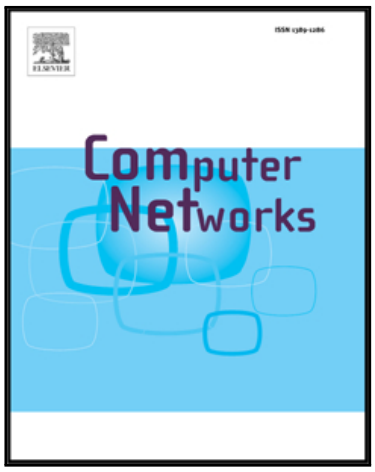
Accepted Manuscript



The role of big data analytics in Internet of Things

Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, Athanasios V. Vasilakos

PII: S1389-1286(17)30259-1
DOI: 10.1016/j.comnet.2017.06.013
Reference: COMPNW 6240

To appear in: *Computer Networks*

Received date: 19 December 2016
Revised date: 3 June 2017
Accepted date: 15 June 2017

Please cite this article as: Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, Athanasios V. Vasilakos, The role of big data analytics in Internet of Things, *Computer Networks* (2017), doi: 10.1016/j.comnet.2017.06.013

The role of big data analytics in Internet of Things

Ejaz Ahmed, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Imran Khan, Abdelmuttlib Ibrahim Abdalla Ahmed, Muhammad Imran, and Athanasios V. Vasilakos

Abstract—The explosive growth in the number of devices connected to the Internet of Things (IoT) and the exponential increase in data consumption only reflect how the growth of big data perfectly overlaps with that of IoT. The management of big data in a continuously expanding network gives rise to non-trivial concerns regarding data collection efficiency, data processing, analytics, and security. To address these concerns, researchers have examined the challenges associated with the successful deployment of IoT. Despite the large number of studies on big data, analytics, and IoT, the convergence of these areas creates several opportunities for flourishing big data and analytics for IoT systems. In this paper, we explore the recent advances in big data analytics for IoT systems as well as the key requirements for managing big data and for enabling analytics in an IoT environment. We taxonomized the literature based on important parameters. We identify the opportunities resulting from the convergence of big data, analytics, and IoT as well as discuss the role of big data analytics in IoT applications. Finally, several open challenges are presented as future research directions.

Index Terms—Internet of things, big data, analytics, distributed computing, smart city.

1 Introduction

The technological advancements and rapid convergence of wireless communication, digital electronics, and micro-electro-mechanical systems (MEMS) technologies have resulted in the emergence of Internet of Things (IoT). According to the Cisco report[1], the number of objects connected to the Internet has exceeded the number of human beings in the world. These Internet-connected objects, which include PCs, smartphones, tablets, WiFi-enabled sensors, wearable devices, and household appliances, form the IoT as shown in Figure 1. Reports show that the number of Internet-connected devices is expected to increase twofold from 22.9 billion in 2016 to 50 billion by 2020 as shown in Figure 2.

Most IoT applications do not only focus on monitoring discrete events but also on mining the information collected by IoT objects. Most data collection tools in the IoT environment are sensor-fitted devices that require custom protocols, such as message queue telemetry transport (MQTT) and data distribution service (DDS). Given that sensors are used in nearly all industries, the IoT is expected to produce a huge amount of data. The data generated from IoT devices can be used in finding potential research trends and investigating the impact of certain events or decisions. These data are processed using various analytic tools [1]. Fig-

- *E. Ahmed, I. Yaqoob, I.A.T. Hashem, and A.I.A. Ahmed are with the Centre for Mobile Cloud Computing Research, Faculty of Computer Science and Information Technology, University of Malaya. (E-mail: ejazahmed@ieee.org, ibraryaqoob@siswa.um.edu.my, targio@siswa.um.edu.my, and abdelmuttlib@siswa.um.edu.my)*
- *I. Khan is working in Schneider Electric Industries, 38TEC, Grenoble, France. (Email: imran@ieee.org)*
- *M. Imran is with the College of Computer and Information Sciences, King Saud University, Saudi Arabia. (E-mail: dr.m.imran@ieee.org)*
- *Athanasios V. Vasilakos is working with the Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Sweden (e-mail: athanasios.vasilakos@ltu.se)*

1. http:www.cisco.comcdamen_usaboutac79docsinnovIoT_IBSG_0411FINAL.pdf

Fig. 1: Big Data Sources in IoT

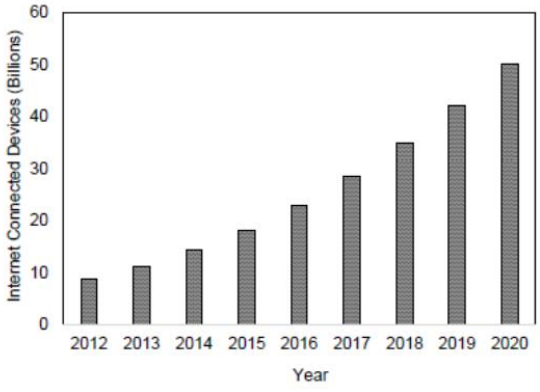


Fig. 2: Number of Internet-Connected Devices[1]

ure 3 illustrates the process of data collection, monitoring, and data analytics[2].

Although IoT has created unprecedented opportunities that can help increase revenue, reduce costs, and ameliorate efficiencies, collecting a huge amount of data alone is insufficient. To generate benefits from IoT, enterprises must create a platform where they can collect, manage, and analyze a massive volume of sensor data in a scalable and cost-effective manner [2]. In this context, leveraging a big data platform that can assist in consuming and reading diverse data sources as well as in accelerating the data integration process becomes vital. Data integration and analytics allow organizations to revolutionize their business process. Specifically, these enterprises can use data analytics tools to transform a huge volume of sensor-collected data into valuable insights. Given the overlapping research trends in these areas, this paper focuses on the recent advances in management of big data and analytics in the IoT paradigm.

The contributions of this paper are as follows:

- We critically review the recent literature.
- We discuss big data processing and analytics platforms in the IoT environment.
- We discuss the key requirements for big data processing and analytics in an IoT environment.
- We taxonomized the literature based on important parameters.
- We discuss the potential opportunities in big data processing and analytics in the IoT paradigm and highlight the role of data analytics in IoT applications.
- We discuss the open research challenges and highlight the vision of big data analytics in IoT as future research directions.

These contributions are given in separate sections from 2-8. We provide concluding remarks in section 9.

2 Recent Advances in IoT-based Big Data and Analytics

Bashir and Gill [3] propose an IoT big data analytics framework to overcome the challenges of storing and analyzing large amount of data originating from smart buildings. The proposed framework is composed of three components which are big data management, IoT sensors, and data analytics. The analytics are performed in real-time in order to be used in different parts of the smart building to manage the oxygen level, smoke/hazardous gases, and luminosity. The framework is implemented in Cloudera

2. http://www.businessinsider.com/how-the-internet-of-things-market-will-grow-2014-10

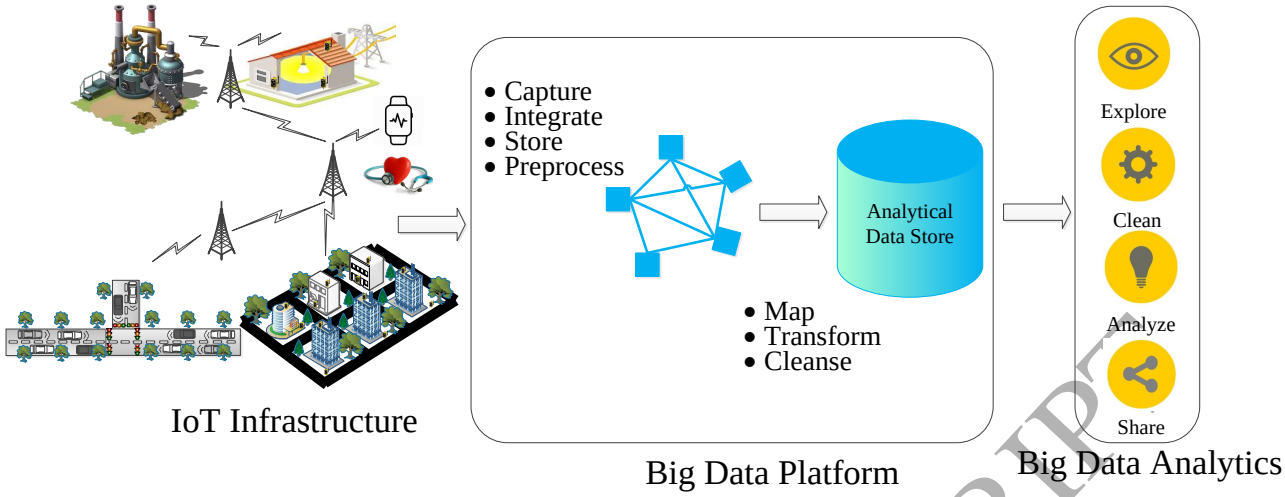


Fig. 3: Big Data Flow in IoT

Hadoop distribution where the analytics is performed using PySpark. The results show that the framework can be utilized for IoT-enabled big data analytics. The proposed framework is specifically designed for smart buildings that should be extended to make it generalize so that it can deal with other IoT applications including smart cities and smart airplanes.

