Leveraging Big Data Analytics in 5G-Enabled IoT and Industrial IoT for the Development of Sustainable Smart Cities

Article *in* Transactions on Emerging Telecommunications Technologies · July 2022

DOI: 10.1002/ett.4618

4 authors:

Suprakash Mukherjee
Birla Institute of Technology and Science Pilani

Shashank Gupta
Birla Institute of Technology and Science Pilani

Oshin Rawlley
Birla Institute of Technology and Science Pilani

Siddhant Jain
Birla Institute of Technology and Science Pilani

Received: 17 November 2021 | Revised: 12 July 2022 | Accepted: 12 July 2022

DOI: 10.1002/ett.4618

SURVEY ARTICLE

WILEY

Leveraging big data analytics in 5G-enabled IoT and industrial IoT for the development of sustainable smart cities

Suprakash Mukherjee | Shashank Gupta | Oshin Rawlley | Siddhant Jain

Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, Rajasthan, India

Correspondence
Shashank Gupta, Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, Rajasthan, India.
Email:
shashank.gupta@pilani.bits-pilani.ac.in

Abstract

There has been an exponential growth in the number of low-cost heterogeneous sensor devices that are connected to the internet in the existing infrastructure of smart cities in the past decade. These sensors and actuator devices are employed in various industries to collect invaluable data that greatly impacts business decisions. State-of-the-art research is being carried out to process the high-volume data collected in high velocity streams with high variability in order to draw meaningful insights to cater to business needs in various domains. Big data analytics finds diversified opportunities in 5G-enabled Internet of Things (IoT) and industrial IoT environment to study the data patterns and produce new results which provides big organizations a conducive environment to take informed decisions. This article presents an exhaustive investigation of the various applications and algorithms of the big data analytics in 5G-driven IoT and industrial IoT systems with a detailed taxonomy of the existing analytical systems and also the challenges specific to the applications in an IoT environment. A holistic understanding on the importance of security and privacy of big data during the development of smart city infrastructure in high velocity streaming process has been explained for a broad understanding of the big data streaming processes. A thorough analysis of important data mining algorithms such as classification, association rule mining, clustering, and prediction methods for big IoT data from the recent literature also has been presented to provide a clear understanding of existing issues in handling big data in 5G-enabled IoT for the development of sustainable smart city infrastructure.

1 | INTRODUCTION

Due to the tremendous technological advances in the recent years in micro-electromechanical systems (MEMS), we have seen an upsurge in the use of low-cost sensors with increased communication and data processing capabilities in conventional Internet of Things (IoT). With low power requirement and multifunctional possibilities these wireless sensors have found various applications starting from health, security applications, commercial and environmental monitoring, industrial sensing, disaster prediction and diagnostics and context-aware applications like smart homes. These technological advancements have led to the advent of 5G-enabled IoT which include various objects that are connected to the Internet like smartphones, PCs, smart watches, cars and so forth.

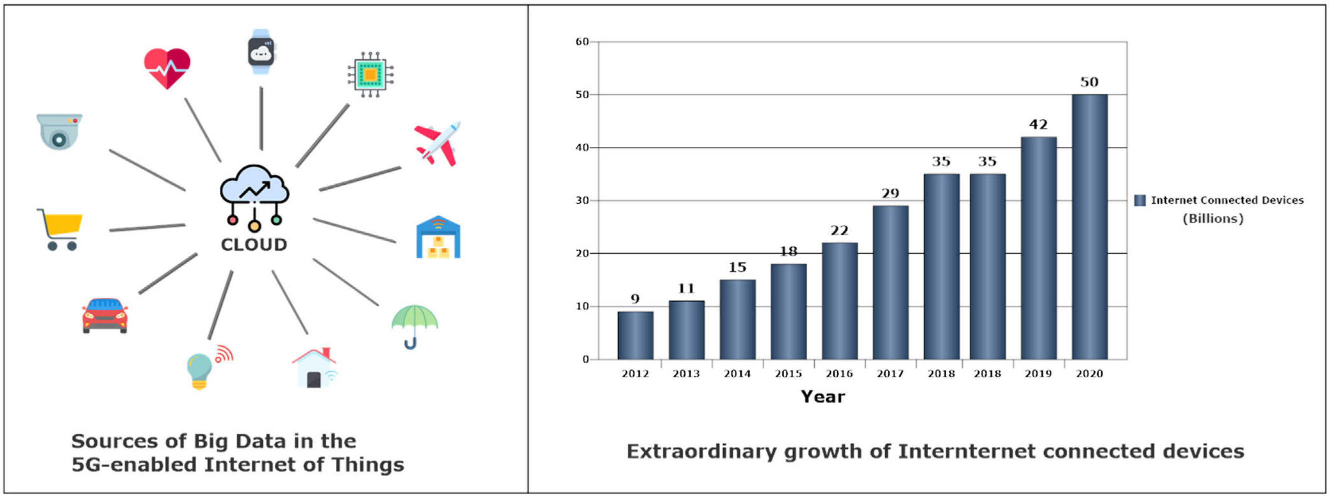


FIGURE 1 Sources of big data and growth of internet connected devices

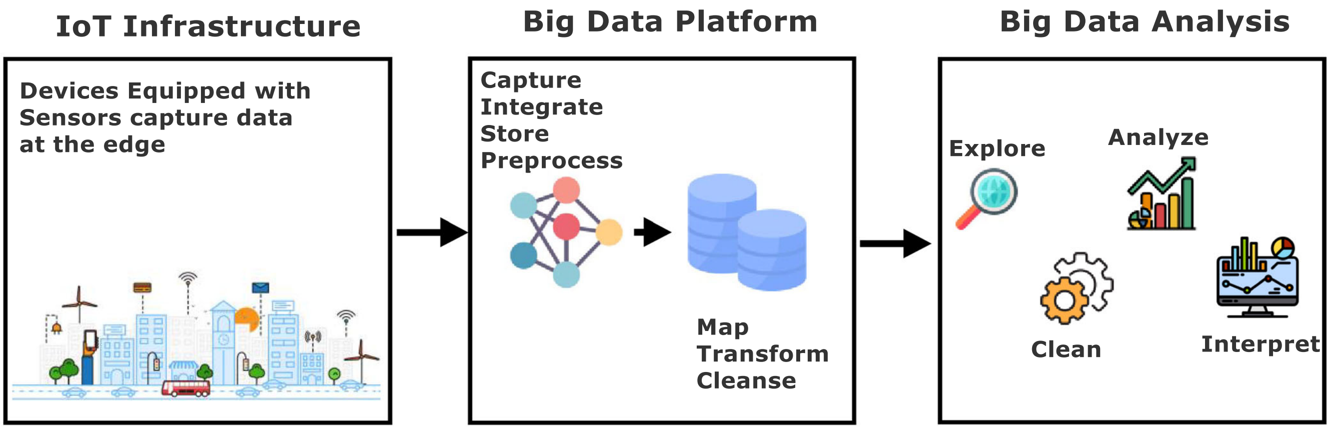


FIGURE 2 Big data analysis process overview

According to survey of Cisco, internet connected objects will outnumber humans. The internet connected components include household appliances, smartphones, sensors, PCs, tablets, and wearable devices. Figure 1 shows the various big data sources and we witness that the number of internet-connected devices has seen an extraordinary growth and large diverse datasets are produced by many internet-connected devices. But in order to make the data useful 5G-enabled big data analytic systems plays a key role in extracting invaluable information that finds its applications in various industries like precision agriculture, smart grid systems, IoT healthcare, smart transport and so forth. The incorporation of Cyber physical systems,[1] cloud computing environment,[2] intelligent automation in production lines,[3] concentric computing,[4] wireless technologies, IoT and big data analytics (BDA) have led to the advancements in Industrial IoT systems to deliver valuable insights from the data captured of various platforms. Unified association of BDA tools and techniques growing into the industrial IoT systems have shown considerable increase in profits and value creation and have created a significant impact in business transformation for the development of sustainable smart cities infrastructure.[5,6]

The rate at which raw data is generated and collected from various devices like sensor data and mobile data, it becomes challenging for the conventional database systems to cater to this issue. Figure 2 illustrates the big data gathering process, monitoring and storage systems, and finally the analysis phase. The IoT device data is divided into data streams which is analyzed on big data platform after pre-processing. It is important to understand the requirement of secure big data streaming from a varied range of devices producing heterogeneous data. Big data security solution approaches are put in place to tackle new and emerging challenges in the big data analytics workflow in the IoT environment. Data mining techniques can be employed in the BDA process under IoT environment in making predictions, pattern recognition, deriving hidden insights from big data and improve in decision making processes.[7] Data mining mechanisms are employed in solving specific problems as well as addressing generic data analytics use cases. Predicting customer shopping behavior and analyzing patterns on social media are some of the uses of big data mining and they apply to many more applications providing predictive and efficient descriptive solutions.

Existing literature review highlighting big data, its advantages and application in 5G-enabled IoT and IIoT systems are limited to specific aspects. None of the existing articles have addressed these issues related to big data analytics in IoT and IIoT system environments in a detailed format, together with discussing the importance of secure big data streaming and relevance of data mining algorithms and techniques in mining big data collected from IoT systems in a single publication. Table 1 presents a detailed contrast of the proposed survey with the existing literature surveys.

1.1 | Overview of 5G technology

Wireless technologies have seen an immense growth all around the world in the recent times. In this context the one of the most challenging and important topic of research in wireless communication: the fifth generation or 5G technology has a tremendous opportunity to be a game changing technology in future. Fifth generation technology standard (5G) is the successor to 4G networks, a technology standard for broadband cellular networks. At present, 4G networks are more prevalent in providing connectivity to mobile devices while 5G began deploying around the globe in 2019 with various cellular companies offering this network. However, according to the GSM Association, 5G is expected to have 1.7billion subscribers by the year 2025.[20] In 5G cellular networks each service area is divided into cells. Radio waves acts as a communication medium for the 5G devices to communicate with the telephone network and internet enabled by a local antenna in the cell. 5G provides faster speeds (up to 10 Gbits/s)[21] with larger bandwidth and also can connect heterogeneous devices while improving the quality of service (QoS). However, hardware devices which are having only 4G capability need to upgrade themselves in order to connect to the 5G network. Table 2 presents the key differences between 4G and 5G and how 4G stacks up against 5G. Considering these differences, we can say that IoT and big data analytics powered by 5G cellular network technology creates the possibility to make denser IoT networks by deploying more number of sensors over 4G. The impacting factors such as: lower latency, higher throughput, and traffic capacity with better spectral efficiency in 5G enabled IoT networks challenges the existing technology framework by providing massive data processing intelligently through optimized communication channels and higher bandwidth on a large scale. In addition to this, 5G technology also empowers efficient edge nodes management by having better network energy efficiency. All these advantages of the 5G cellular network reserves the ability to provide better and advanced applications in IoT, IIoT and hence, in the development of smarter cities and IoT environments.

1.2 | Key contributions of the survey

In this article the authors have presented an exhaustive investigation of the various applications, algorithms and architectures of big data analytics in 5G enabled IoT scenarios such as industrial IoT systems, building smart city infrastructure with big data streaming applications and how big data mining helps in intelligent decision making in the IoT environment. This article also discusses the existing analytical systems with a detailed taxonomy and the challenges specific to the applications in 5G enabled IoT environment. The various components of big data analytics in the 5G-IoT system such as data engineering, data preparation, data analytics, and pipeline management has been discussed. A holistic understanding on the importance of security and privacy of big data during the development of smart city infrastructure in high velocity streaming process has been explained with the help of a case study meeting Industry 4.0 requirement standards for comprehensive understanding of the big data streaming processes. Additionally, a thorough analysis of important data mining algorithms such as classification, association rule mining, clustering, and prediction methods for big IoT data from recent literature has been presented to provide a clear understanding of existing issues in handling big data in 5G-enabled IoT for the development of sustainable smart city infrastructure.

The key contributions of this survey article are listed below.

- This article provides a detailed and extensive review of present day advances in BDA and its applications in the 5G-enabled IoT architecture and outlines the opportunities in the 5G driven IoT space.

TABLE 1 Existing surveys

State of the art	Highlight of the survey	Advantages	Limitations
2021[8]	Reviews the adoption of SDN and network function virtualization (NFV) for securing the IoT network from emerging threats.	Provides a comprehensive survey on security solutions based on SDN, blockchain, NFV for IoT network security.	Limited to IoT security in SDN/NFV
2021[9]	Solves network scalability challenges using blockchain distributed network.	Proposes raft consensus algorithm to increase throughput of blockchain and also introduces zkLedger to solve privacy issues.	Does not provide cover details on big data mining, machine learning application in big data analytics.
2019[10]	Highlights various ways machine learning can be used in wearable wireless body area network (WBAN)	Focuses on the challenges and open research areas for deploying ML models in sensitive networks such as WBAN	Big data streaming and its impact in the IoT environment was not explored
2019[11]	verifying and validating schemes for big data streaming in IoT applications	Streaming and processing of heterogeneous big data in secured manner	Does not provide cover details on big data mining and its applications in big data analytics.
2018[12]	Incorporates big data applications in IoT environment	Big data technologies compared across existing latest techniques in IoT domain	Does not cover issues related to privacy
2018[13]	Deep learning and analysis in big data streams in IoT environment	Uses of deep learning approaches and primary research domains in IOT applications were explored	Application of security framework in big data streaming and its impact is not explored
2018[14]	Blockchain based secure mechanisms for deployment in IoT covered	Latest security challenges and the methods to address them explored	security criterion for big data not explored, limited to security challenges in IoT environment
2017[15]	Concerns of big data security issues studied in detail	Provides exhaustive exploration with respect to primary security concerns for smart city development	Big data streaming and its impact in the IoT environment was not explored
2017[16]	Big data security software platforms and methods investigated	Present big data framework explored in detail highlighting important security concerns	Big data security and privacy challenges were covered
2016[17]	Security attacks and its classification in IoT applications	Exhaustive study of layered based threats	Security threats in big data streaming in IoT application were not explored
2016[18]	Security infrastructure of IoT in cloud and eHealth scenarios reviewed	IoT risks and threats studied in different layers	security threats in big data streaming were not explored, impact of big data streaming not covered
2014[19]	Advantages of privacy in big data and challenges associated studied	Big data processing technologies and its operation was studied	Big data streaming security challenges not explored
Our survey	Presents BDA applications in 5G-enabled IoT and Industrial IoT environment, big data streaming security and use and relevance of big data mining mechanisms in the development of sustainable smart city infrastructures.	Exhaustive study of big data analytics in 5G-driven IoT and IIoT systems with comprehensive understanding of streaming architecture and data mining techniques presenting state-of-the-art technologies and emerging trends.	-

TABLE 2 Key differences between 4G and 5G

Parameter	4G technology standard	5G technology standard
Latency	10–50 ms	1 ms
Throughput	2 Gbps	20 Gbps
Traffic capacity	10 Mbps/m^2	1000 Mbps/m^2
Density	100 k connections/Km^2	1 M connections/Km^2
Spectral efficiency	30 bps/Hz	100 bps/Hz
Network energy efficiency	Baseline	15% savings

- This article elaborates about the BDA architecture in IIoT systems, the various technologies, new age opportunities and their research challenges that IIoT systems are susceptible to.
- The importance of secure big data streaming has been identified and a detailed taxonomy of streaming security has been presented highlighting the research challenges with secure streaming.
- Finally, the importance of data mining techniques in BDA in 5G-enabled IoT environment has been understood by studying various state-of-the art articles from recent literature focusing on big IoT data mining.

1.3 | Outline of the survey

Section 1 of the survey has already discussed the introduction to the topic, significance and impact of 5G enabled IoT systems and the key contributions and value additions of the article. Here, a detailed understanding of the various sections of the survey has been presented to provide the viewers an overview of the topics covered. Section 2 presents the applications of BDA in the internet of things environment. This section presents a detailed discussion on big data analytics architecture. A taxonomy has been presented for classifying the analytics systems and it also focuses on the future opportunities of BDA in 5G-enabled IoT systems. The Section 3 focuses deeper into industrial IoT systems and applications of BDA. Industrial IoT systems are crucial to the development of automated systems in industrial processes. This section further discusses on BDA architecture, technologies and mechanisms and the research challenges in industrial IoT environments. Then Section 4 includes a comprehensive study of big data streaming architecture and security concerns with big data streaming has been discussed. The study presents a bigdata-based healthcare case study to understand the issues with respect to security in big data streaming and the challenges have been discussed. Data mining techniques play a key role in the development of most strategies for big data analytics. Section 5 discusses the data mining mechanisms with detailed study on the various frameworks, architectures, and applications related to big data mining in IoT applications. It also highlights the challenges in big data mining applications in IoT systems. Table 3 lists down the various abbreviations used in the article and their definition.

2 | 5G-ENABLED IOT APPLICATIONS OF BDA

2.1 | IoT BD works

IoT has started automating the actuation of data-driven applications which are based on intelligent technologies. Data acquisition from the sensory devices and analyzing that data on the basis of cloud services are the activities which are driven by the competent technologies on the connected objects. This revolutionary automation of systems has exhibited numerous benefits in real life examples such as smart grids, smart transportation, smart inventory, smart healthcare and so forth. Nevertheless, recent state-of-the-art industrial surveys indicates that there are many data concerning issues which are accountable for sluggish development of IoT in past years. Since huge amount of data is produced through these sensory devices distributed over a larger area, the data-related challenges concerns its acquisition process, integration, storage, and processing processes. Consequently, many researchers focus on managing heterogeneous data in a

TABLE 3 Abbreviations table

IoT	Internet of things
MEMS	Micro-electromechanical systems
IIoT	Industrial Internet of Things
BDA	Big data analytics
ITS	Intelligent traffic system
SFS	Smart factory system
BD	Big data
BDS	Big data streaming
BDM	Big data mining
SFS	Smart factory system
DM	Data mining
LAN	Local area network
WAN	Wide area network
VPN	Virtual private network
SC	Security consideration
ACM	Access control and monitoring
CA	Classification approach
PC	Policy compliance
SSP	Smart security practice
GOV	Governance
RTA	Risk and threat assessment
PaaS	Platform as a service
IaaS	Infrastructure as a service
FOTA	Firmware over the air
CPS	Cyber physical systems

distributed environment having characteristics such as multiple source diverse data, dynamic data, semantically weak data, and inaccurate data.

Moreover, authors have also developed an end-to-end solution for disentangling issues of knowledge graphs in IoT under 5G environment. They have attempted to adopt techniques of block chain management and have smartly produced matching of the relevant concepts and their relations.

In addition, the emergence of IoT and 5G/6G has led to sophisticated communications with tera-hertz of high-speed data. The data communicated should be reliable but however faces some intruder attacks. In this sensitive setting the privacy and security of the data is a vexed issue. Many approaches for preserving the nature and confidentiality of the data are exercised to have secure data communication with seamless connectivity.

BDA has exponentially emerged as an important step in the IoT applications to create better strategic decision-making process.[22] "Connected things" is one of the key features of IoT that helps in efficient data acquisition. In order to avail fast real-time analytical information from unstructured data collected from internet connected group of devices, standard big data technology needs to be implemented and set up with proper infrastructure to handle high volume storage and high velocity processing of varied data. With readily available information quicker decisions and interaction can be facilitated which helps organizations in various ways. IoT and big data both have seen a compelling need in development and research to fuel the fields of business sector and IT.

Figure 3 shows that the relationship between big data and IoT systems can be divided into multiple steps for proper IoT data management. In the first step the IoT data source nodes (sensors) interact with various applications. Complex data generated in the second step has high volume, high velocity and high variability of the data called as big data. These

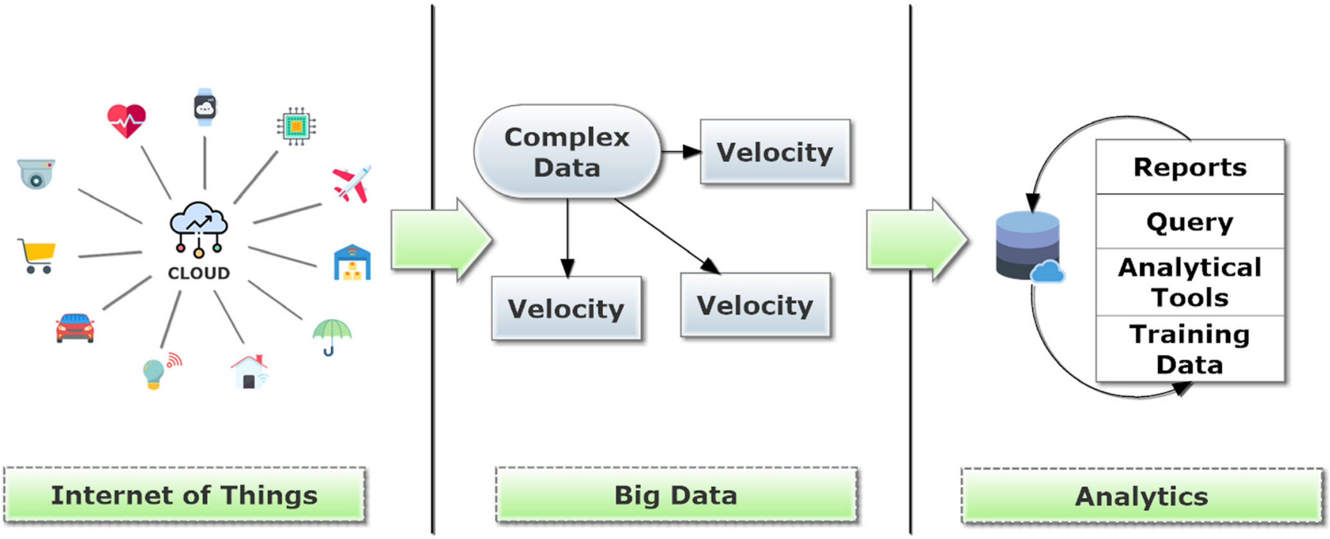


FIGURE 3 BD Analytics and IoT relationship

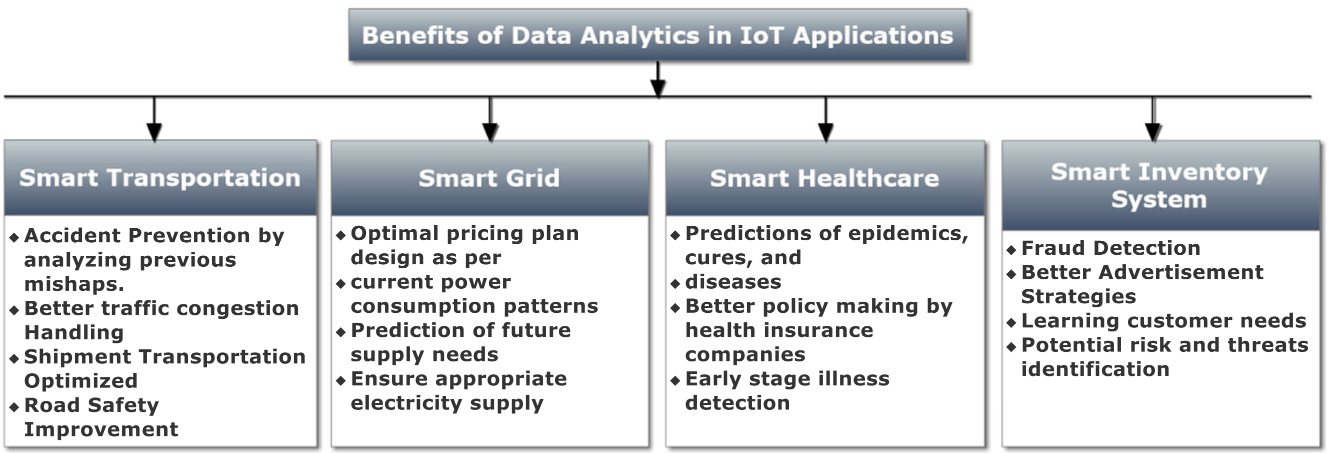


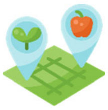
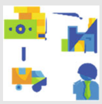
FIGURE 4 Data analytics in 5G-enabled IoT applications and perceived benefits

huge data chunks are stored in distributed storage solutions and analytical systems draw meaningful insights from them through the four layers of analytics. Enterprises are assisted by big data technologies in better decision-making processes with the use of BDA in the IoT infrastructure by providing computation and data storage services. IoT applications such as smart grid systems, smart transportation systems, smart healthcare, and smart inventory systems are a major source of big data.[23-25] Figure 4 summarizes the various applications of BDA in the IoT and their perceived benefits. For instance, intelligent transport system requires vehicular data to be processed in real-time for better road management. Similarly, smart healthcare services demand quick response with carrying highly-sensitive data where big-data analytics needs to be resilient for the reliable operations to function correctly.

Smart grids: Smart grids producing high volumes of data with high velocity, do have access to useful information from this data which is an important factor. In the intelligent grid environment, big data is collected from a variety of sources, some of which are user energy consumption patterns, phasor measurement, and smart meter systems.[26] Appropriate statistics can help in decision making processes to measure the right amount of electricity supply that must be provided to certain customers and predict the changes and spikes in power requirements by various kind of users. Prediction of requirements in the future is another application of big data. By employing big data technologies better strategic decisions can be taken in terms of pricing, by predicting the demand and supply.

Smart transportation: Acquiring data has become a very important task in this modern era of technology where cars and automobiles are connected to the internet and generate large amount of sensory data. BDA can help transport

TABLE 4 BD analytics applications in 5G-driven IoT and their utility

Application		Usefulness	Input type	IoT hardware	BD analytics applications
	Precision agriculture	Extract information like soil moisture level, trunk width of plants, weather forecasting and so forth.	Text and image sources	Sensors	Hadoop
	Smart supply chains	Effective decision making and efficient control leading to supply chain optimization.	Text and image sources	Mobile devices and sensors	Hadoop
	Smart grids	Enables real time monitoring and improves safety and reliability.	Text source	Sensors	Hadoop
	Smart transportation	Helps persons and objects in the transportation system take informed and automated decisions and create new efficiencies.	Text and video sources	Sensors, lidar, camera	Hadoop, Spark
	Smart traffic	Vehicle and pedestrian detection and accordingly control traffic using automated traffic lighting systems for example.	Image and video sources	Cameras	Hadoop, Spark
	Smart metering	Estimation of energy consumption patterns.	Text source	Sensors	Hadoop

authorities determine the cause of traffic accidents (eg, accident and speeding), reduce the number of road accidents, identifying peak traffic times, and produce optimal route plans and control congestion on the road. Intelligent transport data analytics can indirectly improve shipping movements, provide better safety measures, and improve the end user experience by decreasing travel time.

Smart inventory: Extracting valuable intelligence from big data collected from inventory systems can help enterprises in profit generation. Datasets from such systems help in analysis and gain information about market trends. By analyzing seasonal variation in data, it is possible to make product recommendations accurately. Fraud detection is also an application where inventory data analysis can help. Big data analytics also help in strategic advertising and decision-making processes to better understand customer and product relations. Potential threats and opportunities can be identified that aid companies in planning ahead.

Smart healthcare: A tonne of information has been generated by the health care industry in the past few years. Fast paced high volume data generation has led to several challenges in the storage, processing and gathering of useful intelligence related to healthcare. Data analytics can help in identifying diseases such as breast cancer, predicting possibilities of epidemic and its impacts based on historical data of patient details. Family physicians can better recommend treatments based on such data and this data can also be utilized by insurance companies to design policies that fit their business needs. Healthcare IoT and BDA can prove to be crucial in early detection of health problems thereby decreasing fatality rate.

Apart from the various domains where BDA and IoT technologies prove to be useful there are some very popular uses of big data analytics in Internet of Things environment. Precision agriculture, smart supply chain management, smart traffic systems, and smart metering are some of the common use cases for big data analytics in IoT environment. Table 4

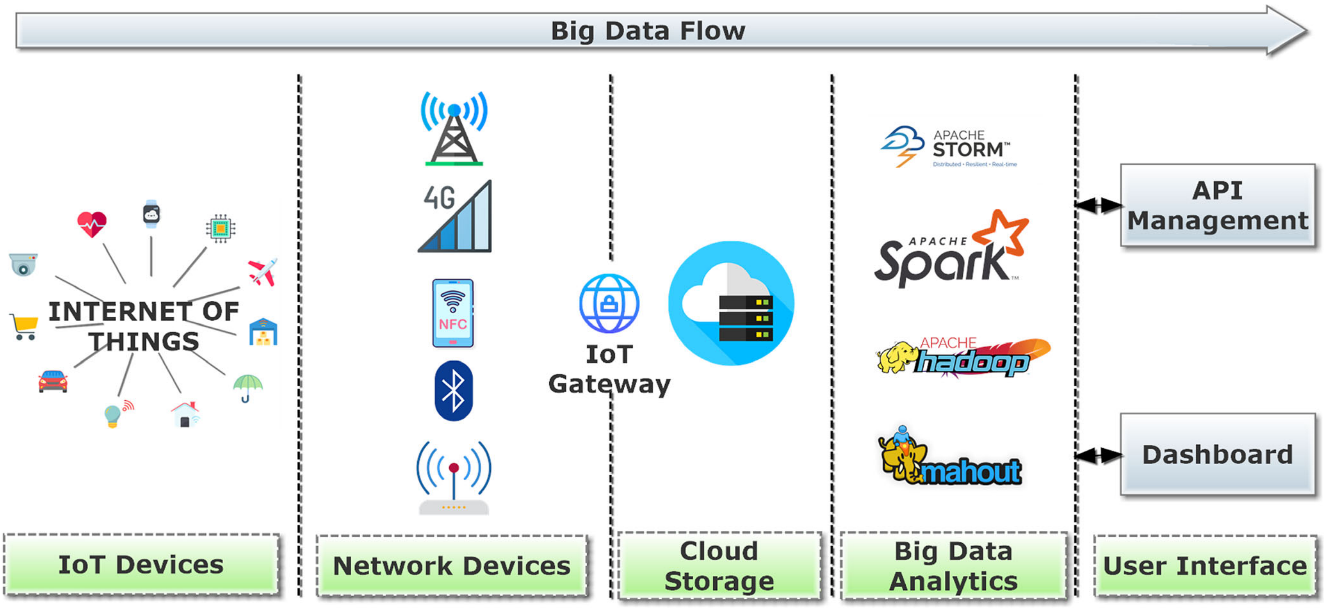


FIGURE 5 5G-enabled IoT BD Analytics architecture

compares these applications based on their utility, the type of data input that is needed, hardware infrastructure and the BDA technologies that can be used for the realization of such implementations.

2.2 | BD analytics and 5G-enabled IoT architecture

IoT architecture is well-defined in different ways based on specific domain. In this subsection a model for data abstraction has been proposed that describes the relationship among various industry verticals in the IoT space such as smart healthcare, smart transportation, smart homes, smart traffic and so forth. Various IoT architecture have been discussed in previous articles.[27-29] With omnipresent sensing techniques and data analytics,[27] presents a cloud-centric framework for IoT.

Figure 5 proposes an architecture that unifies BD analytics with IoT systems and focuses on communications and associated technologies in an IoT environment. In the sensor layer all the objects and devices are connected through wireless communication networks like Wi-Fi, RFID, Zigbee, LoRa or Bluetooth, and various other protocols. The IoT gateway provides an interface for communication among various networks. The upper layers are involved with BDA processes where large amount of sensor data is stored in cloud and is used by various BDA applications. These applications are equipped with API managements and dashboard services that helps the user to interact and perform analysis with processing systems. The overarching goal is to provide a comprehensive business solution, with a holistic architecture and testing system in enterprise environment.

2.3 | BDA taxonomy and existing analytics systems

The taxonomy adopted to categorize BDA for IoT environments has been discussed in this subsection. Different approaches have been categorized on the basis of big-data sources, various system components, big data facilitating technologies, the functional elements and the type of analytic method that has been adopted. Figure 6 describes the taxonomy of BDA solution for IoT environment. The components of the Figure 6 are discussed in detail in subsequent paragraphs.

Big data sources: Big data is collected from IoT systems deployed for applications like city management, industrial manufacturing, intelligent buildings, intelligent traffic systems (ITS) and so forth which make living much easier and secure using connected network of devices. Such large network of connected sensors and internet-connected devices

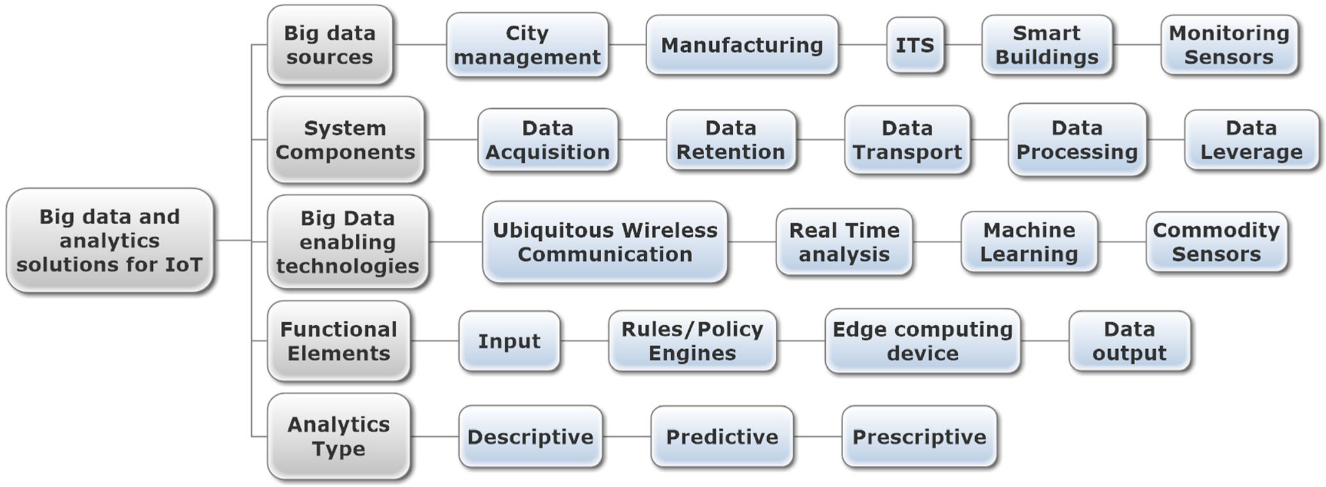


FIGURE 6 BDA taxonomy of BDA solutions for IoT systems

collect huge data which should be smartly handled and analyzed to draw useful information. A good example is the use of big data collected from such settings is seen in smart navigation systems.

System components: In the big data acquisition process, data gathering, cleaning and filtering are the major steps taken before data is saved in persistent storage for analytical purpose. Big data retention policies deal with archiving requirements of the collected data. Data transportation is concerned with the data being transported across various systems to ensure load balancing, make it available for the business continuity and duplication. Big data processing has various challenges related to gathering, storage, analysis and security of high volume, and complex datasets. Big data leverage ensures how an entity can enjoy the potential benefits of big data to improve revenue systems.

BD enabling technologies: There are some ubiquitously used wireless technologies utilized for big data transportation in IoT systems like IEEE 802.11, IEEE 802.15.1, IEEE 802.15.4, IEEE 802.16 to name a few. Instantaneous availability of big data generated by IoT systems is achieved by real-time analytics. Machine learning tools can be used to discover hidden patterns from the data which is of high volume and generated at high rate. Sensors are key entities in IoT systems in the collection and transmission of the data.

Functional elements: A variety of sources provide raw data input to the edge devices. The edge devices preprocess the data by applying certain rules and policies to discover useful information. Such a design comes with its own advantages such as higher rate of output, lesser bandwidth requirements, low cost and so forth.[30-32] Real-time output is made available to the user.

Analytics type: Analytics are of different types such as descriptive, predictive, and perspective in nature. Descriptive analytics has the potential to identify opportunities and threats. Predictive analytics can be used to predict future happenings using data mining techniques and so forth. Suggestions using simulation techniques and other such methods can be provided to stakeholders using prescriptive analytics systems. Analytics systems can also be discussed under categories such as offline analytics, business intelligence, real-time systems, memory level analytics, and massive analytics. Table 5 presents a comparison of the various analytics systems.

2.4 Opportunities of BDA in IoT environments

Internet of Things and its advantages has shown significant impacts in the industry and is one of the profound transitions seen in technology. IoT systems provides tremendous opportunity for big data analytics and its applications in various applications. Various domains such as e-commerce, smart cities, retail & logistics, and healthcare has seen an exponential growth of BDA applications in 5G enabled IoT. These areas are expected to see further rise in the potential applications and value creation by employing big data technologies in the IoT space.

BDA applications have proven to have widespread applications across various industry verticals. One of the key areas where analytics have been useful to create a growth in revenue along with increased sales and consumer base and helped

TABLE 5 Analytics types and their comparison

Existing analytics systems	Use case	Strategies/tools	Benefits
Offline	Suitable in applications which do not require quick response.	Kafka, scribe and so forth	Data acquisition effectively more efficient.
BI analytics	Adopted when data size is larger than the memory.	Data analysis	Online and offline analysis.
Real time	Mostly applicable in case of sensor data, where the data is generated frequently.	HANA	Rapid analytics techniques enable parallel processing of data.
Memory level analytics	Cluster memory is larger than the data size.	MongoDB	Suitable for real time analysis.
Massive analytics	Suitable in applications where data size is larger than the capacity of the BI products or databases.	MapReduce	Offline processing.

in optimization of product and risk mitigation is the e-commerce industry. IoT big data along with well-tailored analytics tool have been able to process big data to produce valuable insights that help in strategic decision making.

Due to the ever evolving cloud computing technology, the analytical capabilities have increased multi-folds and have witnessed dramatic increase in data storing and processing capacities within lower costs. This has opened the doors for big data analytical systems to find new opportunities in improving quality of life in smart cities with improved efficiency gains.

Retail and Logistics have seen and increased use in the deployment of wireless sensor devices to track objects in real-time. Large scale deployment of sensors in logistics produce huge volumes in data that can be processed to gain valuable insights to improve shipment experience, increase efficiency in shipping methods, lower cost and predict trends with supply and demand of goods for profit maximization.

Smart health applications have seen an increasing demand in the past decade and people have become more health conscious than before. Most people having wearable technologies today to track their health in real time, use of health monitoring systems in hospitals to monitor patient health and use of various tools to help healthcare professionals take better decisions prove the relevance of data analytics and IoT in the healthcare domain. Data analytics help in diagnosis and early detection of serious illnesses improving accuracy and better patient care.

3 | BDA APPLICATIONS IN 5G-ENABLED INDUSTRIAL INTERNET OF THINGS

The Industrial IoT (IIoT) system is a network of interconnected sensor nodes, industrial devices, monitoring equipment, measuring instruments that are deployed as a distributed control system in industrial applications to attain higher degree of control and employ automated processes. Industrial IoT networks are crucial in attaining optimal and refined process control models and facilitate efficient data collection methods that have useful information for analysis which potentially benefit by increasing productivity, process efficiency that have perceived economic benefits.

Industrial IoT Systems are governed by seven design principles[33] namely decentralization, interoperability, security, modularity, real-time, service orientation, and visualization. Wireless communication protocols play an important role in realizing interoperability among different technologies in IIoT Systems. Virtualization is an important aspect to deliver efficient services across various IIoT platforms which vary in terms of network, operating systems or applications. Decentralization is important in order to obtain highly distributed systems in IIoT networks ensuring system-wide data storage processing and analysis. Real-time updates in IIoT systems become a critical aspect for various applications that require immediate actuation and control. Modularity in design, security infrastructure, and service-oriented architecture are also adopted as important design principles when implementing IIoT networks.

CPS has led to many more emerging trends in big data in IIoT systems. CPS are the physical machine systems possessing computation power and networking capabilities.[34,35] IIoT systems are driven by IoT devices and CPS systems thus generating huge chunks of data resulting in big data.[36] In IIoT systems, sensor devices are deployed to remotely sense and perform appropriate actuation in an industrial setting.[37] Such sensing devices may be static or mobile and can therefore

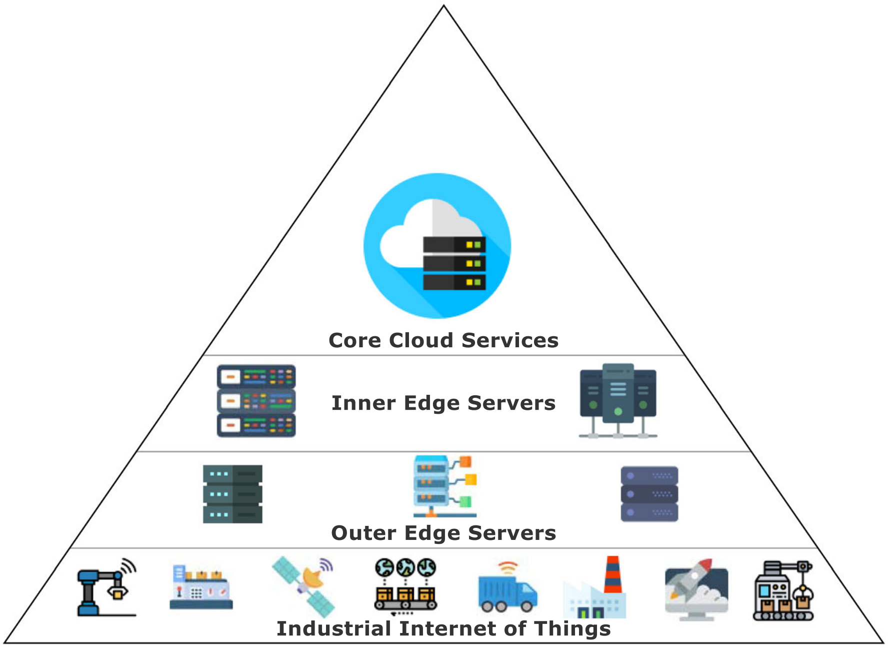


FIGURE 7 Industrial IoTs and multilayer computing resources

have a dynamic topology. Pairing them up with already established CPS platforms can enable them to perform predefined tasks.[38] Big data analytics can provide real-time service provisioning[39] by integrating IoT devices with CPS systems and back-end cloud service.

Recent trends in sensing and processing technologies has paved the way for BD processing. Concentric computing systems offer highly scalable and highly distributed systems having various sensing devices and computing systems in different sizes.[4] With concentric computing model as shown in Figure 7, computation can be performed at various levels such as edge node level, system end points, edge servers, and core cloud computing systems.[5,40,41] Edge node level includes physical devices directly connected to internet where local processing can also be done. These edge nodes are also connected with edge servers for traffic channelization of edge device data. These edge devices are then further connected with cloud services for bigger storage of data.

3.1 | BD analytics architecture in 5G-driven IIoT systems

Concentric computing systems consisting of sensors, smart wearable devices, edge servers, IoT devices, and cloud servers require efficient management of BD analytics processes in automated fashion. This kind of BD analytics infrastructure is established using multistage components which are highly interdependent and lead to smooth execution of the BD analytics process. Figure 8 shows the various components in BD analytics and the components have been explained below.[42]

Data engineering: Engineers establish computing infrastructures that are capable of processing and storage to import, process, integrate, and convert data. IIoT systems generate and integrate large amounts of data from incoming business operations and outgoing customer operations. Raw data in the first stage requires further development in order to improve quality and achieve compliance with IIoT applications. Hence, data conflicts with cleaning methods help to select appropriate data sets in case of historical data or data distribution in the event of data transmission. Data compliance mechanisms are used to ensure largely well-collected data. Data building methods with processing help in order to improve the quality of data by reducing the number of features and thereby modifying data formats for the processing to be uniform.

Data preparation: High volume and high-speed big data which is in raw form is extracted which makes data scientists invest majority of the time on data preprocessing tasks.[43] BD is refined using mathematical techniques to process random, unbalanced, and poorly maintained data points.[44] Additionally, processing the data helps in summarizing large

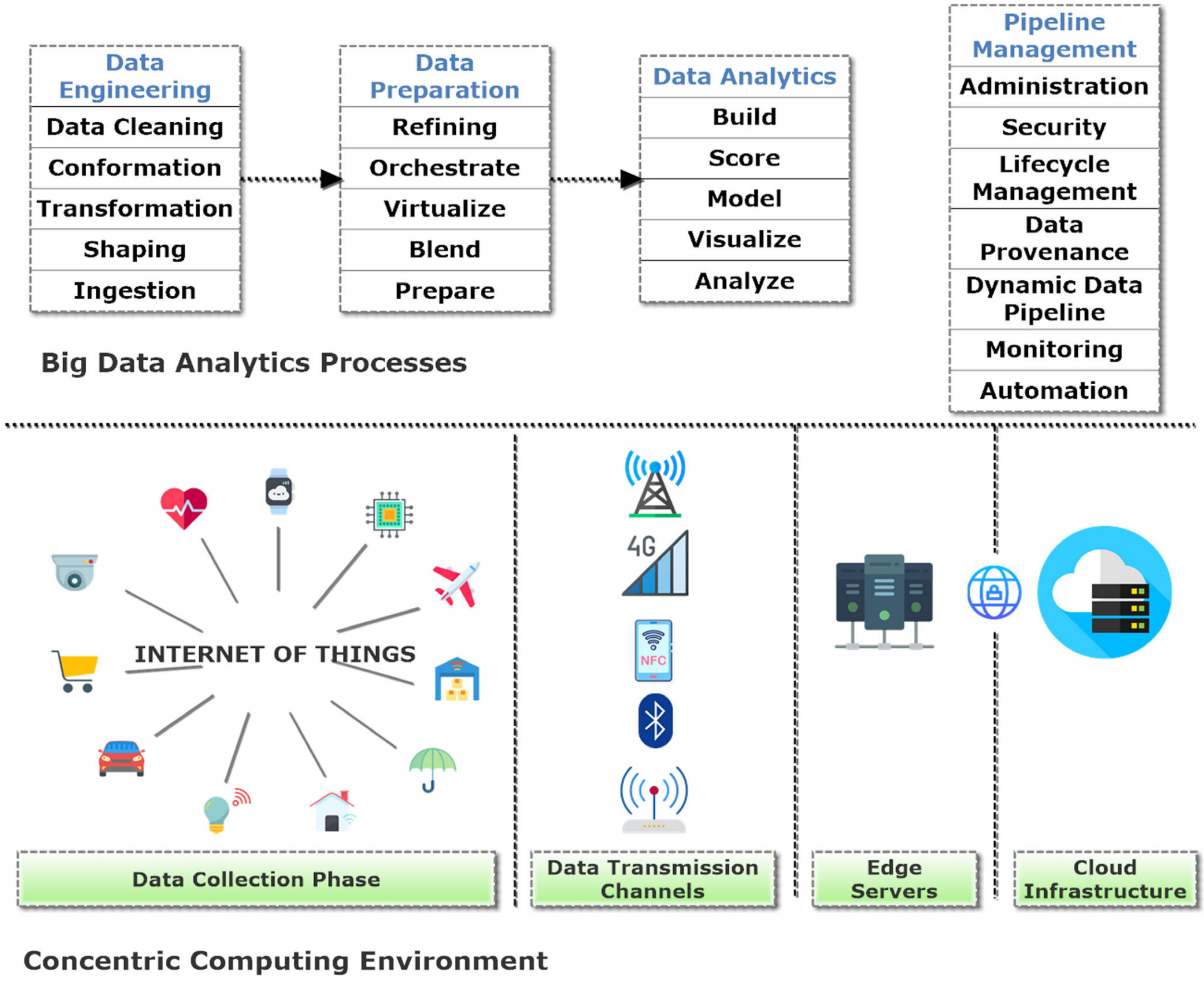


FIGURE 8 Components in BD analytics and management of multistage process of execution[42]

data to reduce the complexity. Thus, variation in the features of big data space is observed in IIoT systems. Finally, the spatial locality of data is utilized to reduce delay time in big-data applications and check traffic in network.[44] The most visible local data infrastructure can address these issues. Although, the process of blending data from multiple sources called data integration, becomes much more difficult. Therefore, it requires noise reduction and cleaning of data which further involves the data scientists in the process. The mechanisms used for obtaining outliers and anomalies are useful in preparing large-scale data for the analysis in consequent stages.[45,46]

Data analytics: Analysis procedures in IIoT systems are performed in several stages.[47] Data scientists produce intelligence and information from well-processed high-quality data. After the model is developed, model measurement tasks are put into execution by providing sample data sets and obtaining and placing attributes in the dataset by ranking them. With future data and well-tuned models that are incorporated into production facilities we can obtain information patterns that help in decision making processes.

Pipeline management: Automated data pipelines lack popularity in the Industrial IoT space as of now, however, big data analytics procedures are performed in a series of steps throughout the duration of data collection, preprocessing, integration, and analysis. A comprehensive method for the management and implementation of BDA processes is required for all levels of concentric computing models. Life cycle management is essential for the implementation of a comprehensive process from data acquisition to information recognition and performance. The emergence of data, that is, determining data ownership for owners of various stacks, also requires intensive care to ensure control of the entire

TABLE 6 Big data analysis in 5G-enabled Industrial IoT systems

State of the art	Goal	Aim	Data source	Pros	Cons	Resolution
2017[38]	Shop floor manufacturing using physical internet	Self-prediction	Batch data	Good prediction rate	Overestimation needs to be managed	Alternate machine learning algorithms
2017[52]	Customize product as per need	Self-configuration and self organization	Streaming data	Exhaustive model for heavy production and customization	Implementation and verification not available	Implementing use cases and verifying feasibility
2017[53]	Rapid and quick maintenance in real time	Self-maintenance	BD	Real time maintenance facilitated	Needs to be validated with streaming data in real time	Real time big data analytics platform
2016[54]	Sustainable microgrid development	Self-prediction	Batch data	Exhaustive model for microgrid analysis of data	Extracting analytical information in industrial microgrids is effort intensive process	Exhaustive big data analytics platform for analytics
2016[55]	Extraction of attributes for large scale customer customization	Self-prediction	HD	Fast paced development with quick response to customer requirement changes	BDA lacking large scale validation in real time	Re-enforcement learning
2016[56]	Extraction of attributes for large scale customer customization	Automated estimation	HD	Fast paced development with quick response to customer requirement changes	Big data analytics lacking large scale verification in real time scenario	Deep learning algorithms in big data analytics
2016[57]	Finding no defect	Self-configuration	BD	Industrial process models created and monitored	Accuracy not implemented	Other machine learning algorithms

system in the data.[48] The continued emergence of data streams leads to information changes that force data pipelines to redefine analytical processes. Data pipelines need constant scrutiny to detect changes, while the whole BDA procedure requires to be redesigned to return superior results.[49] From a security standpoint, cross-platform implementation of the BDA operations requires secure operation on IoT device, CPS, and large data levels.[50]

3.2 | BDA Technologies for IIoT systems

A smart factory system (SFS)[40] is a popular example of IIoT system. Self-awareness, self-care, self-organization, self-reflection, preparation, and self-comparison are some of the important features of SFS and its subprograms.[51] A review of earlier studies is presented in Table 6 showing BDA in the context of SFS and IIoT programs[51] considering the above-mentioned autonomic characteristics.

3.3 | BDA research challenges and opportunities in IIoT systems for sustainable smart cities

BDA will certainly assist businesses in the process of creating value by its involvement in the Industrial IoT space. Increase in efficiency, production cost optimization in industries, streamlining supply chain management, and ensuring greater

TABLE 7 Research challenges and solutions in BDA in IIoT

Sl. No.	Research challenge	Source	Resolution
1	Personal and business data privacy concerns.	Data sharing policies not following industry standards with not up to the mark models used in security deployments.	Privacy preservation techniques and revised policies to store data anonymously should be adopted.
2	Resource discovery and managing heavy data streams leads to scalability challenges.	Heavy data generation and edge device having less computing power.	Preprocessing data at edge device with more efficient algorithms before sending the data over the network.
3	Internal and external threats in security concerns.	Lack of security restrictions on systems and susceptibility.	Smart monitoring with exhaustive analysis and periodic testing of security systems is needed.
4	BD processing challenges include corrupt data in streams, data fusion and latency issues.	Heterogeneous sources of data create integration issues, with connectivity and bandwidth leading to slower transfer rates.	Real time integration of data aided with device centric and concentric processing can be possible solutions.
5	Data corruption and loss during transmission creating issues in communication aspects.	Huge streams of data with congestion is the major cause.	Need for robust connectivity and communication standards to deliver reliable and high throughput connection.
6	Management challenges including installation, supervision and configuration of systems in the IIoT infrastructure.	More cloud cantered architecture.	Focus toward device and build better edge computing systems.
7	Lack of conformity and integration challenges.	Standardization is lacking in adoption of big data analytics in IIoT systems.	Standardized protocols need to be set for regional and industry oriented as well as for global needs.
8	Efficiency challenges in energy, resource limitations.	Continuous sensing by edge devices leading to more energy consumption and unnecessary data generation.	Appropriate sensing and sleep cycles to be implemented for optimal efficiency about energy, bandwidth utilization, computation, and memory.

customized production are some of the key areas where BDA processes will create an impact.[58] Table 7 provides a summary of the existing research challenges and their possible practical solutions for fully utilizing Industrial IoT systems in the BDA. The table shows that there is a huge scope of improvements to be made to fully utilize the possibilities of BDA technologies in the IIoT domain.

3.3.1 | BDA opportunities in IIoT

The origination of BDA technologies in the Industrial IoT environment has given birth to a plethora of opportunities and open challenges. Opportunities have mainly emerged in areas such as artificial intelligence and automation of industrial process.[59,60] AI methods are expected to be the core technology to solve problems in data optimization and analysis[61,62] improving the throughput of industrial processes and increasing efficiency.

BDA processes in augmented technologies and wearable devices have led to enrichment in new age human-machine interaction methodologies. Real-time information from large data systems will create highly autonomous and high productivity systems and will be widely adopted technologies in the IIoT infrastructure.

Security becomes an important factor in the networked IIoT infrastructure. BDA processes are expected to facilitate detection and analysis of cyber-attacks and threats in real-time.[63]

With the recent induction of BDA processes in the Industrial IoT environment, there is a need for defining universal standards for the collection, security, storage, and sharing of big data that is collected from users and identifying the stakeholders and the benefits to pass on to the users to in exchange of the data being collected. Universal standards help in keeping ethical issues in check creating mutual benefits for the industry and its customers.

Several industries indulge in the whole process—starting with obtaining customer data to completing the product/service and procurement management.[2] Collaboration is highly regarded between various industries; however, new policies are needed to find collaborative IIoT systems ensuring interoperability.

A variety of incoming and outgoing data from various sources such as customer data, finance, IoT devices and CPS systems and so forth is creating an evolution in the big data collected at IIoT systems. In already existing solutions these various data source is treated individually to perform analytics processes. Thus, it creates a scope to design an end-to-end analytics solution as a pipeline system which is capable of handling large amounts of data from a variety of data sources at the same time and acquire highly relevant data patterns from all IIoT systems.[64]

Precision manufacturing[65] can be developed further with BDA processes into the picture. Businesses will be able to provide the appropriate products and services to the potential customers by the fragmentation and classification of customer data. Precision production will go a long way in creating value for customers and businesses. Precision manufacturing systems have been existing in healthcare industry[66] which have the potential to be integrated with Industrial IoT systems.[65]

3.3.2 | Future trends and open challenges

BDA processes must be able to facilitate real-time information patterns on interactive single-dashboard big data applications by using industry-wide knowledge. There is a huge implementation challenge with respect to planning, deploying and maintaining such big data processes specific to industry. Hence, correlating different data sources collected from different industrial domains is necessary to smoothly adopt BDA into Industrial IoT environment.

In order to maximize the value creation in BDA applications in Industrial IoT systems, concentric computing systems must be employed to orchestrate the BDA processes by providing computational and optimal storage support systems for varied sources of data. In this regard, BDA applications in addition to concentric computing models must be designed with the consideration of storage efficiency, network mobility, power usage, security and availability of information in real-time.[67,68]

4 | BIG DATA STREAMING IN SUSTAINABLE SMART CITIES

The IoT devices in developing the sustainable smart city has led to the rapid generation of huge chunks of data in the past few years which is still multiplying. This massive amount of data varies with all the important V0's and is thus called big data (referred as BD here onwards). These intelligent devices can perform tasks and follow protocol without the need of human intervention in several application domain.[69] Using cloud has significantly solved the issue of storage of the raw heterogeneous data that is generated by thousands of such devices every second. Since the amount of data is enormously large, it is one of the key research challenges to handle such huge proliferation. Security becomes another major concern for big data apart from storage. BD defines a new generation of technologies and architectures as per the reports by interactive data corporation.[70] BD enables high speed discovery, extraction, and data analytics by being able to capture large data volumes of wide variety of heterogeneous data. The intense growing use of scientific research, business, streaming, social networking, mobile devices, sensor networks, and various other applications has led to the extreme growth in data volumes, data variety and the rate at which the big data is being generated and has made it difficult to handle using traditional data processing systems.

4.1 | BD-streaming architecture

In batch processing we collect a high volume of data from multiple transactions over a given period. Using batch processing, the data so collected can be stored separately depending on whether it is structured data, semi-structured data, or unstructured data. Batch processing however is time taking and is meant to be performed on large volumes of data which is not time sensitive. Hence it does not prove to be useful in low latency applications such as online gaming applications. Figure 9 presents the basic big data streaming (BDS) architecture in the IoT environment. Since, computation performed

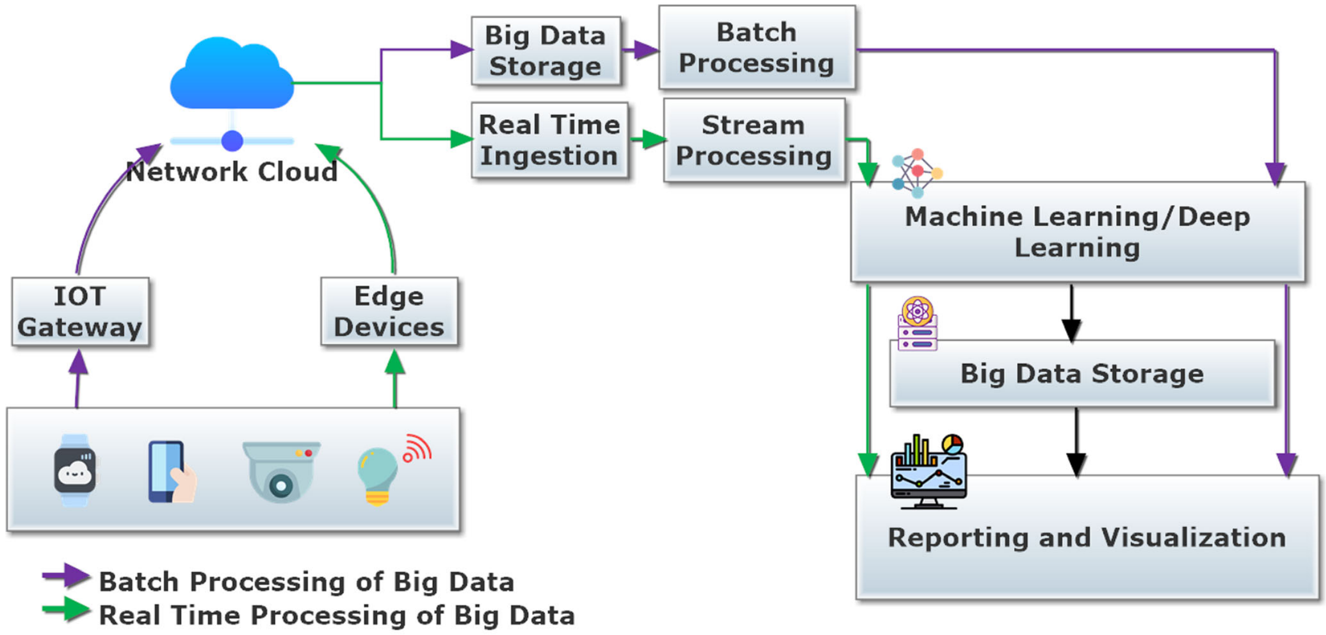


FIGURE 9 BDS architecture

by stream processing takes less time, it seems near real-time processing. High level of accuracy is also achieved by the method which is needed for time sensitive stream processing. For performing analytical tasks and use this method for BD applications containing huge volumes of data, strong and accurate mathematical models need to be employed in such applications. Analytical output in applications like healthcare helps doctors and medical practitioners take appropriate action by monitoring patient's health in very short time.[71] It can be further extended to applications where the patient is attended by the doctor remotely and since the patient's health data is fed live to the doctor through smart wearable healthcare devices.[72]

4.2 | Security concerns with BD-streaming in IoT

There have been dramatic transformations led by big data researches in the business and science fields and has also created the opportunities for facilitating real-time applications and analytics.[19] However, increased use of BD for decision making and other applications has also increased the risks related to security and privacy of data. Big data provides invaluable user data by analyzing and profiling user habits and recording user activities.[73] Security solutions for BD security improves reliability and safety of the system, while the BD privacy solutions protect from undesired interference and access of the data[74] by unauthorized entities. BD contains highly sensitive information which demands to be protected by robust and accurate solutions, thus there is a challenge of privacy and security that arises out of the additional societal value that is derived from BD.[19] Figure 10 suggests a detailed taxonomy of BDS security in IoT ecosystem. The detailed taxonomy is described below.

4.2.1 | Adhering to laws and policies

There are different laws and policies for the data security data privacy in different countries. It is imperative to make rules which are essentially important to protect user sensitive information before accessing it. While electronic observation is permitted under certain conditions,[75] it is not permitted to monitor company staff activities in many companies. In most situations however, the laws are strict enough to ensure user privacy. In case of BD streaming, it needs special attention by the data governance bodies to protect critical information. BD governance plays an important role in the applications that have high volume data transactions from IoT devices[76] and therefore needs transparency, and

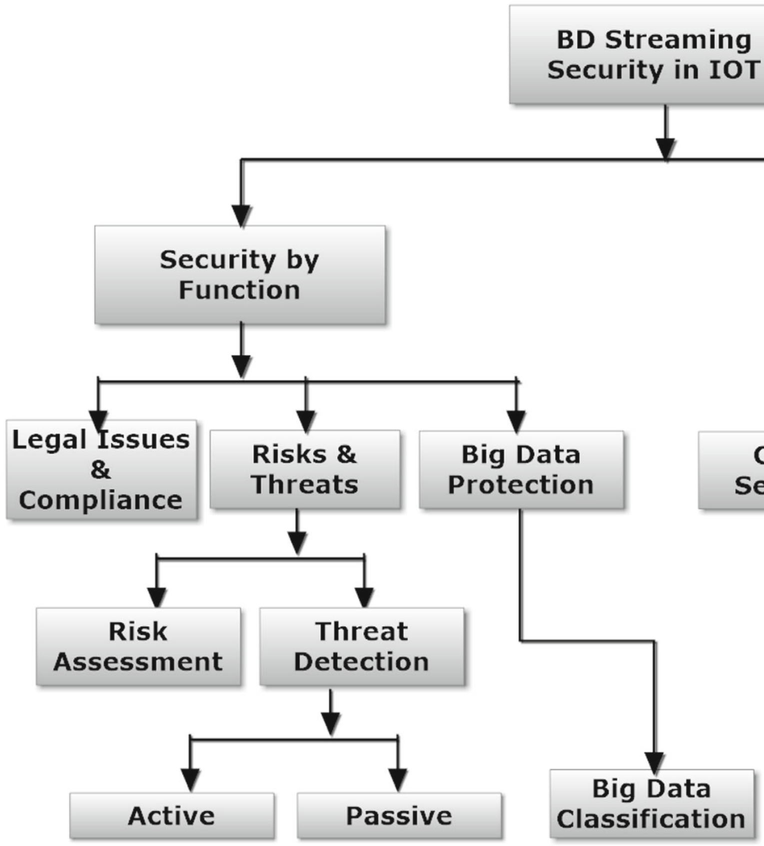


FIGURE 10 Taxonomy of BDS security in IoT ecosystem

timely rule-enforcements to achieve business objectives.[77] It is important to define clear and acceptable data policies with respect to their type and accessibility.[78] Adapted government practices is seen to be balance the risk factors and the value created with BDS and it makes full utilization of the many applications possible such as healthcare and smart grid systems.[79]

4.2.2 | Threats and risks

IoT sources produce enormous volumes of big data and hence need to be protected from threats.[11] Risk assessment is another major criterion to ensure BD security in IoT environment.

Risk assessment: Risk analysis and its evaluation is called risk assessment.[80] Risk assessment provides an approach to handle existing risks associated with BD security and the consequences.[81] Risk assessment encompasses the risks involved with different platforms which may be networks, operating systems, applications, and their interconnectedness.

- Risk analysis: Studies the assets and their vulnerabilities and identifies the risks and threats to evaluate them to avoid possible damages. There are various steps involved in the risk analysis process as described below:
- Resource identification: Identifies the various resources and assets that are crucial to the organization and evaluates them.
- Risk identification: Describes the interconnection between the threats and vulnerabilities. Using various tools and techniques, vulnerabilities can be recognized based on the availability of resources and criticality of the information.[82,83]
- Risk measurement: Specifies the interconnection between various factors of risk which may be threat impact, value of resources, and vulnerability effect.[84]

- Risk evaluation is the activity to estimate the importance of each risk by rating the risk exposures on a scale. Following that a proven method to manage and address the risk is involved.[83] Handling the risk is done in multiple stages as described below:

1. Risk acceptance: Understanding and acceptance of the risk by the organization.
2. Risk avoidance: Assessment based avoidance of the risk by the organization.
3. Risk transfer: Partly or wholly, the risk is transferred to another party.
4. Risk mitigation: Limiting the risk and the reverberations is controlled to be lower than the acceptable limits set by the organization.

It is very crucial to recognize planned safeguards in the risk assessment activity to avoid risking the organization in increased costs and danger which maybe a result of inaccurate risk assessment.[85]

Threat detection: In the IoT space, BD security concept can be expanded to active and passive sources.[86] The attacks such as active attacks, breach in the security and privacy intrusion in the device communication causes disorder in the entire network. For instance, let us say two people A & B are communicating. A is expected to reply to B, but C is replying to B. In this situation C is the active attacker.[87-89] Passive attacks on the other hand involves observation and close observation of the communication channel in the IoT system in an attempt to breach it solely upon observed data. The purpose of this type of attack is not to let the user know about it or break the communication. Such attacks are used to gather information from the communication from two entities. For instance, A sends a message to B which is passively read by the entity C. In such case the entity C is called the passive attacker.[90] In IoT systems there can be various sources of passive data including network location from mobile devices, wireless data, user patterns, GPS location, security certificates, user privileges, travel patterns, physical data of user, and many others. The prevention methods pertaining to passive attacks primarily consist of techniques such as modifying network settings, multilevel authentications, generating alert messages, and confinement to prevent ongoing fraud.[91]

4.2.3 | BD protection

When the confidentiality and authentication is taken into account then handling secure data becomes different form plain text. Authentication protects the BD from tampering, and it has many applications. While confidentiality of data is ensured by data encryption and decryption techniques. Based on big data classification like structured, semi-structured, and unstructured data Alouneh et al[92] suggested an effective classification methodology to ensure BD security. Security protection is also needed in other factors such as DoS protection and access control list.

BD classification-based security architecture is presented in Figure 11. BD is classified into structured, semi-structured, and unstructured data. In the architecture described, the extracted data is first categorized into a class and then the data is passed through several layers where security services. For instance, encryption is checked first whether it must be applied while, authentication services are needed on processed data. In the subsequent layer apt actions and security services are employed based on network topology such as local area network (LAN), wide area network (WAN) or virtual private network (VPN) and so forth. In the last step the final analytic as shown is established on the 10 V's.

To handle security in BD streaming in IoT there are various approaches. Table 8 presents a comparison of the existing methodologies of BD streaming security infrastructure. The various state-of-the-art methods are compared based on multiple parameters. Their advantages and disadvantages are also highlighted.

4.2.4 | Security by platform

Based on platform security in BD streaming can be classified into cloud security, network security, and end-device security as explained below.

Cloud security: Cloud computing came up as solution to the ever increasing volumes of data by providing storage for data which can be accessed anytime from anywhere. It includes features such as high security with visualization and abstraction and is highly scalable.[97] Cloud computing provides three types of services such as software as a service (SaaS), Platform as a service (PaaS), and infrastructure as a service (IaaS). While cloud computing is one of the most promising

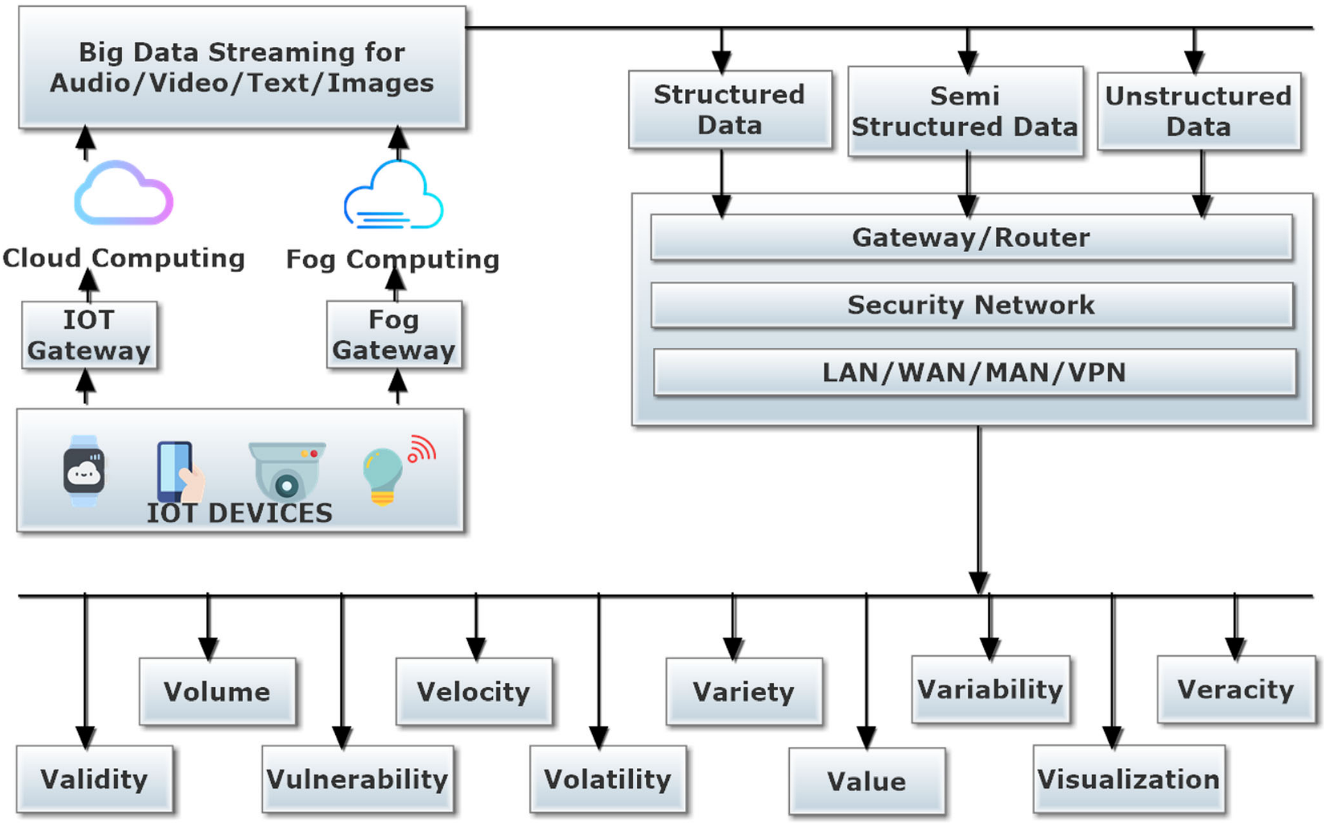


FIGURE 11 Security architecture based on BD classification

technologies, security issues have increased with the incremental popularity of cloud systems. The security issues associated with cloud infrastructure are less with the storage of data and more significantly with the transfer of high voluminous data over the network to the cloud. Mishandle of sensitive information related to patient health, personal portfolio, and location-based data can have serious security and legal consequences. Thus, to improve security infrastructure storing BD near the edge devices of the cloud can prove to be helpful.[93,95] But adopting such policies must ensure that security of edge devices is also taken care of by employing multiple methodologies such as encryption, tight access control, and secured networks.[94] Failing to take such measures makes the devices themselves more vulnerable to breach.

Network security: Network security has been divided into the categories as described below.

- Network monitoring and visibility: It monitors all the devices which are connected to a specific network. It takes into account various factors such as the number of devices which are connected, their types, identifies most active devices and the most passive devices, volume of data transfer and so forth.
- Security analysis over the network: Network flow analysis and security incident analysis is handled over a network. Anomaly detection and device behavior pattern detection are some applications.

End-device security: Endpoint security includes the management of vulnerabilities and patches. There several methods that have been proposed by various research works to handle end-device security. A popular approach FOTA was developed with the aim to update mobile phone firmware. Manufacturers of mobile devices use firmware over the air (FOTA) to regularly update their customers' mobile devices by pushing software bug fixes, software updates, and security patches directly from the service provider.

Types of devices: At device level the end-device security can be categorized into two types namely security-at-low-level and security-at-high-level. IoT systems have a varied impact due to both types of security infrastructure. IoT devices may undergo partial or complete failure as well.[17]

- Low-end device: Attacks affect the entire IoT system since here security involves the low power devices such as IoT devices in smart home applications. Smart devices like washing machines and smart entertainment systems which

TABLE 8 Existing approaches to handle big data streaming in 5G-enabled IoT

State of the art	Highlights	Advantages	Disadvantages	Parameters considered						
				SC	ACM	CA	PC	SSP	GOV	RTA
2018[93]	Protecting the user credentials on the core network of cloud data centers	Relevant enhancement in contrast to existing software developments	Early security forethought requires better architectural overview	✓	✓	✓	✓	✗	✗	✓
2018[92]	Big data complexity is controlled using different classification mechanisms	Classification mechanisms used to resolve unstructured data collected from edge devices	Unstructured attribute types, unstructured data was not explored	✓	✓	✓	✗	✗	✓	✓
2017[94]	Utilized cloud security measures to protect the information	Safeguarding cloud security concerns is focused	Holistic cloud security and architectural knowhow not explored	✓	✓	✗	✓	✓	✗	✓
2017[95]	Concentrates on possible issues by cloud customers and established the compliance models & managed the risk to control security	Security policies investigated and their advocacy over security concerns to consumer	Surging amounts of data makes security concerns more complicated. Such scenarios were not explored	✓	✓	✓	✓	✓	✓	✓
2016[80]	Focuses on information security at organizational level	Offers the administration diverse methodologies for risk assessment and check	Security at user level not focused	✓	✓	✓	✓	✓	✓	✓
2016[17]	Classification of security attacks focused on to be useful to IoT developer community	Better consciousness of the risks associated with security shortcomings	Influence of online data that leads to major security issues in the IoT environment is unexplored	✓	✓	✗	✓	✓	✗	✓
2015[96]	Offers big data security consideration, monitoring and access control	Avoidance of risk and security compromises in big data	Unstructured data not traced	✓	✓	✗	✓	✓	✓	✓
2015[88]	Security Architecture of IoT systems studied, security concerns related to interchange of diverse objects	Better evaluation of the existing challenges and concerns of IoT	Network and hardware-based security for IoT was not considered	✓	✓	✗	✓	✓	✓	✓
2014[90]	Security challenges in sensor networks is the major concern and identification mechanisms of threats	Tackles most of the security issues	Threats may lead to compromise of data confidentiality	✓	✓	✗	✓	✓	✓	✓
2013[87]	Sensor network specifications and security concerns are addressed, and their classification into various security attacks ad management mechanisms	Security attacks taxonomy and measures to handle security breaches increases security infrastructure effectiveness	Unattended area unexplored, where security concerns are most likely	✓	✓	✓	✓	✓	✗	✓
2013[79]	Highlights the big data governance and its related actions	Offers the minimal required information like threat, usefulness and tariff related to governance	Policy does not explore all areas. Only partial policy taken into account for Big Data governance	✓	✗	✗	✓	✓	✓	✓
2013[77]	Highlights the big data governance	Improved governing in big data especially for structured data	Unstructured data complexities	✓	✗	✓	✗	✗	✓	✗

(Continues)

TABLE 8 (Continued)

State of the art	Highlights	Advantages	Disadvantages	Parameters considered						
				SC	ACM	CA	PC	SSP	GOV	RTA
2012[85]	Intrusion response system classification and taxonomy	Information retrieval system is well ordered	Attacks neutralized	✓	✓	✓	✗	✓	✓	✓
2012[84]	Risk mitigation and analysis	Risk component, threat recognition and response system	Security treated as a whole	✓	✓	✓	✓	✓	✓	✓
2011[81]	Security challenges analysis framework and resolution of issues	Security application identifies risks and threats on the network	Level of threat in security policies not explored	✓	✓	✗	✓	✓	✓	✓

have the potential to be remotely monitored and controlled and use low-cost, low power connecting system using an outside radio link come in this category.

- High-end device: IoT devices connected to the internet are accessible from anywhere at any given point of time using any computing device. This kind of security infrastructure requires devices to remove the attacks on IoT system entirely.
- Table 9 gives a detailed comparison of the existing tools and frameworks for BD streaming security management in internet of things environment. Details with respect to various parameters have been provided and their advantage, disadvantages, and techniques have been highlighted.

4.3 | Secure BD-Based healthcare tracking

Due to the recent Internet advancements such as increased interactions over social networks and transmissions in IoT ecosystem, huge volume of real-world data has been generating from various sources. Consumption and generation of data is mainly seen to be from sources like text, video, numeric data, audio, logs to name a few. Such massive data generation has led to BD being not limited to three V's.[96] In this subsection, we will look at the 10 V's which describe different attributes of BD Healthcare 4.0 system. This subsection also describes the security challenges for IoT devices during the BDS process with the help of a case study.

4.3.1 | BD classification in Healthcare 4.0 systems

This subsection describes the 10 V's which describe different attributes of BD healthcare system as per Industry 4.0 standards.

Volume: Doctor records, patient data, nurses and staff data, and other information in a hospital is collected continuously which leads to a huge volume of data to pile up. Patient monitoring equipment and other healthcare gadgets also contribute to the increasing data volumes in smart healthcare systems. The volume is an attribute of big data which is used to estimate storage requirements for healthcare systems.

Variety: In Healthcare 4.0 system the data can be of varied types and from different sources such as smart devices, sensors, cameras and so forth. The data may be text, video recordings, audio recordings, images both 2D and 3D. Variety of data can thus be structured, semi-structured, and unstructured format in the IoT healthcare systems.

Velocity: Since the rate at which data is collected, velocity becomes of the key challenges in Healthcare 4.0 systems. Data is generated and captured at a very faster rates since the sources of healthcare data such as smart health devices and social media have seen an exponential growth in the last decade.

Variability: Variability is an attribute which is used to detect abnormality and irregularity in the data being collected. Irregular rates of incoming data can also be identified. The inconsistencies detected in the big data may affect health condition of the patient.

Vulnerability: Vulnerability of big data is an attribute that helps prevent possible attacks in the Healthcare 4.0 systems. It has a direct relation with the security aspects in the IoT healthcare system.

Volatility: Accurate estimation of cost and complexity is important to design the storage in a Healthcare 4.0 system. Volatility helps in doing so. Volatility is associated with volume and velocity of the streams of data and it denote the significance of data.

Validity: Validity of big data refers to the accuracy of the data stored in the Healthcare 4.0 system. In order to validate data, it needs to be correct and accurate under permissible limits.

Veracity: Measurement of confidence in the data is called veracity. It indicates reliability in terms of the data sources. An example situation where veracity property of big data proves to be useful is when EHR of different patients are recorded over a number of years. Using veracity, we can find the answer to questions like count of operations performed by surgeons in the past 5 years which were successful, the methods and team involved and other details.

Visualization: Since healthcare IoT systems in Industry 4.0 standards have to maintain functionality and scalability along with decreased response time, visualization of BD becomes one of the key challenges.

Value: It is important to give value to custodians for any organization's success. In Healthcare 4.0 system a systematic feedback system with a proper monitoring system is important to maintain the value. Value is created by meaningful information.

TABLE 9 Security management framework for big data streaming in 5G-driven IoT

Software	Open source	Highlight	Key technique	Advantages	Limitations
Apache Apex	✓	Synthesized stream and aggregated processing of big data	Windowing at event time	Scalable system with secure operation and provides fault-tolerant design	NA
Informatica Data Streaming	✗	Systematic aggregation and transmission of streaming data in real time. Batch processing technologies employed for analysis in next stages.	Structured as well as unstructured data stream for efficient data collection	Scalability, real-time pipeline management, spontaneous load balancing, supports geo-spatial data	Proprietary tool
WSO2 complex event processor	✓	Real-time identification of events inside an event, analysis of impacts and actions to be performed.	Processing of complex streams and events	Lightweight application with real-time event detection with alerts and enables monitoring of correlation	Recursive query, external calls
Stream Analitix	✗	Visual based platform for big data analytics providing stream processing with machine learning algorithms, helps achieve real-time and enterprise level software.	Analysis on big data streaming using machine learning algorithms	Visually build big data applications, analysis on data in IoT applications, anomaly detection in real-time	NA
Azure Stream Analytics	✗	Users are able to develop and run applications in real-time analytics on multiple data streams simultaneously from sources which can be sensors, web data, social media and smart devices.	Simultaneous analysis applications in real-time	Scalable and provides real-time analysis on IoT data and applications using Azure IoT-Hub	Expensive services on demand
Apache Samza	✓	Fault-tolerance achieved by Apache Hadoop YARN, Distributed stream processing for messaging, processor isolation, and resource management	Framework for computation is asynchronous	API is simple and easy to use, durable, scalable, and ensures proper coupling of processor	YARN resource manager is the only option
Apache Flink	✓	Data flow achieved with distributed streaming network.	Information streaming programming model	Supports exactly one semantic, Fault tolerant	Requires external data store
Apache Storm	✓	Streams data in real time with task parallel, distributed system	Streaming data, RPC and continuous computation is done in distributed fashion	Fast framework providing high scalability and useful in real-time computations	Topology needs to be decided before hand and change is not advised. Not preferred for scheduled jobs.
Apache Spark	✓	Prompt unified validation utilized in big data applications	Scalable and parallel processing of distributed resilient data set	Easy Operation, by using in-memory computation it works faster than Hadoop for data processing in large scale	Inadequately handled problems in file management, handling delays, and windowing.
SAP event stream processor	✗	End to end processing of real time streaming data from capturing to analysis and response management system.	Event based real-time stream processing	Real-time stream processing in heavy volumes and analysis performed on critical event data to generate immediate response	Expensive services

(Continues)

TABLE 9 (Continued)

Software	Open source	Highlight	Key technique	Advantages	Limitations
IBM Streams	✗	Ingest, process, analyze huge amount of real-time streaming data	Analysis performed in real-time on huge volumes of data	Data parallelism and multiple heterogenous data streaming	Expensive services
Amazon Kinesis	✗	Collection, processing, analysis of real-time streaming data and achieve timely information	Absorbs real-time data into data stores with publisher subscriber model	Fully managed, scalable, real-time. Used in ingesting real-time data like video, website click streams and so forth.	Expensive services
Apache Kafka	✓	Unified, high-throughput, low-latency platform for handling real-time data	Publish subscriber message queue in distributed fashion	Scalable, real-time, fault-tolerant, low latency, batch handling, replicates topic log partitions onto multiple servers	Reduced performance achieved with message tweaking and lack of monitoring tools

4.3.2 | Secure BD-healthcare case study

It is essential to understand the security systems problems for the proper functioning of the IoT devices to stream big data. To demonstrate the same, a case study has been presented which highlights secure healthcare system and the various phases in which health monitoring systems function. Real-time transfer of data is critical in the current system to analyze and correctly predict decisions which impact the healthcare industry. Sensors and actuators especially wearable smart devices are becoming essential medical devices to collect clinical information of people.

Healthcare devices are producing enormous big data continuously. Data such as heart rate, body temperature, blood pressure and so forth is transmitted by a fog layer[98] and alerts and notifications are sent to the monitoring person who may be a doctor or the patient himself. A hybrid big data based secure health care system was proposed by Kaur et al.[99] Figure 12 shows a four-layer architecture for secure monitoring system in Healthcare 4.0. The different stages are explained below.

Data collection phase: It is the first step of the secure monitoring system. In this stage the data is collected from several internal and external sources as shown. This heterogeneous data collected from various sources needs to be secured to protect the user's confidential data which is a necessary requirement of Healthcare 4.0.[100] Hence patient privacy is of utmost importance while gathering the patient's data.

Data transmission & storage phase: After the data is collected, required transformations, and classification of the data is done to get meaningful information from it. This is done by removing noise and duplicated values and improving the overall quality of data before the next stage.

Security & privacy stage: In this stage various validation techniques are employed to ensure safety of collected data.

Applications: Finally, the collected data is used for deriving the important decision-making information used by health practitioners or the patient himself.

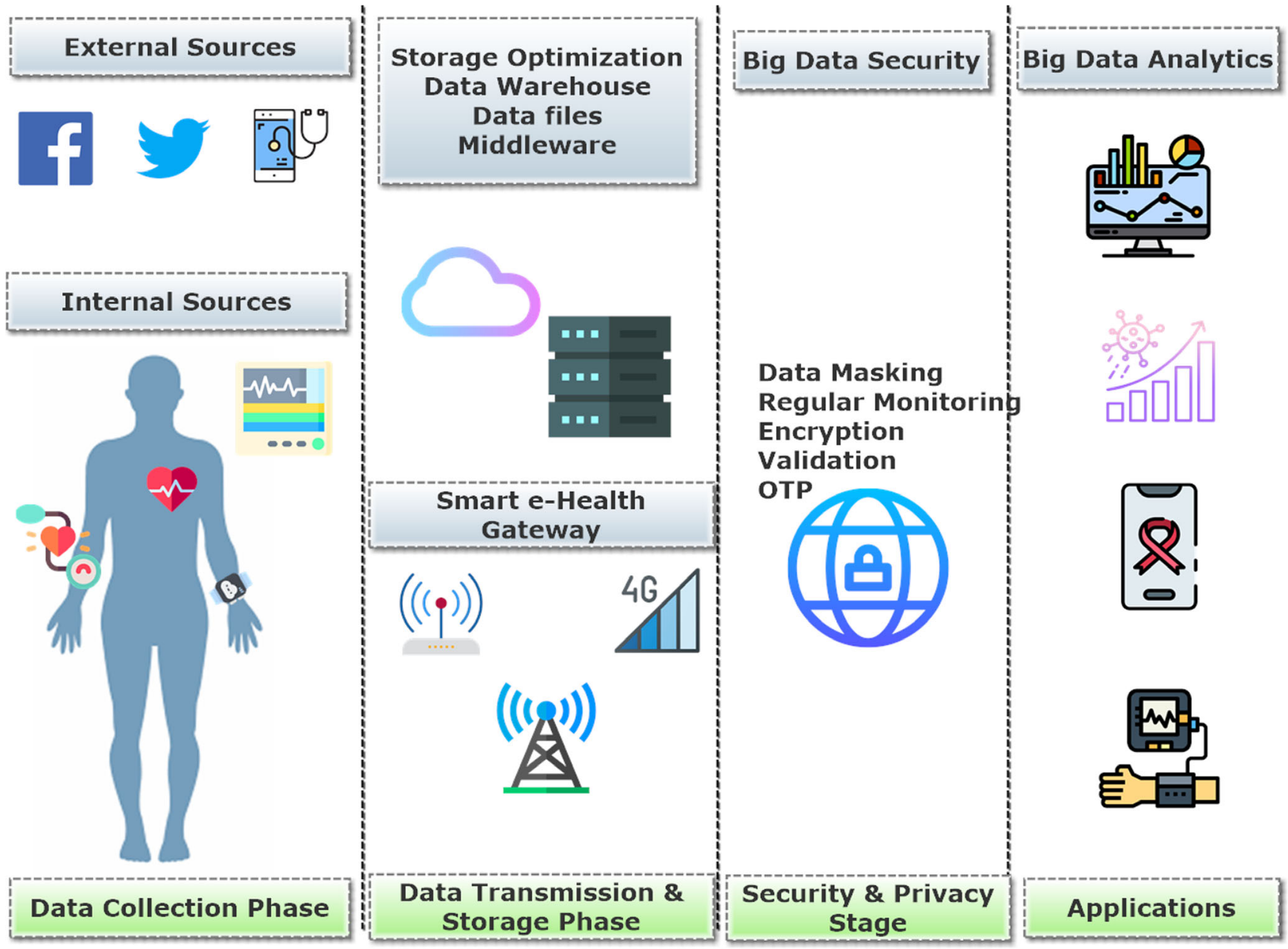


FIGURE 12 Healthcare 4.0 IoT system and BDA

Furthermore, the security challenges and privacy issues in BDS envisaged by the healthcare industries are discussed in detail by Abouelmehdi et al.[101] Also, methods to resolve the issues have been discussed.

4.4 | Security challenges with BD-streaming in IoT

The various challenges associated with secure BD-streaming in the IoT space have been discussed below.

Limitations of resources: Since computation power and memory are the two bottlenecks in the IoT it becomes challenging to deploy robust mechanisms for handling security in the IoT architecture. Keys and certificates need to be exchanged for deploying communication protocols successfully in IoT environment for BD-streaming.[102,103]

Single point of failure: Reliability vs cost trade-off is one factor to note while trying to create highly available IoT elements with real-time performance. Standard approaches need to be designed to address this issue to keep the downtime of application in permissible limits.

Protocol convergence: Security mechanisms implemented in IPV4 network settings should be portable to IPV6 networks.

Blockchain susceptibilities: Blockchain based security systems are also vulnerable to attacks. Since IoT devices have restricted computation power and memory, the hashing power is questionable and so it can allow an intruder to host the blockchain. Effective protocols need to be deployed to dodge race attacks.[104,105] This leaves a scope of enormous improvements in the area of security and privacy management in IoT environment using blockchain technologies.

Security software updating and management: When there a many heterogenous devices it becomes challenging to deliver software updates at scale while ensuring reliability. In addition, concerns related to reliable and secure management, procurement, and data privacy are important. Blockchain is one of the emerging technologies that is being noticed for having potential to provide robust IoT security solutions in the IoT system despite blockchain itself is known to have several research challenges such as efficiency, distribution, collaboration, regulation and so forth.[95]

Interoperability of security protocols: In the previous subsection we have discussed a powerful global security design for Industry 4.0 IoT systems, agreements that need to be made on each layer and must work together seamlessly. An effective level of safety in each layer can be implemented as per design requirement.

Standard verification protocol: BD generation in IoT systems is a continuous process due to large deployment of cheap energy efficient sensor devices. When deploying such massive number of hardware devices, ensuring security mechanisms for packet processing and routing becomes a challenging task. Thus, after the deployment of IoT devices, it is very difficult to manage any vulnerability that is detected and so there is a significant need for standard protocols to for verification processes to ensure security in largely deployed IoT networks.

Heterogeneous devices: IoT devices are naturally diverse and range from low power devices to high power sensor and actuator devices. To manage these large set of heterogeneous devices there is a need of a security structure that can cater to identifying each device that generates big data and dynamically make security judgments and configurations. Hence there is a need of a dynamic framework with standard security protocols that provide intelligent IoT security infrastructure.

5 | BIG DATA MINING IN 5G-DRIVEN IOT ENVIRONMENT

Communication modules needs to be installed within the IoT sensor devices for collecting data by means of embedded communication. The communication protocols that connect a series of devices and smart equipment for the reception and distribution of data are solutions such as Wi-Fi, RF, Bluetooth, GSM and so forth. Billions of devices are getting connected to the internet through various channels and there has been an increasing demand to handle the amount of data that is generated over time. If the increasing demands and rising volume of data is not identified and handled, then it will restrict us in various ways.[106] This section describes the applications of data mining (DM) techniques in the processes of data transfer, processing, and storage.

There are various techniques that are used in processing, transmission, and storage of big data. Figure 13 shows the various factors of big data analytics and the techniques that are popular at each step. Various tools and techniques facilitate data mining applications for analyzing methods to compress the data for optimal retention and storage solutions. Extraction of meaningful insights form the data is the most important step which ultimately makes the data worthy. Data mining techniques have proven to be useful in the analysis of gathered data[107,108] and improving IoT intelligence.[106]

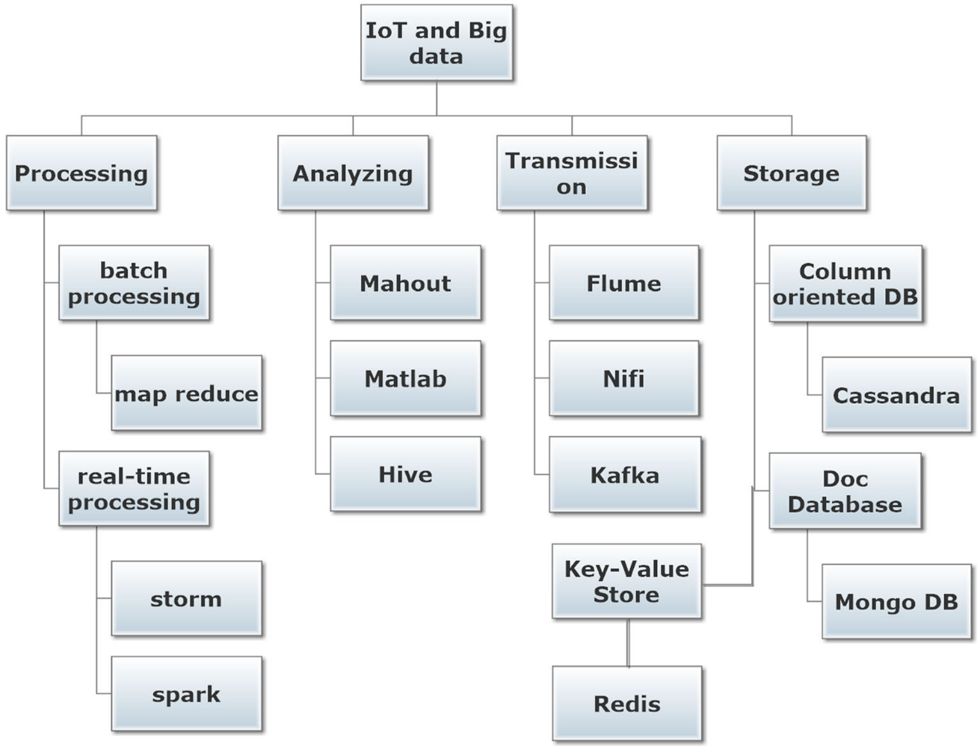


FIGURE 13 Steps in BD analytics and techniques

Data mining mechanisms are still popular and widely distributed in both problem-solving and general data analysis methods. Similarly, machine learning techniques and statistics find application. The emergence of big data also affects the analytical requirements. Even though the requirements for effective methods are present in every outlook of big data management,[109] such as data gathering, persistence, processing, and analysis; big data analytics requires more efficient and faster methods compared to traditional analytics methods that is capable of processing high-volume varied data streamed at high velocity in cost effective ways to meet upcoming demands of the industry.[110]

5.1 | Data mining techniques in big data analytics

Data mining techniques play a key role in the development of most strategies for big data analytics. Available big data analytics techniques are varied, and it is important to understand the appropriate decision-making method that need to be employed in a scenario. It is estimated that more than 50 billion devices that includes smartphones, sensors, game consoles and so forth are going to be connected to the internet via different networks.[111] Thus there has been voluminous increase in the amount of data produced due to the miniaturization of connected devices of the internet.[112] To take advantage of this opportunity, performing analytics on the data is extremely crucial. This subsection introduces the various data mining techniques that can be employed in different use cases of big data analytics. Some of the strategies also meet the needs of big data analytics in IoT environment providing more efficient and optimized solutions. Commonly diverse set of high volume data deliver deeper insights but there may be exceptions as well as more data can have more obscurities.[7] Figure 14 categorizes the various big data analytics methods under various data mining techniques which are classification, association rule, clustering, and prediction. Each section is a DM task and incorporates multiple algorithms and strategies to meet the requirements for data extraction and analysis. These techniques are elaborated in further sections below.

5.1.1 | Classification

It is a supervised learning method that uses prior training data to classify data objects into groups.[113] The objective of classification predictive model is to predict the class for an input object. The class or group are predefined categories as shown

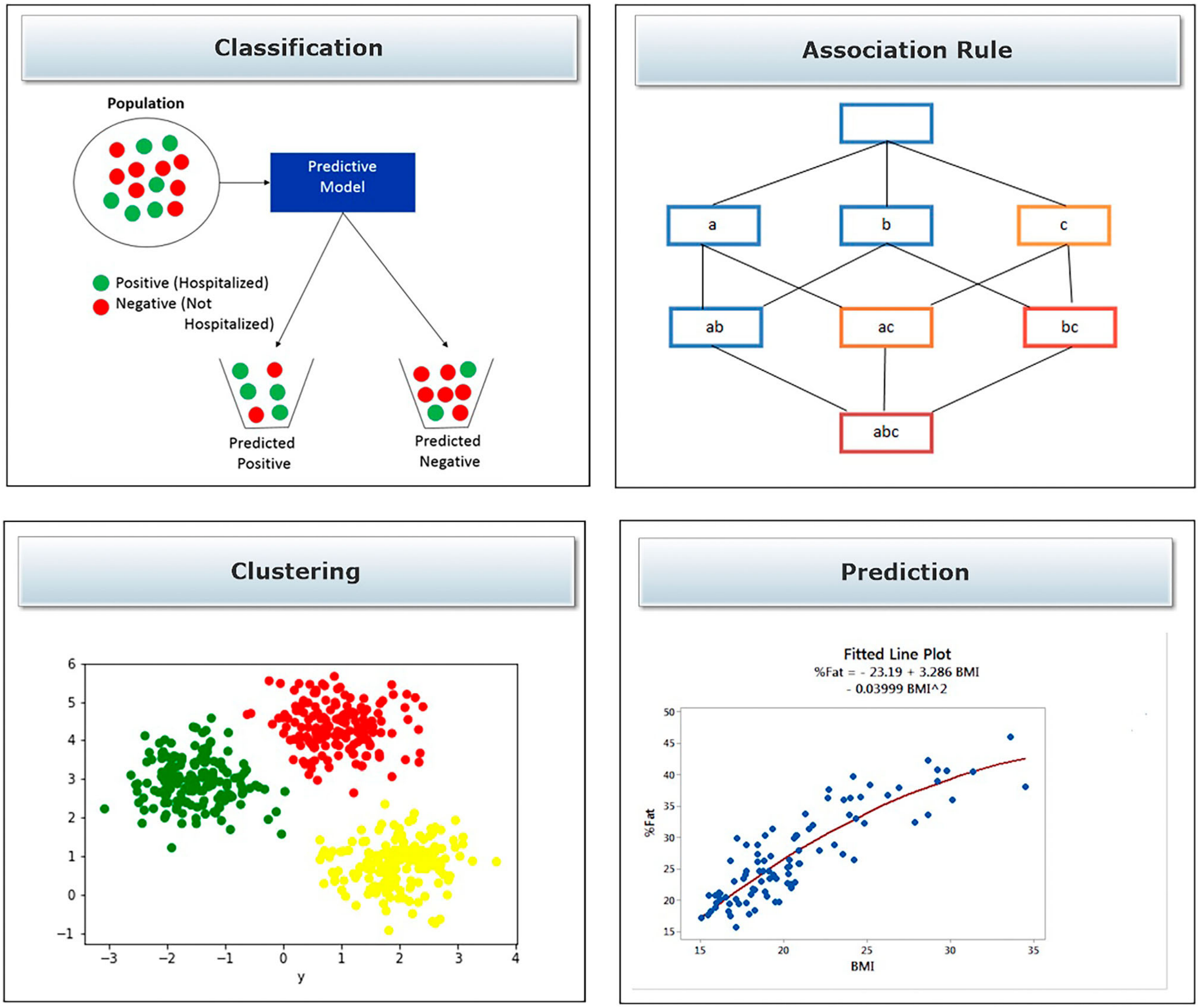


FIGURE 14 Data mining techniques in big data analytics

in Figure 14. Finding unknown or hidden patterns is a major challenge for large IoT data. In addition, extracting important data from big data sets to improve decision-making is a critical task. Model interpretability is offered by Bayesian network which are highly efficient in complex and big data analysis over traditional data formats. SVM is also a classification method uses statistical learning for the analysis of data patterns. SVM finds application in various applications. Similarly, efficient mechanisms based on KNN are used to discover hidden trends from data and find their similarity to a previously described category.[114] Classification is one of the most popular techniques in data mining employed in big data analytics.

5.1.2 | Clustering

Clustering employs un-supervised learning technique and clusters different group of objects based on distinct features.[115] Figure 14 shows how many given objects can be grouped into clusters and that makes data manipulation easier. Partitioning techniques and hierarchical clustering are the popular approaches for clustering. Partitioning clustering requires a data analyst to define the number of clusters and then applying an algorithm like K-means to group the objects to the nearest cluster. Hierarchical clustering on the other hand combines smaller clusters in iterations to form a tree and creates aggregated clusters.

TABLE 10 Data mining techniques in BDA for 5G-driven IoT

	Big data mining methods				
Applications	**Prediction**	**Classification**	**Association rule**	**Clustering**	**Time series**
Speech recognition	ø	✓	ø	ø	✓
NLP	ø	✓	ø	ø	ø
Social network	✓	ø	ø	✓	✓
Market analysis	✓	ø	✓	✓	ø
Industry	ø	✓	✓	✓	ø
e-governance		✓	✓	✓	✓
Healthcare	ø	ø	ø	✓	✓
Medical imaging	ø	✓	ø	✓	✓
Human genetics	ø	ø	ø	✓	ø
Bioinformatics	ø	ø	✓	✓	ø
Disaster management	✓	ø	ø	ø	✓

Note: ✓ = Supported and ø = Unknown.

5.1.3 | Association

Big data analytics finds its prime application and importance in commercial decision making and analysis of the market conditions. Association rule mining[116] identifies the interrelation among different entities in order to analyze market trends, consumer behavior, and predicts product demands. Based on the regularity of the occurrences of data, association rule mining identifies and creates rules. For data processing two mechanisms can be followed in association rule mining. One of the ways can be to identify collaborative associations using algorithms such as MSPS[117] which is a priori-based technique. Another method is to use the technique for analyzing event patterns in continuous data and this approach is temporal sequence analysis.

5.1.4 | Prediction

Historical data is used as training data to obtain results as patterns and behavior in the data in predictive methods. Predictive analytics help in applications such as predicting social media trends and predicting customer buying behavior.[118] Fuzzy algorithms and SVM are popular techniques employed to identify the correlation between independent and dependent entities to obtain regression curves. This can be useful in cases like prediction of natural disasters. Big data analytics requires data representation as a significant factor to reduce dimensionality of big data for efficient decision making. ARMA,[119] wavelet functions,[120] and bitmaps are important research in line with time series analysis.

The most popular data mining techniques which are widely accepted in applications like speech recognition, disaster management, bioinformatics, e-governance, social media and so forth are presented in Table 10. In the table "✓" is used to denote that the data mining method is supported in the application domain while "ø" denotes that it is not very obvious that the method is supported in a particular application scenario. Classification techniques are well-suited for applications like NLP, speech recognition, industry, e-governance and so forth while clustering technique find use in most applications. Prediction techniques mostly suited for disaster management and market analysis and social media by predicting future trends.[121]

5.2 | Case studies: Framework, architecture, and applications of BDM strategies in IoT systems

Data mining techniques has the potential to increase the efficiency of the big data transmission, processing and analysis in IoT and address the storage challenges of huge volumes and high-speed data. This subsection presents an investigation of the recent research that has been carried out in the field of IoT big data and data mining to identify important subjects that are focused in the current research or expected to trend in future.[122] A study of 30 research articles has been summarized and categorized into three categories to provide a detailed comparison of the state of the art in each category. In the first category the articles related to framework have been studied. Architecture is the second category that has been created based on the research articles studied. The third category lists all the research articles under the applications category in which the artifact mainly focuses on the application scenario of their corresponding research.[123] The Tables 11, 12, 13 provide key information about the research artifacts that have been studied in a tabular representation for easy comparison among them. The various parameters include main aim and highlights, whether the research proposes a design or an implementation, the pros and cons of the proposals, the kind of data mining task(s) that has been employed, and knowledge about where the proposed research can be employed.

5.2.1 | BDM framework

This subsection investigates the important studies focusing on the data mining framework that are proposed for an IoT environment. Table 11 provides detailed information on the existing frameworks on big IoT data mining.

The CSF[124] study introduces a semantic intelligence-driven framework for media data. Shared information by social media is extracted and converted to key value format and then using MapReduce and Spark technologies for parallel processing. The Ahab[125] framework is a generic cloud-based fault tolerant solution and it uses HDFS and Lambda like technologies for processing of data. It is a scalable mechanism and finds use in smart traffic implementation of IoT. SBDA[126] framework features a unified framework for semantic web IoT and big data. It has not been implemented to any use case yet and it yet to be tested.

In the IBDA[127] framework data from IoT sensors are analyzed to give output in real-time. This kind of framework is applicable in applications requiring continuous monitoring like smart building applications. The CEP & Self-healing[128] framework features a self-healing IoT capability. Cyber physical systems (CPS) is an area are where the framework finds application however its performance has not been analyzed for increased load. ISES[129] framework features a data processing capability from remotely controlled vehicle. Hive implementation enables data processing in batches, and it features faster query response times compared to traditional mechanisms. The COIB[130] framework has use limited to industrial applications only. It features a cognitive IoT framework and is efficiently scalable with real-time storage using HBase. IoT-StaticDB[131] framework features a generic cluster of statistical databases and provides better performance with faster response. It is being used in routing taxis at Beijing and has proven to be an important method to transform data into meaningful information employing technologies such as Casandra, MongoDB and PostgreSQL.[153-156]

5.2.2 | BDM architecture

This subsection investigates the important studies focusing on the data mining architectures that are proposed for an IoT environment. Table 12 provides detailed information on the existing frameworks on big IoT data mining.

The GOVA[132] architecture features a BDA design plan for smart grid systems. The goal of this architecture is to enable a grid system design with scalable storage systems planning using graph database. Health system architecture[133] provides a health IoT based big-health system application with cloud-based service layer design offering storage, processing, and analysis solution. IV-tier architecture[134] provides an efficient and scalable implementation aimed at smart city and urban planning using big data analytic system. Prediction, processing, classification, and analysis are the key data mining tasks involved. TreSight architecture[135] was proposed with the purpose of improving smart tourism in the Trento city of Italy. It employs FI-WARE technology to implement a content-based recommendation system that is aimed at better decision making for cost optimization.

Behra architecture[136] features a generic architecture for storage and analysis and offers a real-time solution with efficient data management and optimal storage costs with cloud-based computation platform that is well suited to employment for efficient power management applications. Rathore architecture[137] features a real time drug response system by proposing a 5-layer architecture. The various layers are associated with processes such as data collection and filtration, communication, processing, management, and the service layer. It utilizes HDFS with MapReduce and Spark for parallel processing of data. Kholod architecture[106] presents an algorithm for distributed data mining to provide increased processing speeds with efficiency by using cloud and fog systems. It is aimed at applications to process time series data, breast

TABLE 11 Existing frameworks on IoT big data and data mining

Framework name	Methods	Highlight	Intended plan	Pros	Cons	Big data/data mining task	Employment
CSF[124]	HBase, MapReduce, Spark	Crowdsourcing through collective information of social media users	Implementation	Precise and highly efficient unified model	Low accuracy and dependent on results of social users for its input and does not employ any machine learning model.	Storage, processing	Social media
Ahab[125]	HDFS, Lambda, Spark	Cloud based generic solution which is scalable and fault-tolerant	Implementation	Error tolerant online and offline analysis with scalability	Incomplete development	Storage, processing, analyzing	Smart traffic
SBDA[126]	Spark, MLlib spark	Unified framework for semantic web, IoT, and big data	Implementation	Semantic web, data stream	Not applied in any use case	Storage, processing, analyzing	Specific application not mentioned
IBDA[127]	Flume, pySpark, HDFS	Analysis of data from IoT sensors generated in real-time	Implementation	Continuous monitoring	-	Storage, transmission processing, analyzing	Smart buildings
CEP & Self-healing[128]	Kafka, storm	Self-healing IoT capability	Implementation	Self-healing	Performance not investigated for increased load	Transmission, processing	CPS
ISES[129]	Redis, Hive, PostgreSQL	Data processing from remotely controlled vehicle	Implementation	Faster response to queries compared to traditional methods	Data processed in batches because of usage of technologies such as Hive	Storage, processing	Underwater
COIB[130]	HBase	Cognitive IoT framework	Implementation	Scalable with real time storage	Limited to industrial applications	Storage, analyzing	Industrial
IoT-Static DB[131]	PostgreSQL, Casandra, MongoDB	Generic cluster of statistical databases	Implementation	Better performance with faster response	Limited by the existing implemented commands	Analyzing	Routing taxis at Beijing

TABLE 12 Existing architectures on 5G-driven IoT big data and data mining

Architecture name	Methods	Highlight	Intended plan	Pros	Cons	Big data/data mining task	Employment
GOVA[132]	Spark, GraphX	BDA architecture for smart grid	Design	System planning and storage scalability. Using graph database.	Performance depends on data compatibility.	Storage	Smart grid systems
Health system architecture[133]	Cloud	Health IoT based big-health system application	Design	Cloud service layer	-	Storage, processing, analyzing	Health care
IV-Tier architecture[134]	Spark, MapReduce	Smart city development using BDA technologies	Implementation	Efficient and scalable	Performance not compared to non-Hadoop systems	Prediction, processing, classification, analyzing	Smart city and planning urban spaces
TreSight architecture[135]	FI-WARE Technology	FI-WARE technology-based recommender	Implementation	Cost optimization	-	Analyzing	Smart Tourism
Behra architecture[136]	Hive, MapReduce, Pig, HDFS	Generalized architecture for storage and analysis	Design	Real-time and efficient data management, reduced cost of data acquisition, computing at cloud	Security of data not addressed	Prediction, processing, classification, analyzing, storage, transmission	Efficient power management
Rathore architecture[137]	MapReduce, HDFS, Spark	Real-time medical tracking with medical sensors	Implementation	Real-time response system	Hadoop used when data volume is too low	Storage, processing, analyzing	Real time response systems like medical emergency
Kholod architecture[106]	Dxelopes	Algorithm for distributed data mining	Implementation	Increased processing speed and efficiency by using cloud and fog systems	-	Storage, processing, analyzing	Applicable on time series data, breast cancer data, weather and telescope data
Din architecture[138]	Hadoop, MapReduce	Data architectures for big data	Implementation	Parallel processing	Hadoop is more suitable for applications with high throughput	Storage, prediction, analyzing	Health care
Suciu architecture[139]	Slapos, Exaleal cloud view	M2M cloud IoT	Implementation	Early disease detection in grape farm by using M2M	Not intelligent enough big data analysis	Storage, processing	Viticulture
VCC architecture[140]	Riak	E-health solution based on IoT and big data	Design	Riak database	-	Classification, storage, processing	Health care
Sowe architecture[141]	Clustering, knowledge extraction	SCN combined with cloud computing	Implementation	Heterogeneous sensors	Mobile sensing not covered	Analyzing, processing	Air pollution
COT[142]	Open CV, Raspberry Pi	Cooperative big data processing	Implementation	Edge devices in processing parallelly	Lacking collective intelligence algorithm for edge	Analyzing, prediction	Face detection

TABLE 13 Existing applications on IoT big data and DM

Application name	Methods	Highlight	Intended plan	Pros	Cons	big data/data mining task	Employment
Saenko et al[143]	HDFS, YARN, MR	Event based complex processing leading to secure BD processing	Implementation	Real-time and scalable	-	Storage, processing	Monitoring securely
IoT based BD model[144]	HDFS	BD processing for IoT data	Implementation	Parallel query execution reducing time and cost with reduced redundancy	-	Classification, processing	Not mentioned in article
Rizwan et al[145]	Make model	Smart traffic management system in real-time with reduced cost	Implementation	Low-cost and real-time	Inaccuracy of the model in calculating char characteristics	Analysis	Traffic management system in real-time
Dineshkumar et al[146]	HDFS, MR, Arduino, Galileo by Intel	BDA on healthcare IoT	Implementation	Quick response system useful for alert type services	-	Storage, processing	Healthcare IoT
Tang et al[147]	Bayesian	Reliability evaluation system for ensuring product quality	Design	-	Storage mechanisms involved not discussed in detail	Analysis	Index system
Niyato et al[148]	SVMs, Logistics Regression	Optimizes complex operations of the system using large sets of data	Analysis numerically	Proposed model shows to be useful in achieving maximum profit for IoT Big Data service providers	-	Analysis	Analysis of market model
Alam analysis[149]	KNN, SVM, ANN and so forth	Data mining methods studied in IoT domain	Simulation	Investigates efficacy of data mining tools and techniques	Not tested with large data set	Analysis	UCI repository of data
Khorshed[150]	Random forest, LMT, REPTree	Security threat identification	Programming	-	Accuracy is not cent-percent	Analysis	Cyber-attacks identification
Rathore et al[151]	HDFS, Spark, Giraph	Graph-oriented methodology proposed	Implementation	Increased scalability with real-time and efficient processing of graphs	-	Clustering, storage, processing, analysis	Smart transport system
Kang et al[152]	MongoDB shard key	MongoDB based RFID sensor data repository	Simulation	Scalable horizontally	Complex query logic with shard extension not supported	Descriptive, storage, processing	Supply chain management

cancer data, weather, and telescope data. Din architecture[138] proposes a dynamic bigdata based architecture that provides parallel processing providing storage, prediction, and analytical solution. It finds application in healthcare IoT systems. Such architecture[139] features the M2M cloud IoT remote telemetry platform for radio communications, remote control and efficient low power long lasting applications like viticulture. It is employed for disease prediction on grape farms and improve overall product quality of the farm. VCC[39] architecture proposed a method for tracking chronic patients and mainly contributes by providing an e-health solution based on IoT and big data. Riak database, a distributed storage is used for scalability. The research is motivated for storage rather than data analysis. Sowe[141] architecture is used for management of complex heterogenous sensors and the services. It uses SCN in the cloud and is deployed mainly to manage data collected by the sensors to monitor air pollution. COT[142] architecture highlights cooperative big data processing. This platform consists of the COT server, analysis server, task manager, COT scheduler, and edge analysis that allows message collection, filtering, and preprocessing.

5.2.3 | BDM applications

This subsection investigates the important studies focusing on the data mining applications that are proposed for an IoT environment. Table 13 provides detailed information on the existing frameworks on big IoT data mining.

Saenko et al[143] proposed an event based complex parallel processing system for secure big IoT data streaming applications. The system offers proper scalability and real time processing of security events. The IoT based model proposed by Bera[144] offers parallel query execution for big IoT data processing and thus reduces time and cost with reduced redundancy. Primary data mining techniques employed in this method is classification and processing. Rizwan[145] employed big data mining techniques in IoT environment for smart traffic management. The system presents an intelligent low-cost solution for real-time gathering of traffic data from various devices and perform analysis to identify traffic congestion and derive suggestive actions. Dineshkumar et al[146] proposes a real-time analysis on health data of multiple patients that is collected and stored on the cloud. Analysis is done on HDFS using MapReduce. Tang et al[147] proposes a design for reliability evaluation system for ensuring product quality covering database management, analysis on test data and serves online test requests.

Niyato et al[148] describes an analytical study to propose a plan to optimize complex operations like pricing on big data based market model. The model has potential benefits in achieving maximum profits for IoT big data service providers. In the research presented by Alam et al,[149] various data mining algorithms have been simulated to draw a comparative understanding of their efficacy and applicability in the IoT domain. Khorshed et al[150] presents an investigation for detection of cyber-attacks identification. As per the research, 93.9% of the activities and cyber-attacks can be identified by random forest algorithm in a complex environment such as IoT. Rathore et al[151] performed an experimentation using graph-oriented methodologies in big IoT data. Multiple sensors identify the vehicle speed and location information to analyze and obtain real time output. A MongoDB based RFID sensor data repository has been proposed by Kang et al[152] which is generated from a simulation setting of supply chain of automotive parts. The model offers horizontal scalability however it has a complex query logic and shard extension is not supported.[157]

5.3 | Big data mining challenges in IoT

Data mining techniques have proven to be efficient in providing tailored predictive and descriptive analytical solutions for BD and can be extended for new data.[158] Emergence of cloud computing technologies and IoT big data enabling mechanisms to have led to new challenges of data investigation and information gathering.[159] Figure 15 presents and overview of the primary research challenges with respect to Big IoT data mining and data processing.

The highly exhaustive amounts of reads and writes performed on high volume data at high velocity creates issues in heterogeneous communication, data extraction, integration and analysis of data. The challenges are bigger due to highly varied big IoT data which is collected as they impose new requirements in the data mining process.[160-162] Big data is also likely to have more ambiguities and irregularities and abnormalities with outliers and thus require more preprocessing stages. Extracting meaningful information is of utmost importance since it's the sole purpose of data mining and to do so from large data volumes with accuracy adds to the issues of data mining with big IoT data. Obtaining meaningful and accurate insights from complex heterogeneous data requires rigorous analysis of data characteristics and identifying the correlation among different data instances.[163-166]

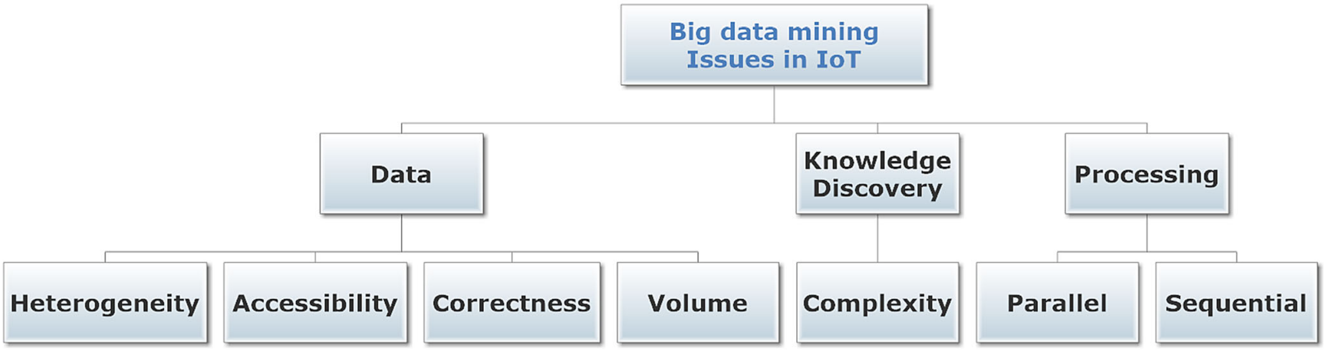


FIGURE 15 Big data mining issues in IoT

The optimization of the big data is done by using various sequential and parallel query execution models to achieve quicker results.[167] Many researches directed in big data mining have been optimized to improve selective factors like improving a data discovery from single knowledge source, implementing algorithms for multisource architectures, streaming data[168] analysis, dynamic data mining and so forth. Methods like parallel association rule mining[169] and parallel K-means have been in use. But such methods limit the compatibility of these algorithms with the latest state-of-the-art platforms with parallel execution capabilities. Parallel computing systems also come with the challenges of synchronization during information exchange among various data mining methods. Data mining bottlenecks have become one of the key challenges in data analytics of big data collected from IoT systems and it must be addressed.[170-173]

6 | DISCUSSION

In this manuscript, the authors have elaborately discussed the mushrooming growth of 5G-enabled IoT devices as technology moves toward Industry 4.0.[174] The digitization of cyber-physical systems has enabled interconnection and communication globally which has resulted in the generation of huge amount of unstructured data. Therefore, the authors have talked about the necessity to organize and preprocess this burgeoning unstructured data into useful information so that suitable information-driven IoT devices can be deployed to function efficiently. Furthermore, the authors have provided an extensive exploration of different BD algorithms & applications in 5G environment. Moreover, the state-of-the-art analytical systems and their pertaining challenges specific to the IoT applications are discussed. In addition, the article also covers the BD security aspect which is vital while conducting varied data streaming processes. Lastly, the issues related to information mining processes are elucidated in this article for better insights on the existing issues of handling big data in 5G-enabled IoT for the development of sustainable smart city infrastructure.[175]

6.1 | Domain applicability of BDA in 5G-IoT

To understand the BDA technological advances and its applications in the 5G-enabled IoT architecture Figure 4 list out various applications of BDA in the IoT and their perceived benefits. The Figure 4 shows what impacting domains are needed to be addressed for a comprehensive functioning of the smart city potential. Many smart units such as smart grids, ITS, smart healthcare, smart inventory system will enable the automation of many such processes in the real-time operation enabling a digital ecosystem.[176]

6.2 | Architectural and functional overview of BDA

Moreover, the BDA of the sensor data generated by the IIoT systems can be logically viewed in the Figure 5. In addition, Figure 6 also presents detailed BDA taxonomy of BDA Solutions in IoT systems. The big data flows from the physical systems to the other fellow physical systems through some gateways or transmitters called network devices such as routers, gateways and so forth.[177] These network devices are responsible for routing and publishing messages to the other end. There are various wireless communication networks which act as carriers of the message transmitted on the network such as Wi-Fi, RFID which are discussed in the Section 2.2. The sensor generated heavy data is stored on centralized cloud storages and on the edge devices installed in the near proximity of the sensors. These edge devices store and also process the relevant at the intermediate level of the IoT network hierarchy. Further processing and analysis of this data is done on cloud servers through which productive insights are delivered to the users via various web-based dashboard using different APIs. The components of hierarchical process of BDA execution is shown in Figure 8.

6.3 | Research and security challenges

BDA will certainly assist businesses by increasing efficiency, optimizing production cost in industries, streamlining supply chain management and so forth but there are some research challenges concerning to these process which are discussed in the Table 7. Section 6.3 addresses these research challenges by expressing a huge scope of improvement in these challenges to fully utilize the possibilities of BDA technologies in the IIoT domain.[178] Moreover, the BDA technologies have enormous amount of opportunities springing up to realize cyber-physical systems in real-time.

BDA processes initiation has also emphasized on the need of the requirement of various security standards for operating on networked IIoT infrastructure.[179-181] The security architecture shown in the Figure 11 is built keeping in mind the BDA classification and Table 9 discusses the different security management frameworks for big data streaming in 5G-driven IoT.

Lastly, Section 5 discusses the applications of data mining in 5G-driven IoT environment. Figure 14 enlists an overview of some strategies that will meet the needs of BDA in IoT environment to facilitate more efficient and optimized solutions. Table 10 discusses various applications in which different data mining techniques play role. Finally, the different case studies and prospective future challenges are talked about in the subsequent sections.

7 | CONCLUSION

5G-enabled IoT has tremendous potential and applications when big data generated by devices in an IoT network are put into analysis to find meaningful insights since they derive actionable intelligence. In this work, we have specifically analyzed and studied recent literature in the context of big data processing and 5G-driven IoT analytics solutions that create an impact in IoT and Industrial IoT systems for the development of sustainable smart city solutions. First, the research presents a taxonomy of the big data analytics solutions and studies the various 5G-driven IoT analytics systems in recent literature. Second, we learnt that different algorithms are suited toward different needs of smart cities infrastructure and applications due to their own pros and cons and that there are some common challenges that upcoming research work must focus on. Third, we learnt about the security concerns, threats and risks, protection methodologies and security management frameworks, and the research challenges. Proper implementation of law and governance is one of the preferred solutions to big data streaming security and we need well-designed policies in 5G-enabled sustainable smart city infrastructure. Fourth, a detailed comparative evaluation of recent state-of-the-art data mining frameworks, architectures and implementations for big data analytic solutions has been presented based on various parameters. Finally, we can conclude that the various technologies, methods and mechanisms for BDA in 5G-enabled IoT systems covered in this work, can help developers choose an existing algorithm or create a new one as per the need by working on the advantages of the already existing systems and eliminating their shortcomings.

CONFLICT OF INTEREST

The authors do not have any competing conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Shashank Gupta https://orcid.org/0000-0002-2124-9388

REFERENCES

1. Akkaya I, Derler P, Emoto S, Lee EA. Systems engineering for industrial cyber–physical systems using aspects. *Proc IEEE*. 2016;104(5):997-1012.
2. Wollschlaeger M, Sauter T, Jasperneite J. The future of industrial communication: automation networks in the era of the internet of things and industry 4.0. *IEEE Ind Electron Mag*. 2017;11(1):17-27.
3. Pedersen MR, Nalpantidis L, Andersen RS, et al. Robot skills for manufacturing: from concept to industrial deployment. *Robot Comput-Integr Manuf*. 2016;37:282-291.
4. Howard P. Concentric analytics for IoT [Online], 2017.
5. Rehman MH, Khan AR, Batool A. Big data analytics in mobile and cloud computing environments. *Innovative Research and Applications in Next-Generation High Performance Computing*. Pennsylvania: IGI Global; 2016:349-367.
6. Ehret M, Wirtz J. Unlocking value from machines: business models and the industrial internet of things. *J Mark Manag*. 2017;33(1–2):111-130.
7. Tsai C-W. Big data analytics: a survey. *J Big Data*. 2015;2(1):1-32.
8. Babiker Mohamed M, Matthew Alofe O, Ajmal Azad M, Singh Lallie H, Fatema K, Sharif T. A comprehensive survey on secure software-defined network for the internet of things. *Trans Emerging Tel Tech*. 2022;33(1):e4391. doi:10.1002/ett.4391
9. Dhar Dwivedi A, Singh R, Kaushik K, Rao Mukkamala R, Alnumay WS. Blockchain and artificial intelligence for 5G-enabled internet of things: challenges, opportunities, and solutions. *Trans Emerging Tel Tech*. 2021;e4329. doi:10.1002/ett.4329
10. Al-Turjman F, Baali I. Machine learning for wearable IoT-based applications: a survey. *Trans Emerging Tel Tech*. 2019;e3635. doi:10.1002/ett.3635
11. Kumari A, Tanwar S, Tyagi S, Kumar N. Verification and validation techniques for streaming big data analytics in internet of things environment. *IET Netw*. 2019;8(3):155-163. doi:10.1049/iet‐net.2018.5187
12. Ge M, Bangui H, Buhnova B. Big data for internet of things: a survey. *Futur Gener Comput Syst*. 2018;87:601-614.
13. Mohammadi M, Al-Fuqaha AI, Sorour S, Guizani M. Deep learning for IOT big data and streaming analytics: a survey. *IEEE Commun Surv Tutorials*. 2018;20(4):2923-2960.
14. Khan MA, Salah K. IOT security: review, blockchain solutions, and open challenges. *Futur Gener Comput Syst*. 2018;82:395-411.
15. Joshi N, Kadhiwala B. Big data security and privacy issues-a survey. Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT); 2017:1-5; Vellore, India.
16. Chandra S, Ray S, Goswami RT. Big data security: survey on frameworks and algorithms. Proceedings of the 2017 IEEE 7th Int Advance Computing Conference (IACC); 2017:48-54; Telangana, India.
17. Nawir M, Amir A, Yaakob N, Lyn OB. Internet of things (IOT): taxonomy of security attacks. 2016 3rd Int Conference on Electronic Design (ICED); 2016:321-326; Phuket, Thailand.
18. Ida IB, Jemai A, Loukil, A. A survey on security of IOT in the context of ehealth and clouds. Proceedings of the 2016 11th International Design & Test Symposium (IDT); 2016:25-30; Hammamet, Tunisia
19. Matturdi B, Zhou X, Li S, Lin F. Big data security and privacy: a review. *China Commun*. 2014;11(14):135-145.
20. Positive 5G outlook post COVID-19: what does it mean for avid gamers? Forest interactive. Retrieved November 13, 2020.
21. Hoffman C. What is 5G, and how fast will it be? How-to geek website. how-to geek LLC. Archived from the original on January 24, 2019. Retrieved January 23, 2019.
22. Ahmed E, Yaqoob I, Hashem IAT, et al. The role of big data analytics in internet of things. *Comput Netw*. 2017;129(part 2):459-471. doi:10.1016/j.comnet.2017.06.013
23. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. *J Internet Serv Appl*. 2015;6(1):1.
24. Ahmed E, Imran M, Guizani M, et al. Enabling mobile and wireless technologies for smart cities: part 2. *IEEE Commun Mag*. 2017;55(3):12-13.
25. Suciu G, Suciu V, Martian A, et al. Big data, internet of things and cloud convergence–An architecture for secure e-health applications. *J Med SystJ Med Syst*. 2015;39(11):1-8.
26. Hashem IAT, Chang V, Anuar NB, et al. The role of big data in smart city. *Int J Inf Manag*. 2016;36(5):748-758.
27. Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of things (IoT): a vision, architectural elements, and future directions. *Futur Gener Comput Syst*. 2013;29(7):1645-1660.
28. Duan R, Chen X, Xing T. A QoS architecture for IOT. Proceedings of the 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing; 2011:717-720; IEEE.
29. Zhang Y. ICN based architecture for IoT. IRTF Contrib; 2013.
30. Ahmed E, Rehmani MH. Mobile edge computing: opportunities, solutions, and challenges. *Futur Gener Comput Syst*. 2017;70:59-63.
31. Ahmed A, Ahmed E, A survey on mobile edge computing. Proceedings of the 2016 10th International Conference on Intelligent Systems and Control (ISCO); 2016:1-8; IEEE.
32. Jararweh Y, Doulat A, AlQudah O, Ahmed E, Al-Ayyoub M, Benkhelifa E. The future of mobile cloud computing: integrating cloudlets and mobile edge computing. Proceedings of the 2016 23rd International Conference on Telecommunications (ICT); 2016:1-5; IEEE.
33. Wang K. Intelligent predictive maintenance (ipdm) system–industry 4.0 scenario. *WIT Trans Eng Sci*. 2016;113(1):259-268.
34. Wang L, Törngren M, Onori M. Current status and advancement of cyber-physical systems in manufacturing. *J Manuf Syst*. 2015;37(Part 2):517-527.
35. Zhou K, Liu T, Liang L. From cyber-physical systems to industry 4.0: make future manufacturing become possible. *Int J Manuf Res*. 2016;11(2):167-188.

36. Lee J, Ardakani HD, Yang S, Bagheri B. Industrial big data analytics and cyber-physical systems for future maintenance & service innovation. *Proc CIRP*. 2015;38:3-7.
37. Cheng L, Wang T, Hong X, Wang Z, Wang J, Liu J. A study on the architecture of manufacturing internet of things. *Int J Model Identif Control*. 2015;23(1):8-23.
38. Zhong RY, Xu C, Chen C, Huang GQ. Big data analytics for physical internet-based intelligent manufacturing shop floors. *Int J Prod Res*. 2017;55(9):2610-2621.
39. Hossain M, Muhammad G. Cloud-assisted industrial internet of things (iiot)–enabled framework for health monitoring. *Comput Netw*. 2016;101:192-202.
40. Georgakopoulos D, Jayaraman PP, Fazia M, Villari M, Ranjan R. Internet of things and edge cloud computing roadmap for manufacturing. *IEEE Cloud Comput*. 2016;3(4):66-73.
41. Gupta H, Dastjerdi AV, Ghosh SK, Buyya R. iFogSim: a toolkit for modeling and simulation of resource management techniques in internet of things. *Edge Fog Comput Environ*. 2016;42:1-22.
42. Rehman MH, Ahmed E, Yaqoob I, Hashem IAT, Imran M, Ahmad S. Big data analytics in industrial iot using a concentric computing model. *IEEE Commun Mag*. 2018;56(2):37-43.
43. Press G. Cleaning big data: most time-consuming, least enjoyable data science task, survey says, Online, 2016.
44. Wiener P, Stein M, Seebacher D, Bruns J, Frank M, Simko V, Zander S, Nimis J. Biggis: a continuous refinement approach to master heterogeneity and uncertainty in spatio-temporal big data (vision paper). Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems; 2016:8; ACM, New York.
45. Bai M, Wang X, Xin J, Wang G. An efficient algorithm for distributed density-based outlier detection on big data. *Neurocomputing*. 2016;181:19-28.
46. Stojanovic L, Dinic M, Stojanovic N, Stojadinovic A. Big-data-driven anomaly detection in industry (4.0): an approach and a case study. Proceedings of the 2016 IEEE International Conference on Big Data (Big Data); 2016:1647-1652; IEEE.
47. Niño M, Sáenz F, Blanco JM, Illarramendi A. Requirements for a big data capturing and integration architecture in a distributed manufacturing scenario. Proceedings of the 2016 IEEE 14th International Conference on Industrial Informatics (INDIN); 2016:1326-1329; IEEE.
48. Wang R, Sun D, Li G, Atif M, Nepal S. Logprov: logging events as provenance of big data analytics pipelines with trustworthiness. Proceedings of the IEEE Conference on Big Data; 2016.
49. Yamanishi K, Miyaguchi K. Detecting gradual changes from data stream using MDL-change statistics. Proceedings of the 2016 IEEE International Conference on Big Data (big data); 2016:156-163; IEEE.
50. Cheng X, Fang L, Hong X, Yang L. Exploiting mobile big data: sources, features, and applications. *IEEE Netw*. 2017;31(1):72-79.
51. Lee J. Smart factory systems. *Inform-Spektrum*. 2015;38(3):230-235.
52. Yang C, Lan S, Shen W, Huang GQ, Wang X, Lin T. Towards product customization and personalization in iot-enabled cloud manufacturing. *Clust Comput*. 2017;20:1717-1730.
53. Wan J, Tang S, Li D, et al. A manufacturing big data solution for active preventive maintenance. *IEEE Trans Ind Inf*. 2017;13:2039-2047.
54. Gamarra C, Guerrero JM, Montero E. A knowledge discovery in databases approach for industrial microgrid planning. *Renew Sustain Energy RevRenew Sustain Energy Rev*. 2016;60:615-630.
55. Saldivar AAF, Goh C, Chen WN, Li Y. Self-organizing tool for smart design with predictive customer needs and wants to realize industry 4.0. Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC); 2016:5317-5324; IEEE.
56. Saldivar AAF, Goh C, Li Y, Chen Y, Yu H. Identifying smart design attributes for industry 4.0 customization using a clustering genetic algorithm. Proceedings of the 2016 22nd International Conference on Automation and Computing (ICAC); 2016:408-414; IEEE.
57. Zurita D, Delgado M, Carino JA, Ortega JA, Clerc G. Industrial time series modelling by means of the neo-fuzzy neuron. *IEEE Access*. 2016;4:6151-6160.
58. Rehman MH u, Yaqoob I, Salah K, Imran M, Jayaraman PP, Perera C. The role of big data analytics in industrial internet of things, future generation. *Comput Syst*. 2019;99:247-259. doi:10.1016/j.future.2019.04.020
59. Cao J, Cui H, Shi H, Jiao L. Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on mapreduce. *PLoS One*. 2016;11(6):e0157551.
60. Dopico M, Gomez A, De la Fuente D, García N, Rosillo R, Puche J. A vision of industry 4.0 from an artificial intelligence point of view. Proceedings on the International Conference on Artificial Intelligence (ICAI), the Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2016:407.
61. Berger-Wolf TY. Research: computational ecology, social network analysis, data science. *J Parallel Distrib Comput*. 2016;93:66-86.
62. Rehman MH, Liew CS, Abbas A, Jayaraman PP, Wah TY, Khan SU. Big data reduction methods: a survey. *Data Sci Eng*. 2016;1:265-284.
63. Martellini M, Abaimov S, Gaycken S, Wilson C. Assessing cyberattacks against wireless networks of the next global internet of things revolution: industry 4.0. *Information Security of Highly Critical Wireless Networks*. New York: Springer; 2017:63-69.
64. de Farias CM, Pirmez L, Fortino G, Guerrieri A. A multi-sensor data fusion technique using data correlations among multiple applications, Future Gener. *Comput Syst*. 2019;92:109-118.
65. Wu P-Y, Cheng C-W, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. Omic and electronic health record big data analytics for precision medicine. *IEEE Trans Biomed Eng*. 2017;64(2):263-273.
66. Wan J, Tang S, Li D, et al. Reconfigurable smart factory for drug packing in healthcare industry 4.0. *IEEE Trans Ind Inf*. 2019;15(1):507-516.

67. Yaqoob I, Hashem IAT, Ahmed A, Kazmi SA, Hong CS. Internet of things forensics: recent advances, taxonomy, requirements, and open challenges. *Future Gener Comput Syst*. 2019;92:265-275.
68. Yaqoob I, Ahmed E, Rehman MH u, et al. The rise of ransomware and emerging security challenges in the internet of things. *Comput Netw*. 2017;129:444-458.
69. Tanwar S, Patel P, Patel K, Tyagi S, Kumar N, Obaidat MS. An advanced internet of thing based security alert system for smart home. Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS); 2017:25-29; Dalian, China.
70. Gantz J, Reinsel D. Extracting value from chaos. *IDC iView*. 2011;2011(1142):1-12.
71. Vora J, DevMurari P, Tanwar S, Tyagi S, Kumar N, Obaidat MS. Blind signatures based secured ehealthcare system. Proceedings of the 2018 International Conference on Computer, Information and Telecommunication Systems (CITS); 2018:1-5; Alsace, Colmar, France.
72. Vora J, Italiya P, Tanwar S, et al. Ensuring privacy and security in Ehealth records. Proceedings of the 2018 International Conference on Computer, Information and Telecommunication Systems (CITS); 2018:1-5; Alsace, Colmar, France.
73. John-Walker S. Big data: a revolution that will transform how we live, work, and think. *Int J Internet Mark Advert Rev Market Commun*. 2014;33:181-183.
74. Ardagna C, Damiani E. Business intelligence meets big data: an overview on security and privacy. Proceedings of the NSF Workshop on Big Data Security and Privacy; 2014:1-6; Dallas, TX
75. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S. The rise of big data on cloud computing: review and open research issues. *Inf Syst*. 2015;47:98-115.
76. Loshin D. Data governance for big data analytics: considerations for data policies and processes. *Big Data Analytics*. Boston, Burlington, MA: Morgan Kaufmann; 2013:39-48.
77. Malik P. Governing big data: principles and practices. *IBM J Res Dev*. 2013;57(3/4):1-13.
78. Abawajy J. Symbioses of big data and cloud computing. Opportunities Challenges; 2013.
79. Tallon PP. Corporate governance of big data: perspectives on value, risk, and cost. *Computer*. 2013;46(6):32-38.
80. Shameli-Sendi A, Aghababaei-Barzegar R, Cheriet M. Taxonomy of information security risk assessment (ISRA). *Comput SecurComput Secur*. 2016;57:14-30.
81. Saleh ZI, Refai H, Mashhour A. Proposed framework for security risk assessment. *J Inf Secur*. 2011;2(2):85.
82. Slater WF III. Vulnerability Management; 2016.
83. Michael K. Security risk management: building an information security risk management program from the ground up; 2012.
84. Ross-Ronald S. Guide for conducting risk assessments', NIST SP-800-30rev1; 2012
85. Shameli-Sendi A, Ezzati-Jivan N, Jabbarifar M, Dagenais M. Intrusion response systems: survey and taxonomy. *Int J Comput Sci Netw Secur*. 2012;12(1):1-14.
86. Curry S, Kirda E, Schwartz E, Stewart WH, Yoran A. Big data fuels intelligence-driven security. *RSA Secur Brief*. 2013.
87. Sabeel U, Maqbool S. Categorized security threats in the wireless sensor networks: countermeasures and security management schemes. *Int J Comput Appl*. 2013;64(16):19-28.
88. Hossain MM, Fotouhi M, Hasan R. Towards an analysis of security issues, challenges, and open problems in the internet of things. Proceedings of the 2015 IEEE World Congress on Services (SERVICES); 2015:21-28; IEEE, New York.
89. Mayzaud A, Badonnel R, Chrisment I. A taxonomy of attacks in RPLbased internet of things. *Int J Netw Secur*. 2016;18(3):459-473.
90. Alam S, De D. Analysis of security threats in wireless sensor network. arXiv Preprint arXiv:14060298, 2014.
91. Mahmood T, Afzal U. Security analytics: big data analytics for cybersecurity: a review of trends, techniques and tools. Proceedings of the 2013 2nd National Conference on Information Assurance (NCIA); 2013:129-134; Rawalpindi, Pakistan.
92. Alouneh S, Al-Hawari F, Hababeh I, Ghinea G. An effective classification approach for big data security based on GMPLS/MPLS networks. *Secur Commun Netw*. 2018;2018:1-10.
93. Yu W, Yu P, Wang J, et al. Protecting your own private key in cloud: security, scalability and performance. Proceedings of the 2018 IEEE Conference on Communications and Network Security (CNS); 2018:1-2; IEEE; Beijing, China
94. Aljawarneh SA, Alawneh A, Jaradat R. Cloud security engineering: early stages of SDLC. *Futur Gener Comput Syst*. 2017;74:385-392.
95. Kalaiprasath R, Elankavi R, Udayakumar DR. Cloud security and compliance – A semantic approach in end to end security. *Int J Smart Sens Intell Syst*. 2017;8(5):482-494.
96. Lafuente G. The big data security challenge. *Netw Secur*. 2015;2015(1):12-14.
97. Kaleeswari C, Maheswari P, Kuppusamy K. A brief review on cloud security scenarios. *Int J Sci Res Sci Technol*. 2018;4:46-50.
98. Vora J, Tanwar S, Tyagi S, et al. FAAL: fog computing-based patient monitoring system for ambient assisted living. Proceedings of the 19th IEEE International Conference on eHealth Networking, Applications and Services (Healthcom); 2017:1-6; Dalian, China
99. Kaur P, Sharma M, Mittal M. Big data and machine learning based secure healthcare framework. *Proc Comput Sci*. 2018;132:1049-1059.
100. Kumari A, Tanwar S, Tyagi S, Kumar N. Fog computing for healthcare 4.0 environment: opportunities and challenges. *Comput Electr Eng*. 2018;72:1-13.
101. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. *J Big Data*. 2018;5:1.
102. Barkham R, Bokhari S, Saiz A. *Urban Big Data: City Management and Real Estate Markets*. New York, NY: GovLab Digest; 2018.
103. Kamalinejad P, Mahapatra C, Sheng Z, Mirabbasi S, M. Leung VC, Guan YL. Wireless energy harvesting for the internet of things. *IEEE Commun Mag*. 2015;53(6):102-108.
104. Li X, Jiang P, Chen T, Luo X, Wen Q. A survey on the security of blockchain systems. *Futur Gener Comput Syst*. 2020;107:841-853. doi:10.1016/j.future.2017.08.020

105. Belle I. The architecture, engineering and construction industry and blockchain technology. (DADA 2017). Proceeding of the International Conference on Digital Architecture at Nanjing University; 2017:279-284; Nanjing, China.
106. Kholod I, Kuprianov M, Petukhov I. Distributed data mining based on actors for internet of things. Proceedings of the 2016 5th Mediterranean Conference on Embedded Computing (MECO), 2016.
107. Siddiqa A, Hashem IAT, Yaqoob I, et al. A survey of big data management: taxonomy and state-of-the-art. *J Netw Comput Appl*. 2016;71:151-166.
108. Chen F. Data mining for the internet of things: literature review and challenges. *Int J Distrib Sens Netw*. 2015;2015:14.
109. Siddiqa A. A survey of big data management: taxonomy and state-ofthe-art. *J Netw Comput Appl*. 2016;71:151-166.
110. Bifet A, Holmes G, Kirkby R, Pfahringer B. MOA: massive online analysis. *J Mach Learn Res*. 2010;11:1601-1604.
111. Shadroo S, Rahmani AM. Systematic survey of big data and data mining in internet of things. *Comput Netw*. 2018;139:19-47. doi:10.1016/j.comnet.2018.04.001
112. Marjani M. Big IoT data analytics: architecture, opportunities, and open research challenges. *IEEE Access*. 2017;5:5247-5261. doi:10.1109/ACCESS.2017.2689040
113. Estivill-Castro V. Why so many clustering algorithms: a position paper. *ACM SIGKDD Explor Newslett*. 2002;4(1):65-75.
114. Larose DT. K-Nearest Neighbor Algorithm. *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, New Jersey: Wiley; 2005:90-106.
115. Srivastava K. Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment. *Int J Comput Theory Eng*. 2013;5(3):520.
116. Gosain A, Bhugra M. A comprehensive survey of association rules on quantitative data in data mining. Proceedings of the IEEE Conference Infernational Communication Technology; April 2013:1003-1008.
117. Fitzwater M. Efficient mining of maximal sequential patterns using multiple samples. Proceedings of the SIAM International Conference Data Mining; 2005:1.
118. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manag*. 2015;35(2):137-144.
119. Kalpakis K, Gada Puttagunta DV. Distance measures for effective clustering of ARIMA time-series. Proceedings of the IEEE International Conference on Data Mining (ICDM); December 2001:273-280.
120. Ryan D. *High Performance Discovery in Time Series: Techniques and Case Studies*. Berlin, Germany: Springer; 2013.
121. Sheth K, Patel K, Shah H, Tanwar S, Gupta R, Kumar N. A taxonomy of AI techniques for 6G communication networks. *Comput Commun*. 2020;161:279-303.
122. Aggarwal S, Kumar N, Tanwar S. Blockchain-envisioned UAV communication using 6G networks: open issues, use cases, and future directions. *IEEE Internet Things J*. 2020;8(7):5416-5441.
123. Ray PP. A review on 6G for space-air-ground integrated network: Key enablers, open challenges, and future direction. *J King Saud Univ Comput Inf Sci*. 2021.
124. Guo K, Tang Y, Zhang P. CSF: crowdsourcing semantic fusion for heterogeneous media big data in the internet of things. *Inf Fusion*. 2017;37:77-85.
125. Vögler M, Schleicher JM, Inzinger C, Dustdar S. Ahab: a cloud-based distributed big data analytics framework for the internet of things. *Softw Pract Exp*. 2017;47(3):443-454.
126. Sezer OB, Dogdu E, Ozbayoglu M, Onal A. An extended IoT framework with semantics, big data, and analytics. Proceedings of the IEEE International Conference on Big Data (Big Data); 2016.
127. Bashir MR, Gill AQ. Towards an IoT big data analytics framework: smart buildings systems. Proceedings of the 2016 IEEE 18th International con-Ference on 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.
128. Dundar B, Astekin M, Aktas MS. A big data processing framework for self-healing internet of things applications. Proceedings of the 12th International Conference on Semantics, Knowledge and Grids (SKG); 2016.
129. Berlian MH, Sahputra TER, Ardi BJW, Dzatmika LW, Besari ARA, Sudibyo RW. Design and implementation of smart environment monitoring and analytics in real-time system framework based on internet of underwater things and big data. Proceedings of the 2016 International Electronics Symposium (IES); 2016.
130. Mishra N, Chung-Chih L, Hsien-Tsung C. A cognitive oriented framework for IoT big-data management prospective. Proceedings of the IEEE International Conference on Communiction Problem-Solving; 2014.
131. Ding Z, Gao X, Xu J, Wu H. IOT-statistic general statistical database cluster mechanism for big data analysis in the internet of things. Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things; 2013.
132. Dang-Ha TH, Roverso D, Olsson R. Graph of virtual actors (GOVA): a big data analytics architecture for IoT. Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp); 2017.
133. Ma Y, Wang Y, Yang J, Miao Y, Li W. Big health application system based on health internet of things and big data. *IEEE Access*. 2017;5:7885-7897.
134. Rathore MM, Ahmad A, Paul A, Rho S. Urban planning and building smart cities based on the internet of things using big data analytics. *Comput Netw*. 2016;101:63-80.
135. Sun Y, Song H, Jara AJ, Bie R. Internet of things and big data analytics for smart and connected communities. *IEEE Access*. 2016;4:766-773.
136. Behera RK, Gupta S, Gautam A. Big-data empowered cloud centric internet of things. Proceedings of the International Conference on Man and Machine Interfacing (MAMI); 2015.

137. Rathore MM, Ahmad A, Paul A, Wan J, Zhang D. Real-time medical emergency response system: exploiting IoT and big data for public health. *J Med Syst*. 2016;40(12):283.
138. Din S, Ghayvat H, Paul A, Ahmad A, Rathore MM, Shafi I. An architecture to analyze big data in the internet of things. Proceedings of the 9th International Conference on Sensing Technology (ICST); 2015 .
139. Suciu G, Vulpe A, Fratu O, Suciu V. M2M remote telemetry and cloud IoT big data processing in viticulture. Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC); 2015.
140. Páez DG, Aparicio F, de Buenaga M, Ascanio JR. Big data and IoT for chronic patients monitoring, Ubiquitous computing and ambient intelligence. In: Hervás R, Lee S, Nugent C, Bravo J, eds. *Personalisation and User Adapted Services*. Cham: Springer; 2014:416-423.
141. Sowe S, Kimata T, Dong M, Zettsu K. Managing heterogeneous sensor data on a big data platform: IoT services for data-intensive science. Proceedings of the IEEE 38th Annual International Computers, Software and Applications Conference; 2014; Sweden.
142. Lee Y-J, Park HD, Min O. Cooperative big data processing engine for fast reaction in internet of things environment: greater than the sum of its parts. In: Kim K, Wattanapongsakorn N, Joukov N, eds. *Mobile and Wireless Technologies 2016. Lecture Notes in Electrical Engineering*. Singapore: Springer; 2016.
143. Saenko I, Kotenko I, Kushnerevich A. Parallel processing of big heterogeneous data for security monitoring of IoT networks. Proceedings of the 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP); 2017.
144. Bera A, Kundu A, De Sarkar NR, Mou D. *Experimental Analysis on Big Data in IOT-Based Architecture*. Singapore: Springer; 2017.
145. Rizwan P, Suresh K, Babu MR. Real-time smart traffic management system for smart cities by using internet of things and big data. Proceedings of the International Conference on Emerging Technological Trends (ICETT); 2016.
146. Dineshkumar P, SenthilKumar R, Sujatha K, Ponmagal RS, Rajavarman VN. Big data analytics of IoT based health care monitoring system. Proceedings of the IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering; 2016; Varanasi, India.
147. Tang TJ, Chung A, Zhao A, Yang J, Zhang J. An loT inspired semiconductor reliability test system integrated with data-mining applications. Proceedings of the 2016 2nd International Conference on Cloud Computing and Internet of Things (CCIOT); 2016 .
148. Niyato D, Alsheikh MA, Wang P, Kim DI, Han Z. Market model and optimal pricing scheme of big data and internet of things (IoT). Proceedings of the 2016 IEEE International Conference on Communications (ICC); 2016.
149. Alam F, Mehmood R, Katib I, Albeshri A. Analysis of eight data mining algorithms for smarter internet of things (IoT). *Proc Comput Sci*. 2016;98:437-442.
150. Khorshed MT, Sharma NA, Kumar K, Prasad M, Ali ABMS, Xiang Y. Integrating internet-of-things with the power of cloud computing and the intelligence of big data analytics: a three layered approach. Proceedings of the 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE); 2015; Fiji.
151. Rathore MM, Ahmad A, Paul A, Jeon G. Efficient graph-oriented smart transportation using internet of things generated big data. Proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS); 2015.
152. Kang Y, Park I, Rhee J, Lee Y. MongoDB-based repository design for IoT-generated RFID/sensor big data. *IEEE Sens J*. 2016;16(2):485-497.
153. Irram F, Ali M, Naeem M, Mumtaz S. Physical layer security for beyond 5G/6G networks: emerging technologies and future directions. *J Netw Comput Appl*. 2022;103431:103431.
154. Khan LU, Saad W, Niyato D, Han Z, Hong CS. Digital-twin-enabled 6G: vision, architectural trends, and future directions. *IEEE Commun Mag*. 2022;60(1):74-80.
155. Lu Y, Maharjan S, Zhang Y. Adaptive edge association for wireless digital twin networks in 6G. *IEEE Internet Things J*. 2021;8(22):16219-16230.
156. Sun W, Zhang H, Wang R, Zhang Y. Reducing offloading latency for digital twin edge networks in 6G. *IEEE Trans Veh Technol*. 2020;69(10):12240-12251.
157. Lv Z, Lou R, Li J, Singh AK, Song H. Big data analytics for 6G-enabled massive internet of things. *IEEE Internet Things J*. 2021;8(7):5350-5359.
158. Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello CAC. A survey of multiobjective evolutionary algorithms for data mining: part I. *IEEE Trans Evol Comput*. 2014;18(1):4-19.
159. Hu T, Chen H, Huang L, Zhu X. A survey of mass data mining based on cloud-computing. Proceedings of the Anti-Counterfeiting Security, and Identification; 2012:1-4.
160. Sun Y. Mining knowledge from interconnected data: a heterogeneous information network analysis approach. *Proc VLDB Endowm*. 2012;5(12):2022-2023.
161. Chen M, Yang LT, Kwon T, Zhou L, Jo M. Itinerary planning for energy-efficient agent communications in wireless sensor networks. *IEEE Trans Veh Technol*. 2011;60(7):3290-3299.
162. Zhang D. A taxonomy of agent technologies for ubiquitous computing environments. *Trans Internet Inf Syst*. 2012;6(2):547-565.
163. Sodhro AH, Zahid N, Wang L, et al. Toward ML-based energy-efficient mechanism for 6G enabled industrial network in box systems. *IEEE Trans Ind Inform*. 2020;17(10):7185-7192.
164. Hsu CH, Manogaran G, Srivastava G, Chilamkurti N. Guest editorial: 6G-enabled network in box (NIB) for industrial applications and services. *IEEE Trans Ind Inform*. 2021;17(10):7141-7144.
165. Verma S, Kaur S, Khan MA, Sehdev PS. Toward green communication in 6G-enabled massive internet of things. *IEEE Internet Things J*. 2020;8(7):5408-5415.
166. Ji B, Wang Y, Song K, et al. A survey of computational intelligence for 6G: key technologies, applications and trends. *IEEE Trans Ind Inform*. 2021;17(10):7145-7154.

167. Malik UM, Javed MA, Zeadally S, ul Islam S. Energy efficient fog computing for 6G enabled massive IoT: recent trends and future opportunities. *IEEE Internet Things J*. 2021.
168. Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. *IEEE Trans Knowl Data Eng*. 2014;26(1):97-107.
169. Wu X, Zhang S. Synthesizing high-frequency rules from different data sources. *IEEE Trans Knowl Data Eng*. 2003;15(2):353-367.
170. Guo F, Yu FR, Zhang H, Li X, Ji H, Leung VC. Enabling massive IoT toward 6G: a comprehensive survey. *IEEE Internet Things J*. 2021;8(15):11891-11915.
171. Bansal M, Chana I, Clarke S. A survey on iot big data: current status, 13 v's challenges, and future directions. *ACM Comput Surv (CSUR)*. 2020;53(6):1-59.
172. Silva BN, Khan M, