, in 2012. He has been pursuing the Ph.D. degree in computer science with the University of Malaya, Malaysia, since 2013. He is also a Bright Spark Program Research Assistant. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are most downloaded in top journals. His research interests include big data, mobile cloud, Internet of Things, cloud computing, and wireless networks.