C. Lee et al. [4] propose an IoT-based cyber physical system that supports information analysis and knowledge acquisition methods to improve productivity in various industries. This system, which focuses on industrial big data analytics, integrates various data analytics components in the form of reconfigurable and interchangeable modules to meet different business needs. The authors also provide a new context intelligence framework that can help handle industrial informatics based on the sensors, locations, and unstructured data for big data mining. A case study is also performed to illustrate the design of the proposed cyber physical system.

P. Rizwan et al. [5] study the strengths and weaknesses of various traffic management systems. They propose a low cost, real-time traffic management system that deploys IoT devices and sensors to capture real-time traffic information. Specifically, low-cost traffic detection sensors are embedded in the middle of the road for every 500 or 1000 meters. The collected data are then sent to analytics tools to analyze traffic density and to provide solutions via predictive analytics. Compared with the existing systems, the proposed system provides a better alternative method for managing traffic.

Q. Zhang et al. [6] propose Firework, a new computing paradigm that allows distributed data processing and sharing in an IoT-based, collaborative edge environment. Firework combines physically distributed data by providing virtual data views to end users using predefined interfaces. These interfaces come in the form of a set of functions and a set of datasets. Firework aims to minimize data access latency by moving the processing closer to the data producers in the edge network. Firework instance has multiple stakeholders who must register their datasets and corresponding functions that are abstracted as data views. These data views are available to all participants of the same framework instance such that they can merge multiple data views into a single job to perform detailed data analytics. They illustrate such concept by performing case studies of *connected health* and *find the lost*.

M. M. Rathore et al. [7] propose a smart city management system based on IoT that exploits big data and analytics. The data are collected by deploying different sensors, including weather and water sensors, vehicular networking sensors, surveillance objects, smart

home sensors, and smart parking sensors. An architecture for the system and its model for implementation have also been designed. The proposed system is implemented using the MapReduce Hadoop ecosystem in a real environment. The implementation process involves several steps, including data generation, data gathering, data combining, data categorization, data preprocessing, and decision making. Spark over Hadoop is used for the efficient processing of big data. Smart systems are utilized as sources of city data to develop a smart city as an implemented system. However, the developed smart system is yet to be deployed and its accuracy remains untested.

B. Ahlgren et al. [8] discuss the significance of using IoT to deliver services for improving the lives of citizens, including transportation, air quality, and energy efficiency. The authors emphasize that IoT-based systems must be based on open data and standards, including interfaces and protocols, to enable third-party innovations by mitigating manufacturer lock-ins. Based on this idea, the authors design and develop a GreenIoT platform in Sweden to determine the advantage of open platforms and open data for the development of smart cities. However, some guidelines regarding the procurement of an open IoT infrastructure, including common data formats and open application programming interfaces (APIs), must be devised.

O. B. Sezer et al. [9] propose an augmented framework that integrates semantic web technologies, big data, and IoT. The key requirements for the proposed framework are analyzed, and the conceptual design of the envisioned IoT system is proposed based on the analysis results. The conceptual framework comprises five layers, namely, data acquisition, extract-transform-load (ETL), semantic-rule reasoning, learning, and action. The data acquisition layer, which collects data from different sources, can be considered as an input layer to the framework. The ETL layer provides sensor drivers to transform the data received from different types of sensors. The semantic-rule reasoning supports a reasoning engine to make inferences from the resource description framework (RDF) data received from the ETL layer. The learning layer extracts several features from the data and forms machine-learning-based models. The action layer provides predetermined actions for the output of the learning layer.

B. Cheng et al. [10] design GeeLytics, an edge analytics platform that performs real-time data processing at the network edges and in the cloud. This platform addresses the geo-distributed and low-latency analytics resulting from the large amounts of IoT data. GeeLytics is designed to support dynamic stream processing topologies by taking into account the system characteristics of heterogeneous edge/cloud nodes, and the current system workload.

H. Wang et al. [11] discuss the challenges and opportunities resulting from IoT and big data for the maritime cluster. They also develop a new framework for integrating industrial IoT with big data and analytics technologies. Implementing such framework can help increase output and productivity as well as allow whole clusters to continue acting as leaders in the global maritime industry.

Prez and Carrera [12] conduct a comprehensive study of the performance characterization of the servIoTicy API. They specifically focus on the state-of-the-art infrastructure for hosting IoT workloads in the cloud with an aim to provide multi-tenant data stream processing capabilities, advanced querying mechanisms, multi-protocol support, and software solutions by combining advanced data-centric technologies. Another study [13] partially solves the big data storage problem by proposing AllJoyn Lambda, a software solution that integrates AllJoyn in the Lambda architecture that is used for big data storage and analytics.

A. J. Jara et al. [14] conduct a survey to highlight the existing solutions and challenges to big data that are posed by cyber-physical systems. Their study focuses on cloud security and the heterogeneous integration of data from multiple sources. They highlight the need for developing sophisticated data discovery mechanisms and for performing real-time stream data processing.

Z. Ding et al. [15] propose a general statistical database cluster mechanism for big data analysis in the IoT paradigm (IOT-StatisticDB). They input statistical functions on IOT-StatisticDB via statistical operators inside the database management systems (DBMS) kernel. The statistical analysis is performed in a distributed and parallel fashion using multiple servers.

C. Vuppalapati et al. [16] examine the role of big data in healthcare and find that body sensors generate massive amounts of health-related data. Two challenges are analyzed in this context, namely, integrating these massive data points with electronic health records (EHR) and presenting these data to doctors in real time. Based on these observations, they propose a sensor integration framework that suggests a scalable cloud architecture that can provide a holistic approach to the EHR sensor system. Apache Kafka and Spark are used to process large amounts of data in a real-time manner. Although visualizing the health of patients in real time can help detect urgent situations, this model lacks a security solution.

A. Ahmad et al. [17] analyze human behavior by using big data and analytics in the social IoT paradigm [18]. They propose an architecture that comprises three operational domains. They also analyze an ecosystem that is created by smart cities and big data. Collaborative filtering techniques can be used in the future to accurately analyze human behavior.

D. Arora et al. [19] utilize big data and analytics techniques to classify network-enabled devices. They also analyze the performance of four machine learning algorithms, such as k-nearest neighbor (KNN), NaveBayes (NB), support vector machines (SVM), and random forest. The experimental results show that the NB algorithm yields the lowest accuracy among all classifier models, while the random forest algorithm yields the highest accuracy. Meanwhile, the accuracy of KNN and SVM are closely related to that of the random forest algorithm.

I. L. Yen et al. [20] investigate the potential of service discovery and composition techniques in solving real-world problems based on the data generated through IoT. They examine how various technologies, such as data analytics and artificial intelligence, can be used in the smart world to derive situational facts and to take actions accordingly. They propose a gaming-based crowdsourcing platform to make use of human intelligence for the successful completion of certain control tasks. In the future, proactive monitoring and diagnosis mechanisms with a combination of big data mining must be devised to ensure safety in the smart physical world.

R. P. Minch et al. [21] perform an exploratory research about location privacy in the era of IoT, big data, and analytics. They identify, classify, and describe privacy issues and reveal the possible pain points in the context of big data and analytics. They suggest that a reliable framework for securing privacy in a context-aware environment must be developed in the future.

A. Mukherjee et al. [22] propose an IoT framework for the effective execution of data parallel analytic jobs. They aim to identify a suitable analytical algorithm that can cope up with the requirements of processing and analyzing large amounts of data. Their qualitative analysis generates promising results because of the high effectiveness of the parallel analytic algorithms in an IoT environment. Future studies must address those issues that hinder the implementation of this model in the presence of fog computing.

A. Mukherjee et al. [23] publish a report regarding Condor, a grid framework for data-parallel execution in the IoT paradigm. Their experimental results reveal that Condor has a better scalability and CPU utilization for data-parallel jobs compared with a traditional three-tier, server-based architecture.

H. R. Arkian et al. [24] propose MIST, a fog-based data analytics scheme with a cost-efficient resource provisioning optimization approach that can be used for IoT crowd sensing applications. This scheme aims to reduce the latency of service provisioning in traditional cloud computing frameworks. The experimental results show that the MIST fog-based scheme outperforms traditional cloud computing as the number of applications that demand

real-time services increases. Some possible extensions of this work are as follows: (a) adding a selective sensing module to the fog layer, (b) enriching the architecture with privacy-preserving data analytics capabilities, and (c) considering the mobility of data generators and data consumers in the resource provisioning part.

M. M. Rathore et al. [25] propose a system that deals with several problems in a smart city environment, such as enabling objects to react with respect to context, minimizing the cost of collecting data generated by smart devices, and obtaining insights into the data if these data are collected and processed in real time. The proposed system has a four-tier architecture, where the bottom tier is responsible for data generation and collection, the intermediate tier 1 enables communication among sensors, relays, base stations, and the Internet, the intermediate tier 2 is responsible for data management and processing using the Hadoop framework, and the top tier is responsible for applying data analysis techniques and generating results. The implementation results show that the proposed system is more scalable and efficient in terms of throughput and processing time than the current systems. However, this system lacks an intelligent decision-making technique that can cope with big data in an IoT environment.

F. Alam et al. [26] examine the applicability of eight data mining algorithms, including SVM, KNN, NB, C4.5, C5.0, linear discriminant analysis (LDA), artificial neural network (ANN), and deep learning ANN (DLANN), for IoT-generated data. These algorithms are also compared in terms of their confusion matrix, classification accuracy, and execution time. With regard to classification accuracy, C4.5, C5.0, ANN, and DLANN outperform SVM, KNN, NB, and LDA. However, C4.5, C5.0, and ANN are very similar in terms of classification accuracy. Meanwhile, NB and LDA have the fastest execution time, with LDA having a slightly better processing time than NB. The authors are planning to conduct a detailed study on larger and diverse IoT datasets in the future.

M. H. Berlian et al. [27] introduce a framework for monitoring and analyzing large amounts of data that are generated through the Internet of Underwater Things (IoUT). They utilize MapReduce to process these data, and find that MapReduce greatly shortens the query execution time compared with SQL. Despite the many advantages of this framework, testing the applicability of the scalable trust management protocol with IoUT applications and developing trust-based admission control for IoUT systems still need to be addressed in the future.

D. Mourtzis et al. [28] reveal that the adoption of IoT in the manufacturing industry can transform traditional systems into modern ones. Moreover, such transformation leads to a data production process that turns industrial data into industrial big data, which are rendered useless without analytics power. Adopting data analytics can empower enterprises to derive new data-driven strategies that can easily manage competitive pressure. They also demonstrate how the IoT paradigm can be implemented in a simple case of a company with almost 100 machines.

R. Ramakrishnan et al. [29] analyze the current energy development in India and determine the benefits that can be obtained through cloud computing and analytics. They also advocate that the usage of analytics can improve energy conservation, reduce operation costs, and empower customers.

3 Big Data Processing and Analytics Platforms

This section investigates the big data processing and analytics platforms that can be used for large amounts of IoT-generated data. In IoT, the big data processing and anlytics can be performed closer to data source using the services of mobile edge computing [30], [31], cloudlets [32] and fog computing [33].

3.1 Apache Hadoop

First used by Yahoo! and Facebook, Hadoop [34] is an open source data processing platform that stores and processes large amounts of data on a cluster of commodity hardware. The Hadoop architecture contains several components, of which the most important are the

Hadoop Distributed File System (HDFS) and the MapReduce programming model. HDFS is used to store the data, while MapReduce is used to process these data in a distributed manner [35]. Despite its many advantages, Hadoop lacks encryption at the storage and network levels, has a limited flexibility, is considered unsuitable for small data sets, and has a high I/O overhead.

3.2 1010data

1010data [36] consists of a columnar database and mostly deals with semi-structured data, such as IoT data. Aside from its data visualization, reporting, and integration capabilities, this tool provides advanced analytic services, including optimization and statistical analysis. 1010data is also very supportive for large-scale infrastructure. This tool also works in a centralized fashion and applies access controls to interact with back-end systems. 1010data can satisfy customer demand through its advanced analytic capabilities. However, 1010data is considered ineffective in terms of data extraction, transformation, and loading.

3.3 Cloudera Data Hub

Cloudera introduced the Enterprise Data Hub [37], a Hadoop-based framework for big IoT data processing and analytics that can be utilized as a central point in managing massive amounts of IoT data from enterprises. To achieve reliability, high performance, security, and data-access control, the Cloudera Data Hub combines Cloudera Manager, Navigator, and its backup and recovery components. However, this tool does not have its own hardware and software systems and merely relies on third parties when identifying serious privacy and security concerns.

3.4 SAP-Hana

SAP-Hana [38] is an in-memory platform for performing big IoT data analytics and addressing transactional needs. SAP supports various distributed solutions to accommodate big unstructured data. Hana accesses big data through Hive, while SAP uses Sybase IQ to provide a columnar DBMS. Hana also has a built-in analytics library for containing, spatial processing, and supporting R language and text analytics libraries. Apart from its low latency, SAP-Hana can also analyze both text and unstructured data. However, in this tool, all data in a row must be read even though only the data from a few columns are required to be accessed. Moreover, the capabilities of SAP-Hana are not strong enough compared with those of other solutions.

3.5 HP-HAVEn

HP introduced the Hadoop Autonomy Vertica Enterprise (HAVEn) [39] security, a new big IoT data platform architecture for a large number of HP systems that can be used with any number of applications. HP provides reference hardware configurations for the major distributors of the Hadoop software. Autonomy's IDOL software provides search and exploration services for unstructured data. Vertica is an analytical DBMS for a massively parallel processing columnar database that aims to accelerate the analysis of big structured datasets. HP HAVEn is currently collaborating with several companies to complement legacy enterprise data warehouses. HP also introduced a "Flex-Zone" to facilitate the exploration of large datasets before defining the database scheme. The only drawback of HP-HAVEn is that an increment in the number of tenants generates a large database catalog where the lock holding and release processes in all operations are decelerated.

3.6 Hortonworks

Hortonworks [40] focuses on building a big IoT data analytics and management platform based on Hadoop. The Hortonworks Data Platform (HDP) has a free open source software distribution and focuses on the improvement of Hive. However, with its HDP plugin, Hortonworks cannot reduce the number of node-groups or hosts per node group in the generated cluster.

3.7 Pivotal big data suite

The Pivotal big data suite (Pivotal BDS) [41], which is usually deployed in a public cloud, comprises three solutions, namely, Pivotal HDB, Pivotal Greenplum, and Pivotal GemFire, all of which are delivered under a single license. Pivotal is an analytical database that combines massively parallel processing (MPP)-based analytics performance with robust ANSI SQL compliance and helps in performing predictive analytics on data that are stored in HDFS using SQL syntax and other related tools. Pivotal Greenplum is an open source MPP analytical database that is used for performing rapid analytics on voluminous amounts of data and provides high query performance on petabyte-scale data volume. Pivotal GemFire is an in-memory data grid that is designed to support high volumes of operational and transactional applications. Despite its many benefits, Pivotal BDS is still in its infancy and its wide adoption is mired by many unresolved issues.

3.8 Infobright

A tool specifically designed for solving data management and analytic problems [42], Infobright can analyze up to 50 terabytes of data. With its high compression and data skipping ratio, Infobright is considered suitable for machine-generated data, such as IoT data. Infobright mostly works with Hadoop or high-scale data warehouses. The data skipping technology and columnar design of this tool ensure that only the concerned data will be used in each query. These data are also indexed automatically without the need of any partitioning and tuning. However, all queries cannot be answered optimally using the Infobright optimizer.

3.9 MapR

MapR [43] supports big data and analytics as well as adopts several components of Hadoop to improve its performance (e.g., replacing HDFS with an NFS-like network file system to achieve security and high availability). MapR also has its own system recovery approach. Recently, MapR added LucidWorks Search and stream processing options into Hadoop to enhance its predictive capabilities and enable fast processing. However, MapR has a higher complexity compared with Hadoop.

4 Requirements

The requirements of big data and analytics in IoT have exponentially increased over the years and promise dramatic improvements in decision-making processes. As a result, the demands of adapting data analytics to big data in IoT have increased as well, thereby changing the way that data are collected, stored, and analyzed. Big data and analytics have great potential for extracting meaningful information from the data produced by sensor devices. The general requirements for big data and IoT define the functional and nonfunctional specifications for data analytics. This section presents the key requirements for big data and analytics in the IoT environment. These requirements play an important role in improving IoT services through analytics.

4.1 Connectivity

The IoT paradigm is gradually leading to the ubiquitous connectivity of intelligent sensor-equipped objects in a smart environment. One of the key requirements of IoT is to provide a reliable connectivity for big data and analytics to facilitate the combination and integration of huge volumes of machine-generated sensor data. Thus, numerous objects around us have a great potential to be connected to high-performance computing infrastructures to enhance IoT services. Moreover, with the growing presence of WiFi and 4G-LTE wireless Internet access, the evolution toward ubiquitous information and communication networks is already evident [44]. However, a seamless connection among different objects in smart cities [45], such as IoT, cloud computing, big data, and analytics, must be established before embedding intelligence into our environment.

4.2 Storage

The continuous rapid growth of a large number of IoT-enabled objects has resulted in the storage of massive amounts of heterogeneous data in low-cost commodity hardware on a real-time basis. The key requirements of big data storage in IoT include handling very large amounts of unstructured data and providing low latency for analytics. Moreover, the applications of big data technologies for IoT can enable efficient data storage and processing in order to produce information that can enhance different smart city services. The spectrum of IoT data sources includes sensor data, smartphones, and social media that are modeled in different ways and use various communication protocols and interfaces. Most IoT services are based on M2M communication protocols, which require handling a large number of streams and directly benefit from the widely distributed storage capacities of cloud computing infrastructure [46].

4.3 Quality of services

The resource management of IoT sensors and mobile devices is the primary requirement for quality of service (QoS) to effectively analyze a huge amount of data. Although many studies have attempted to meet the QoS requirement, how to unify and integrate the QoS architecture into IoT to support big data and analytics warrants further research [47]. The QoS provided by an IoT network must be reliable and must guarantee a mobile and efficient transfer of data from those sources where big data are generated. The QoS support in this network is extremely important to big data and analytics. However, to create a reliable network, many emerging networking technologies must be introduced into IoT to enable real-time event transfer and improve big data processing capabilities.

4.4 Real-time analytics

Streaming analytics has rapidly emerged as a key IoT initiative for timely decision-making processes [48]. One of the most prominent features of IoT is its real-time or near-time communication of information regarding "connected things". Big data and analytics in IoT require streaming events on the fly and storing streaming data in an operational database. Given that much of these unstructured data are streamed directly from web-enabled "things", big data implementations must perform analytics with real-time queries to help organizations obtain insights quickly, rapidly make decisions, and interact with people and other devices in real time [15].

4.5 Benchmark

Big data and analytics have attracted much attention from the academia and various organizations, and many organizations have started pursuing IoT businesses as well. However, these organizations face some challenges in storing and analyzing vast amounts of data that are collected through sensors in an IoT environment. Solving these problems requires a deep understanding that can be achieved by using a big data and analytics platform. Benchmark plays an important role in this context by providing organizations with a way to judge the quality of big data and analytics solutions. An excellent system benchmark can also provide simple and straightforward comparisons of various solutions.

5 Taxonomy

Figure 4 shows the thematic taxonomy of big data and analytics solutions that are designed for IoT systems. These solutions are categorized based on the following attributes: a) big data sources, b) system components, c) big data enabling technologies, d) functional elements, and e) analytics type.

5.1 Big data sources

Big data are generated by an infrastructure that is deployed to run various IoT applications, including city management, manufacturing, intelligent transport systems (ITS), smart building, and monitoring sensors.

The city management uses connected cameras, sensors, and actuators to make the lives of citizens secure and convenient. However, these

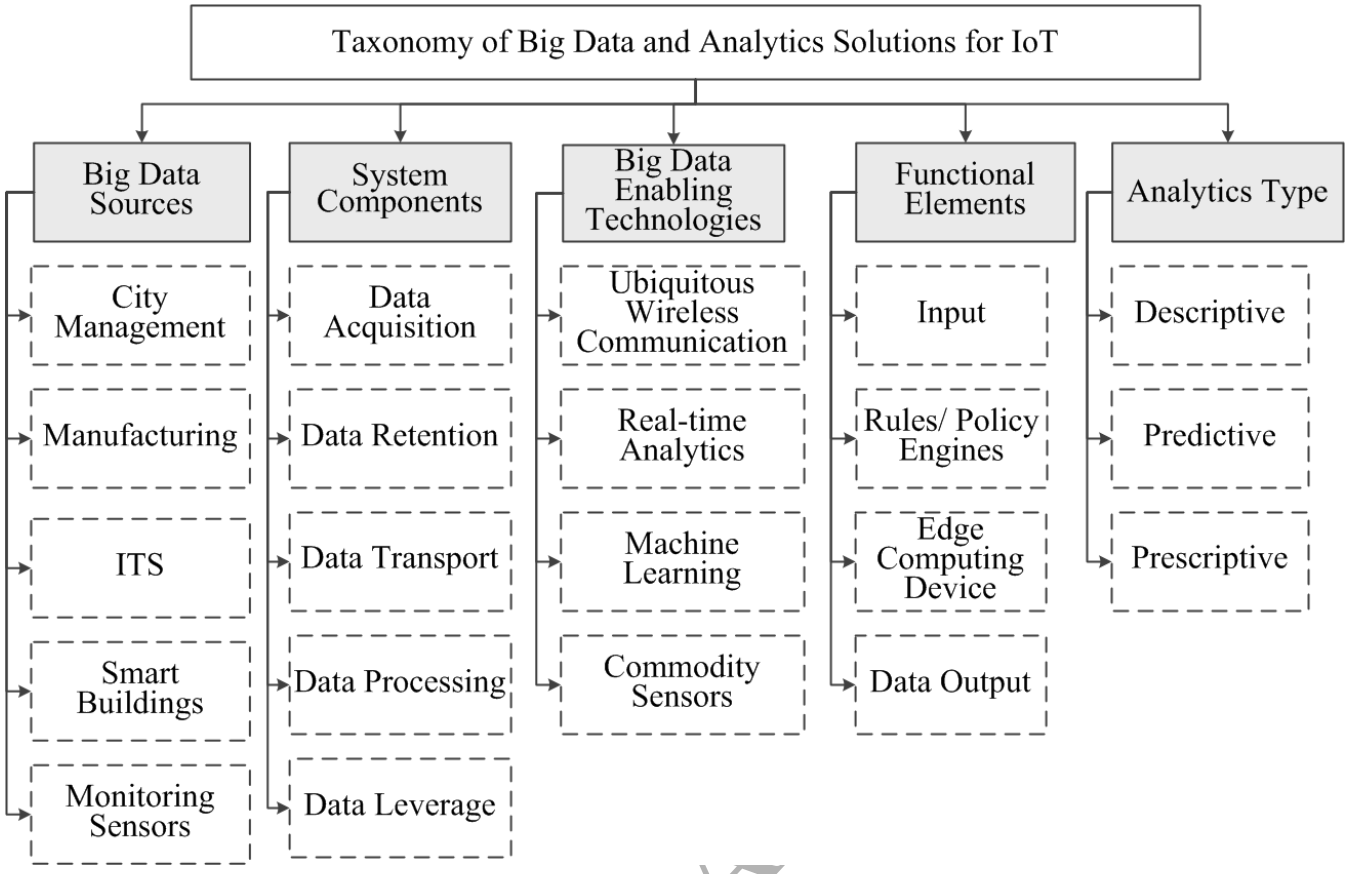


Fig. 4: Taxonomy of Big Data and Analytics Solutions for IoT Systems

devices generate a bulk of data that must be managed and analyzed in real time to obtain relevant insights. Similarly, the manufacturing industry has deployed IoT devices that continuously generate a huge amount of data to maximize the productivity and efficiency of its operations. To obtain insights from these data, big data and analytics solutions have been used in designing and testing new products, optimizing services and marketing, minimizing defects, and improving yields.

Along with big data and analytics, the proliferation of sensors [49], connected vehicle technologies [50], [51], and IoT [52] have resulted in the creation of intelligent transportation systems, thereby significantly increasing the amount of real-time big data that must be communicated, aggregated, analyzed, and managed. The ITS can take advantage of big data and analytics to enhance the decision-making capabilities of its users.

The use of big data solutions in smart building has the same goal as that in the other application domains. The relevant information is extracted from a wide range of existing data and then provided to decision makers for service management and to the users of the building.

Big data in the IoT environment are commonly used for the collection and storage of monitoring sensor data, performing data analytics, making forecasts, and generating alerts if abnormal deviations are detected.

5.2 System Components

Big data and analytics solutions usually comprise five system components, namely, data acquisition, data retention, data transport, data processing, and data leverage.

Big data acquisition involves collecting, filtering, and cleaning the data before they are transferred into the data warehouse. This component is commonly governed by four attributes, namely, volume, variety, velocity, and value. Big data retention deals with the extant

policies and requires the management to meet big data archival requirements. Various big data retention policies involve privacy and legal concerns against economics to identify archival rules, retention time, data formats, and encryption methods. The big data must be transported across different data sites to guarantee load balancing, business continuity, and replication.

Big data is a term used for large and complex datasets that cannot be processed by traditional software. The key challenges involved in big data processing are related to capturing, storage, analysis, search, updating, visualization, and privacy. Big data leverage involves ensuring how a business can reap benefits from their data to increase their revenue.

5.3 Big Data Enabling Technologies

The big data enabling technologies in the IoT context are related to ubiquitous wireless communication, real-time analytics, machine learning, and data capturing elements, such as commodity sensors and embedded systems.

The key ubiquitous wireless communication technologies that are used for transporting big data in IoT include IEEE 802.15.4, IEEE 802.11, IEEE 802.15.1, and IEEE 802.16.

Real-time analytics make the big data generated by IoT devices ready to use as they enter the system. Real time can be defined as a level of computer responsiveness that is either instantaneous or nearly instantaneous.

Unlike traditional analytic tools, machine learning can exploit the hidden insights in big data and extract values from big data sources with minimal human interaction. Machine learning is well suited in the IoT context because of the different data sources and the huge amount and variety of data involved.

The big data in IoT are collected by using several sensors and actuators. These sensor technologies have key roles in collecting and transmitting data to the nearby edge resources for further processing.

5.4 Key Elements

The big data and analytics solutions for IoT comprise four key elements, namely, input, rules or policy engines, edge computing devices, and data output.

The raw data are collected from different resources and transferred to edge analytics systems. These systems are based on a rules/policy engine that defines and applies rules to the input data in order to obtain insights. The edge computing device is another key player in the operation of the analytics system. Processing the data that are generated by IoT devices on the edge devices can bring several advantages, such as low latency, minimal bandwidth consumption, data integrity, security, and low cost [30], [31], [53]. These data are also made available to the consumer in real time.

5.5 Analytics Type

Analytics can be divided into three types, namely, desrciptive analytics, predictive analytics, and perspective analytics. Descriptive analytics, which defines "what has happened or what is happening," helps find new business opportunities and challenges. Predictive analytics, which defines "what will happen and why it will happen,"is enabled by using various technologies, such as text/web/data mining, to accurately predict future conditions and states. Prescriptive analytics, which defines "what should I do and why should I do it," utilizes simulation, expertise, and decision support systems to investigate various choices and provide suggestions to decision makers.

6 The Role of Data Analytics in IoT Applications

Big data technologies can offer data storage and processing services in an IoT environment, while data analytics allow business people to make better decisions. IoT applications are the major sources of big data. This section explains the role of big data and analytics in different IoT applications, including smart grids, smart healthcare, smart transportation, and smart inventory systems [44], [54], [55]. Table 1 summarizes the benefits of big data and analytics in IoT applications.

TABLE 1: Benefits of Data Analytics for IoT Applications

IoT Application	Benefits of Data Analytics
Smart Transportation	(a) Reduce the number of accidents by looking into the history of the mishaps (b) Minimize traffic congestion (c) Optimize shipment movements (d) Ensure road safety
Smart Healthcare	(a) Predict epidemics, cures, and disease (b) Help insurance companies make better policies (c) Pick up the warning signs of any serious illnesses during their early stages
Smart Grid	(a) Help design an optimal pricing plan according to the current power consumption (b) Predict future supply needs (c) Ensure an appropriate level of electricity supply
Smart Inventory System	(a) Detect fraudulent cases (b) Strategically place an advertisement (c) Understand customer needs (d) Identify potential risks

6.1 Smart Transportation

Finding valuable information has become a key concern in this modern age of technologies where vehicles are connected to the Internet and generate large amounts of data. Data analytics can help transport management authorities to find out the history of road mishaps (e.g., under what circumstances did the accident occur and at what speed were the drivers driving during the mishap), minimize the number of road accidents, determine the time when the traffic load reaches its peak, and prepare an optimal route plan that can help minimize traffic congestion.

The analytics of smart transport data can indirectly optimize shipment movements, improve road safety, and enhance end-to-end user experience in terms of delivery time.

6.2 Smart Healthcare

Over the past few years, voluminous amounts of data have been created in the healthcare sector. However, such rapid increase in data production has created challenges in extracting valuable information from big healthcare datasets that can help predict epidemics and find cures for various diseases. Data analytics can help healthcare specialists analyze a large amount of patient data and learn the history of a disease (in the case of family doctors). Insurance companies may also use data analytics when making policies. Healthcare specialists may also detect serious illnesses at their early stages and subsequently prevent the loss of life.

6.3 Smart Grid

Smart grids rapidly generate data, and finding useful information from these data has become imperative. In a smart grid environment, large amounts of data are collected from various sources, such as the power utilization habits of users, phasor measurement data for situational awareness, and energy consumption data measured by widespread smart meters, to name a few [55]. Proper analytics can help decision makers measure the appropriate level of electricity supply that they must provide to their customers. Analytics may also help business people predict electricity demands in the near future. The strategic objectives of specific organizations can also be met through proper analytics (e.g., pricing plans that are consistent with supply, demand, and production models).

6.4 Smart Inventory System

Finding useful information from large amounts of inventory systems data can help business owners generate more profit. The analytics of inventory-systems-generated datasets can help one acquire knowledge about market trends. Product recommendations can be generated after analyzing seasonal variations. The analytics of inventory data can also help detect fraudulent cases. Analytics may aid advertisers in strategically placing their advertisements. Predictive analytics can help people make valuable decisions and understand further their customers and products. Data analytics can also

help companies identify their potential risks and opportunities.

7 Opportunities

The current IoT environment provides the following opportunities for effective big data and analytics.

7.1 Decision making

The proliferation of IoT devices, smart phones, and social media offers decision makers with an opportunity to extract valuable information about their users, to predict future trends, and detect fraud. Big data can generate significant value by making information transparent and usable to organizations, thereby helping them expose variability and boost their performance. Much of the data generated through IoT and various analytics tools create a large number of opportunities for organizations. These tools leverage predictive modeling, classification, and clustering techniques to offer various data mining solutions. Mining IoT can also improve the decision-making habits of individuals using big data.

7.2 Improved Efficiency

The processing and data storage demands of advanced analytics applications have limited their adoption in many domains. However, such barriers are beginning to fall because of IoT. Big data technologies, such as Hadoop and cloud-based mining tools, offer substantial advantages in terms of cost reduction compared with traditional mining techniques. Moreover, traditional analytics techniques require data to be in a certain format, which is difficult to achieve when using IoT data. However, using current big data technologies that build around low-cost community hardware can help improve analytics capability and reduce processing costs.

7.3 Independence from Data Silos

The advent of IoT and enabling technologies such as cloud computing has allowed the removal of data silos in different domains. Typically, each type of data is considered useful only for its domain, but cross-domain data have emerged as efficient solutions to different problems [56]. Different types of data, such as runtime data, device metadata, commercial data, retail data, and enterprise data, can now be used because of the various enabling technologies that complement IoT, including big data, cloud, semantic web, and data storage technologies.

7.4 Value Added Applications

Deep learning [57], machine learning [58], and artificial intelligence [59] are key technologies that provide value added applications using IoT and big data. Before the emergence of IoT and cloud computing, massive amounts of data and computation power are unavailable for certain applications, thereby preventing them from using such technologies. Different data analytics platforms [60], business intelligence platforms [61], visualization applications [62], and analytics applications [63] have recently emerged and helped industries and organizations transform their operations, improve their productivity and diagnostics, and increase their agility. Such level of detail was not possible before the emergence of IoT.

8 Open Research Challenges

IoT systems have the potential to solve many problems, but numerous challenges remain unaddressed. The solutions to some of these challenges are yet to be provided by big data and analytics solutions themselves, while others require concentrated efforts from the IoT community, hardware and platform vendors, governments, and policy makers.

8.1 Exploiting the Temporal Usefulness of IoT

IoT data have a profound impact on the digitized world. However, these data have a temporal aspect that can be useful in making real-time decisions, improving quality, and providing an excellent user experience. For example, a consumer-oriented organization can combine available consumer data with daily parking lot

occupation data to offer incentives to their customers or manage their inventory in a proactive manner on a daily or seasonal basis. In typical IoT solutions, the insights from the IoT data are often either time consuming or not put into use immediately. This trend changes into a proactive one to make correlations, derive insights, and find seasonal, emerging, and diminishing patterns using IoT data [64]. In many critical industrial applications, these correlations, insights, and patterns can help increase operational efficiency and achieve effective control in real time. Therefore, we must implement solutions that can handle data at the device or gateway level where the IoT data from devices, sensors, and processes are initially received. Exploiting semantically annotated data [65] or using a rules engine to locally process information are potential avenues to explore in future research. Applying semantics is particularly useful because of its capability to provide the required abstractions, whereas annotated data still retain their semantics when pushed to IoT/cloud platforms for analytics.

8.2 Adding Semantics to IoT Data

The usefulness of any type of data can be enhanced by adding metadata to their contexts and meanings. This practice is particularly important in IoT by helping users process and utilize heterogeneous IoT data at the device, gateway, and cloud levels with different scopes and granularities. One option is to base the solutions on their ontology, which is a formal representation of concepts and the relationships among these concepts. Therefore, ontology can be used to create vocabularies of metadata for annotating IoT data at the source or near the source. Given that ontologies are very easy to share and link, they can provide the right context and meanings of IoT data in an open manner. Ontologies are also useful for integrating IoT data from multiple domains [66]. Although several efforts have been made to create general and domain-specific ontologies, more efforts are required in some areas, especially in the industrial world, to create specific ontologies for linking and sharing data across different domains and businesses. Another option is to use non-ontology vocabularies, such as the Haystack project [67], which focuses on defining metadata tags for annotation in the building automation domain.

However, this option lacks integration with other similar vocabularies. One may also encounter several disjoint vocabularies unless they evolve into ontologies that can be linked and shared across domains. Another option is to use open standards, such as the one from Hypercat consortium [68] that uses a standard catalogue format to encode metadata as RDF triples and link them together by using URLs. However, such efforts are yet to be extended to the global level.

8.3 Diversity Issues

The IoT paradigm has heterogeneous protocols, standards, and platforms. The industrial world also faces IT and OT integration issues that demonstrate much technological fragmentation. The current protocols have several initiatives, including CoAP, MQTT, XMPP, DDS, STOMP, HTTP, and AMQP. Although the IoT paradigm does not have a universal protocol, multiple protocols may co-exist because of the different requirements and their intended uses. Therefore, IoT systems may be unable to support multiple protocols in an extensible way. Intelligent gateway solutions, such as that proposed in [69], must provide seamless integration and interoperability between various protocols. In terms of standards, several organizations, such as ITU-T, IETF, ISO/IEC, IEEE, ETSI, oneM2M, and 3GPP, have shown some efforts. While we may assume that all these standardization activities will provide interoperability (or some form of it), they may lead to a higher ambiguity because instead of having a broad scope, they all provide specific and isolated solutions that only cover their own domains [70].

In terms of IoT platforms, several initiatives have been launched to generate profit from IoT by providing connectivity, data storage, big data analysis, predictions, and machine learning. The big industry players have achieved

much progress in offering diverse IoT platforms with a rich feature set. IBM Watson, Microsoft Azure, GE Predix, Cisco Jasper, and PTC ThingWorx are examples of enterprise-grade platforms that face a vendor lockdown. Open source IoT platform initiatives, such as thingsboard.io, Kaa, and DeviceHive, are few good examples in this regard.

8.4 Security Challenges

A major hindrance in the broad integration of IoT in industries lies in its security. Several challenges, such as the recent Dyn attack [71], underscore the importance of having secure IoT devices, platforms, and applications which otherwise can lead to major catastrophes, such as the successful execution of a massive DDOS attack. These attacks can have devastating effects on the businesses of many critical industries, threaten national security, and even directly or indirectly affect human lives. The IT professionals in these industries have their hands full with the security issues of BYOD [72], [73] and the implementation of on-site cloud infrastructures in their organizations. Therefore, IoT security issues only add to their worries. Security is also not the first topic in the current IoT discussions and is still largely treated as a compulsory yet secondary subject. Such disregard can be attributed to the lack of organizational policies and the ambiguities in government laws [74]. To guarantee a successful implementation of IoT, solving these security issues must be given priority in the IoT realm. These issues not only require technical solutions but also the appropriate enforcement of policies and guidelines. The views of all stakeholders in IoT must also be considered.

8.5 Data Management Issues

IoT data are valuable assets. With the exponential increase in the number of IoT devices, systems, and processes, new approaches, such as Data Lakes [75], have emerged to handle big data. Data Lakes stores structured and unstructured data without any pre-conceived notion of how these data will be used afterward. This approach does not have apply scheme mapping or query languages and can store any data without restrictions. However, Data Lakes introduces few problems. First, given that any data can be inserted, data swaps may occur in the future [76]. To avoid such problem, we must have oversights for data quality, impose metadata inclusion, and ensure data provenance. Second, using Data Lakes may lead to a loss of agility, which is especially true for large organizations that intend to use a large pool of data for quick analysis and decision making yet are unable to do so efficiently because they must go through several steps before extracting something meaningful from the data. These organizations must instead make a clear distinction between those data that can be used for decision making in near real time and those data that can be used to derive business strategies. The latter data type is more suitable for storage in Data Lake because these data will not be used immediately.

8.6 Data Provenance

Data provenance is linked to the authenticity and integrity of the data as well to their traceability to determine the owners and modifiers of the data at each step [77]. However, given that big data provides deep insights and analytics that may lead to some form of autonomous actuation in the real world, we must ensure that the data used for making such actuation are coming from a legitimate source. Several large-scale initiatives, including smart cities and smart health, plan to make use of big data and analytics, thereby making this issue even more critical. Although much of the current studies on IoT have focused on data management, only few have tried to address the data provenance issue, such as [78]. Having the ability to trace data ownership in IoT can be beneficial for monetization purposes when different actors share their data [79]. Existing studies in the IoT domain, such as [80], can be used as basis for devising technical solutions to this issue.

8.7 Data Governance and Regulation

One critical aspect of IoT data is related to data governance and regulating its use by different entities. Providing unsupervised or uncontrolled access to data introduces privacy concerns and hampers the participation of private owners, such as citizens who share their data from the sensors installed in their homes or in public places for monitoring purposes [81]. We must provide IoT device owners with options and tools to specify their preferences and prioritize/limit the use of data from their devices [82]–[84]. Future studies must also focus on developing policy frameworks to identify the stakes and concerns of data owners, data consumers, and all the other actors between these two. The input from regulatory authorities or governments will be necessary, but care must be taken to not have centralized control over the data. Data owners must be given more power to allow them to make decisions within the scope of the overall policy framework. The general public must be made aware of their role and must be given easy-to-use tools for sharing their data with other parties.

9 Conclusion

IoT is one of the biggest sources of big data, which are rendered useless without analytics power. IoT interacts with big data when voluminous amounts of data are needed to be processed, transformed, and analyzed in high frequency. This work specifically focuses on the big data context. First, we investigate the recent literature on big data processing and analytics solutions for IoT. Second, we identify the numerous requirements for big data and analytics in IoT. Third, we taxonomized the literature. Fourth, we determine the various opportunities that are brought about by big data. Fifth, we highlight the role of data analytics in IoT applications. Sixth, we present the open research challenges that must be addressed in the future. Seventh, we conclude that the existing big data solutions in the IoT paradigm are still in their infancy and the challenges associated with them must be solved in the future.

Acknowledgment

This work is supported by the Deanship of Scientific Research at King Saud University through Imran's Research Group No. (RG # 1435-051)

References

[1] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data: From beginning to future," *International Journal of Information Management*, vol. 36, no. 6, pp. 1231–1247, 2016.

[2] F. J. Riggins and S. F. Wamba, "Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics," in *Proceedings of 48th Hawaii International Conference on System Sciences (HICSS'15)*. IEEE, 2015, pp. 1531–1540.

[3] M. R. Bashir and A. Q. Gill, "Towards an iot big data analytics framework: Smart buildings systems," in *High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on*. IEEE, 2016, pp. 1325–1332.

[4] C. Lee, C. Yeung, and M. Cheng, "Research on iot based cyber physical system for industrial big data analytics," in *Industrial Engineering and Engineering Management (IEEM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1855–1859.

[5] P. Rizwan, K. Suresh, and M. R. Babu, "Real-time smart traffic management system for smart cities by using internet of things and big data," in *Emerging Technological Trends (ICETT), International Conference on*. IEEE, 2016, pp. 1–7.

[6] Q. Zhang, X. Zhang, Q. Zhang, W. Shi, and H. Zhong, "Firework: Big data sharing and processing in collaborative edge environment," in *Hot Topics in Web Systems and Technologies (HotWeb), 2016 Fourth IEEE Workshop on*. IEEE, 2016, pp. 20–25.

[7] M. M. Rathore, A. Ahmad, and A. Paul, "Iot-based smart city development using big data analytical approach," in *Automatica (ICA-ACCA), IEEE International Conference on*. IEEE, 2016, pp. 1–8.

[8] B. Ahlgren, M. Hidell, and E. C.-H. Ngai, "Internet of things for smart cities: Interoperability and open data," *IEEE Internet Computing*, vol. 20, no. 6, pp. 52–56, 2016.

[9] O. B. Sezer, E. Dogdu, M. Ozbayoglu, and A. Onal, "An extended iot framework with semantics, big data, and analytics," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1849–1856.

[10] B. Cheng, A. Papageorgiou, F. Cirillo, and E. Kovacs, "Geelytics: Geo-distributed edge analytics for large scale iot systems based on dynamic topology," in *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*. IEEE, 2015, pp. 565–570.

[11] H. Wang, O. L. Osen, G. Li, W. Li, H.-N. Dai, and W. Zeng, "Big data and industrial internet of things for the maritime industry in northwestern norway," in *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, 2015, pp. 1–5.

[12] J. L. Pérez and D. Carrera, "Performance characterization of the servioticy api: an iot-as-a-service data management platform," in *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*. IEEE, 2015, pp. 62–71.

[13] M. Villari, A. Celesti, M. Fazio, and A. Puliafito, "Alljoyn lambda: An architecture for the management of smart environments in iot," in *Smart Computing Workshops (SMARTCOMP Workshops), 2014 International Conference on*. IEEE, 2014, pp. 9–14.

[14] A. J. Jara, D. Genoud, and Y. Bocchi, "Big data for cyber physical systems: an analysis of challenges, solutions and opportunities," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on*. IEEE, 2014, pp. 376–380.

[15] Z. Ding, X. Gao, J. Xu, and H. Wu, "Iot-statisticdb: a general statistical database cluster mechanism for big data analysis in the internet of things," in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE, 2013, pp. 535–543.

[16] C. Vuppalapati, A. Ilapakurti, and S. Kedari, "The role of big data in creating sense ehr, an integrated approach to create next generation mobile sensor and wearable data driven electronic health record (ehr)," in *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 293–296.

[17] A. Ahmad, M. M. Rathore, A. Paul, and S. Rho, "Defining human behaviors using big data analytics in social internet of things," in *Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on*. IEEE, 2016, pp. 1101–1107.

[18] E. Ahmed and M. H. Rehmani, "Introduction to the special section on social collaborative internet of things," p. 382384, 2017.

[19] D. Arora, K. F. Li, and A. Loffler, "Big data analytics for classification of network enabled devices," in *Advanced Information Networking and Applications Workshops (WAINA), 2016 30th International Conference on*. IEEE, 2016, pp. 708–713.

[20] I.-L. Yen, G. Zhou, W. Zhu, F. Bastani, and S.-Y. Hwang, "A smart physical world based on service technologies, big data, and game-based crowd sourcing," in *Web Services (ICWS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 765–772.

[21] R. P. Minch, "Location privacy in the era of the internet of things and big data analytics," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2015, pp. 1521–1530.

[22] A. Mukherjee, H. S. Paul, S. Dey, and A. Banerjee, "Angels for distributed analytics in iot," in *Internet of Things (WF-IoT), 2014 IEEE World Forum On*. IEEE, 2014, pp. 565–570.

[23] A. Mukherjee, S. Dey, H. S. Paul, and B. Das, "Utilising condor for data parallel analytics in an iot context-an experience report," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2013 IEEE 9th International Conference on*. IEEE, 2013, pp. 325–331.

[24] H. R. Arkian, A. Diyanat, and A. Pourkhalili, "Mist: Fog-based data analytics scheme with cost-efficient resource provisioning for iot crowdsensing applications," *Journal of Network and Computer Applications*, vol. 82, pp. 152–165, 2017.

[25] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Computer Networks*, vol. 101, pp. 63–80, 2016.

[26] F. Alam, R. Mehmood, I. Katib, and A. Albeshri, "Analysis of eight data mining algorithms for smarter internet of things (iot)," *Procedia Computer Science*, vol. 98, pp. 437–442, 2016.

[27] M. H. Berlian, T. E. R. Sahputra, B. J. W. Ardi, L. W. Dzatmika, A. R. A. Besari, R. W. Sudibyo, and S. Sukaridhoto, "Design and implementation of smart environment monitoring and analytics in real-time system framework based on internet of underwater things and big data," in *Electronics Symposium (IES), 2016 International*. IEEE, 2016, pp. 403–408.

[28] D. Mourtzis, E. Vlachou, and N. Milas, "Industrial big data as a result of iot adoption in manufacturing," *Procedia CIRP*, vol. 55, pp. 290–295, 2016.

[29] R. Ramakrishnan and L. Gaur, "Smart electricity distribution in residential areas: Internet of things (iot) based advanced metering infrastructure and cloud analytics," in *Internet of Things and Applications (IOTA), International Conference on*. IEEE, 2016, pp. 46–51.

[30] E. Ahmed and M. H. Rehmani, "Mobile edge computing: Opportunities, solutions, and challenges," pp. 59–63.

[31] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Intelligent Systems and Control (ISCO), 2016 10th International Conference on*. IEEE, 2016, pp. 1–8.

[32] U. Shaukat, E. Ahmed, Z. Anwar, and F. Xia, "Cloudlet deployment in local wireless networks: Motivation, architectures, applications, and open challenges," *Journal of Network and Computer Applications*, vol. 62, pp. 18–40, 2016.

[33] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[34] J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi, "Big data analysis using apache hadoop," in *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*. IEEE, 2013, pp. 700–703.

[35] I. A. T. Hashem, N. B. Anuar, A. Gani, I. Yaqoob, F. Xia, and S. U. Khan, "Mapreduce: Review and open challenges," *Scientometrics*, pp. 1–34, 2016.

[36] V. Morabito, "Managing change for big data driven innovation," in *Big Data and Analytics*. Springer, 2015, pp. 125–153.

[37] A. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, and A. G. Parameswaran, "Datahub: Collaborative data science & dataset version management at scale," *arXiv preprint arXiv:1409.0798*, 2014.

[38] 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 Record*, vol. 40, no. 4, pp. 45–51, 2012.

[39] S. Burke, "Hp haven big data platform is gaining partner momentum," *CRN [online] http://www. crn. com/news/applications-os/240161649*, 2013.

[40] (2017, Accessed on 3rd June) Hortonworks. [Online]. Available: https://hortonworks.com/

[41] Y. Zhuang, Y. Wang, J. Shao, L. Chen, W. Lu, J. Sun, B. Wei, and J. Wu, "D-ocean: an unstructured data management system for data ocean environment," *Frontiers of Computer Science*, vol. 10, no. 2, pp. 353–369, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11704-015-5045-6

[42] D. Slezak, P. Synak, J. Wróblewski, and G. Toppin, "Infobright analytic database engine using rough sets and granular computing," in *Granular Computing (GrC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 432–437.

[43] (2017, Accessed on 3rd June) Mapr. [Online]. Available: https://mapr.com/

[44] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," *Journal of Internet Services and Applications*, vol. 6, no. 1, p. 1, 2015.

[45] E. Ahmed, M. Imran, M. Guizani, A. Rayes, J. Lloret, G. Han, and W. Guibene, "Enabling mobile and wireless technologies for smart cities: Part 2," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 12–13, 2017.

[46] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, S. Halunga, and O. Fratu, "Big data, internet of things and cloud convergence–an architecture for secure e-health applications," *Journal of medical systems*, vol. 39, no. 11, pp. 1–8, 2015.

[47] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, "Network architecture and qos issues in the internet of things for a smart city," in *Communications and Information Technologies (ISCIT), 2012 International Symposium on*. IEEE, 2012, pp. 956–961.

[48] R. Tönjes, P. Barnaghi, M. Ali, A. Mileo, M. Hauswirth, F. Ganz, S. Ganea, B. Kjærgaard, D. Kuemper, S. Nechifor *et al.*, "Real time iot stream processing and large-scale data analytics for smart city applications," in *poster session, European Conference on Networks and Communications*, 2014.

[49] E. Ahmed, S. Ali, A. Akheenzada, and I. Yaqoob, "Cognitive radio sensor networks: Bridging the gap for network," *Cognitive Radio Sensor Networks: Applications, Architectures, and Challenges: Applications, Architectures, and Challenges. IGI Global*, p. 160, 2014.

[50] S. A. A. Shah, E. Ahmed, F. Xia, A. Karim, M. A. Qureshi, I. Ali, and R. M. Noor, "Coverage differentiation based adaptive tx-power for congestion and awareness control in vanets," *Mobile Networks and Applications*, pp. 1–12.

[51] I. Yaqoob, I. Ahmad, E. Ahmed, A. Gani, M. Imran, and N. Guizani, "Overcoming the key challenges to establishing vehicular communication: Is sdn the answer?" *IEEE Communications Magazine*, 2017.

[52] I. Yaqoob, E. Ahmed, I. A. T. Hashem, A. Ahmed, A. Gani, M. Imran, and M. Guizani, "Internet of things architecture: Recent advances, taxonomy, requirements, and open challenges," *IEEE Wireless Communications*, 2017.

[53] Y. Jararweh, A. Doulat, O. AlQudah, E. Ahmed, M. Al-Ayyoub, and E. Benkhelifa, "The future of mobile cloud computing: integrating cloudlets and mobile edge computing," in *Telecommunications (ICT), 2016 23rd International Conference on*. IEEE, 2016, pp. 1–5.

[54] N. Bessis and C. Dobre, *Big data and internet of things: a roadmap for smart environments*. Springer, 2014.

[55] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, 2016.

[56] A. Brring, S. Schmid, C. K. Schindhelm, A. Khelil, S. Kbisch, D. Kramer, D. L. Phuoc, J. Mitic, D. Anicic, and E. Teniente, "Enabling iot ecosystems through platform interoperability," *IEEE Software*, vol. 34, no. 1, pp. 54–61, Jan 2017.

[57] X. W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[58] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.

[59] O. Etzion, "When artificial intelligence meets the internet of things," in *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems*. ACM, 2015, pp. 246–246.

[60] V. O. Safonov, "Example of a trustworthy cloud computing platform in detail: Microsoft azure," *Trustworthy Cloud Computing*, pp. 147–270, 2016.

[61] J. Vidal-García, M. Vidal, and R. H. Barros, "Computational business intelligence, big data, and their role in business decisions in the age of the internet of things," in *The Internet of Things in the Modern Business Environment*. IGI Global, 2017, pp. 249–268.

[62] Y. Jeong, H. Joo, G. Hong, D. Shin, and S. Lee, "Aviot: Web-based interactive authoring and visualization of indoor internet of things," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 3, pp. 295–301, 2015.

[63] M. Strohbach, H. Ziekow, V. Gazis, and N. Akiva, "Towards a big data analytics framework for iot and smart city applications," in *Modeling and processing for next-generation big-data technologies*. Springer, 2015, pp. 257–282.

[64] S. Aljawarneh, V. Radhakrishna, P. V. Kumar, and V. Janaki, "A similarity measure for temporal pattern discovery in time series data generated by iot," in *Engineering & MIS (ICEMIS), International Conference on*. IEEE, 2016, pp. 1–4.

[65] C. El Kaed, I. Khan, H. Hossayni, and P. Nappey, "Sqeniot: Semantic query engine for industrial internet-of-things gateways," *Submitted IEEE GLOBECOM*, 2016.

[66] T. Banerjee and A. Sheth, "Iot quality control for data and application needs," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 68–73, 2017.

[67] A. A. Bhattacharya, D. Hong, D. Culler, J. Ortiz, K. Whitehouse, and E. Wu, "Automated metadata construction to support portable building applications," in *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 2015, pp. 3–12.

[68] T. Jaffey, J. Davies, and P. Beart, "Hypercat 3.00 specification," *Hyper-cat Limited*, 2016.

[69] P. Desai, A. Sheth, and P. Anantharam, "Semantic gateway as a service architecture for iot interoperability," in *Mobile Services (MS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 313–319.

[70] A. Meddeb, "Internet of things standards: who stands out from the crowd?" *IEEE Communications Magazine*, vol. 54, no. 7, pp. 40–47, 2016.

[71] C. Chris Mishler and C. CIA, "The future of the internet of things," *Strategic Finance*, vol. 97, no. 5, p. 62, 2015.

[72] S. Tanimoto, S. Yamada, M. Iwashita, T. Kobayashi, H. Sato, and A. Kanai, "Risk assessment of byod: Bring your own device," in *Consumer Electronics, 2016 IEEE 5th Global Conference on*. IEEE, 2016, pp. 1–4.

[73] K. Hajdarevic, P. Allen, and M. Spremic, "Proactive security metrics for bring your own device (byod) in iso 27001 supported environments," in *Telecommunications Forum (TELFOR), 2016 24th*. IEEE, 2016, pp. 1–4.

[74] V. A. Almeida, D. Doneda, and J. de Souza Abreu, "Cyberwarfare and digital governance," *IEEE Internet Computing*, vol. 21, no. 2, pp. 68–71, 2017.

[75] H. Fang, "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem," in *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on*. IEEE, 2015, pp. 820–824.

[76] R. Hai, S. Geisler, and C. Quix, "Constance: An intelligent data lake system," in *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016, pp. 2097–2100.

[77] W. She, I.-L. Yen, F. Bastani, B. Tran, and B. Thuraisingham, "Role-based integrated access control and data provenance for soa based net-centric systems," in *Service Oriented System Engineering (SOSE), 2011 IEEE 6th International Symposium on*. IEEE, 2011, pp. 225–234.

[78] B. Glavic, "Big data provenance: Challenges and implications for benchmarking," in *Specifying big data benchmarks*. Springer, 2014, pp. 72–80.

[79] Y. Zhang and J. Wen, "An iot electric business model based on the protocol of bitcoin," in *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015, pp. 184–191.

[80] Q. H. Cao, I. Khan, R. Farahbakhsh, G. Madhusudan, G. M. Lee, and N. Crespi, "A trust model for data sharing in smart cities," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–7.

[81] I. Khan, F. Belqasmi, R. Glitho, N. Crespi, M. Morrow, and P. Polakos, "Wireless sensor network virtualization: Early architecture and research perspectives," *IEEE Network*, vol. 29, no. 3, pp. 104–112, 2015.

[82] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 31–40, 2016.

[83] C. H. Suen, R. K. Ko, Y. S. Tan, P. Jagadpramana, and B. S. Lee, "S2logger: End-to-end data tracking mechanism for cloud data provenance," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*. IEEE, 2013, pp. 594–602.

[84] M. B. Jones, B. Ludäscher, T. McPhillips, P. Missier, C. Schwalm, P. Slaughter, D. Vieglais, L. Walker, and Y. Wei, "Dataone: A data federation with provenance support," in *Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings*, vol. 9672. Springer, 2016, p. 230.

Ejaz Ahmed worked at Centre for Mobile Cloud Computing Research (C4MCCR), University of Malaya, Malaysia. Before that, he has worked as Research Associate in CogNet (Cognitive Radio Network) Research Lab SEECS, NUST Pakistan from December 2009 to September 2012, and in CoReNet (Center of Research in Networks and Telecom), CUST, Pakistan, from January 2008 to December 2009. His research experience spans over more than ten years. He is associate editor of *IEEE Communication Magazine*, *IEEE Access*, and *Wiley Wireless Communications and Mobile Computing*, *Elsevier Journal of Network and Computer Applications*, and *KSII TIIS*. He has also served as a Lead Guest Editor/Guest Editor and Chair/Co-chair in international journals and international conferences, respectively. His areas of interest include Mobile Cloud Computing, Mobile Edge Computing, Internet of Things, Cognitive Radio Networks, and Smart Cities. He has successfully published his research work in more than fifty international journals and conferences.

Ibrar Yaqoob received his Ph.D. degree in Computer Science from the University of Malaya, Malaysia, in 2017. He earned 550 plus citations, and 50 plus impact factor during his Ph.D. candidature. He worked as a researcher at Centre for Mobile Cloud Computing Research (C4MCCR), University of Malaya. His research experience spans over more than three and half years. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are very famous and among the most downloaded in top journals. His research interests include big data, mobile cloud, the Internet of Things, cloud computing, and wireless networks.

Ibrahim Abaker Targio Hashem received his Ph.D. degree in Computer Science from the University of Malaya, Malaysia, in 2017. He received his M.S. degree in computing in 2012, Malaysia, and the B.E. degree in computer science in 2007, Sudan. Hashem obtained professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation). He worked as a Tutor at CISCO Academy, University of Malaya. His main research interests include big data, cloud computing, distributed computing, and network.

Imran Khan is working as innovation project leader in Schneider Electric. He is leading the design and specification of data and information management architectures for sustainable energy management in various industrial domains. He received Ph.D. degree in Computing and Networks from Institut Mines-Tlcom, Tlcom SudParis jointly with UPMC Paris VI, France, M.S. degree in Multimedia and Communication from M.A. Jinnah University, Pakistan and B.S. degree in Computer Science from COMSATS Institute of IT, Pakistan. During his Ph.D. he worked as collaborating researcher at Concordia University, Montreal, Canada to lead a 3 year Cisco funded project. He was also involved in several European research projects funded by ITEA2 and H2020. During M.S. Imran was member of Center of Research in Networks and Telecom (CoReNeT) and worked on projects funded by the French Ministry of Foreign Affairs and the Internet Society (ISOC). He has number of publications in peer reviewed conferences and journals, and has also contributed to the IETF standardization activities. His current research interests are Internet of Things (IoT), data and information management using semantic web technologies, cloud and edge computing, software defined automation and wireless sensor networks.

Abdelmuttlib Ibrahim Abdalla Ahmed (abdelmuttlib@siswa.um.edu.my) received his B.Sc. degree in computer science from OIU, Sudan, and his M.S. degree in computer science from IIUI, Pakistan. He is currently pursuing a Ph.D. degree at the University of Malaya. His research Interest areas include trust and reputation systems, security and digital forensics, Internet of Things, mobile and cloud computing, and vehicular networks.

Muhammad Imran (cimran@ksu.edu.sa) is an assistant professor in the College of Computer and Information Science, King Saud University. His research interests include mobile ad hoc and sensor networks, WBANs, IoT, M2M, multihop wireless networks, and fault-tolerant computing. He has published a number of research papers in peer reviewed international journals and conferences. His research is financially supported by several grants. He is serving as a Co-Editor-in-Chief for EAI Transactions on Pervasive Health and Technology. He also serves as an Associate Editor for the Wireless Communication and Mobile Computing Journal (Wiley), the Inderscience International Journal of Autonomous and Adaptive Communications Systems, Wireless Sensor Systems (IET), and the International Journal of Information Technology and Electrical Engineering. He has served/serves as a Guest Editor for IEEE Communications Magazine, IJAACS, and the International Journal of Distributed Sensor Networks. He has been involved in a number of conferences and workshops in various capacities such as a Program Co-Chair, Track Chair/Co-Chair, and Technical Program Committee member. These include IEEE GLOBECOM, ICC, AINA, LCN, IWCMC, IFIP WWIC, and BWCCA. He has received a number of

awards such as an Asia Pacific Advanced Network fellowship.

Athanasios V. Vasilakos currently Professor at Lulea University of Technology, Sweden. He is also General Chair of the European Alliances for Innovation. His research interests include Cloud Computing, Smart Grid, Energy Security and Harvesting, Social Networks, Mobile/Wireless Networks, IoT, Sensor Networks. He has authored or coauthored over 250 technical papers in major international journals and conferences. Moreover, he is author/co-author of five books and more than 20 book chapters. He served or is serving as an Editor or/and Guest Editor for many technical journals, such as the IEEE Transactions on Network and Service Management, IEEE Transactions on Cloud Computing, IEEE Transactions on Cybernetics, IEEE Transactions on Information Forensics and Security. Moreover, he has served as General Chair, Technical Program Committee Chair for many international conferences.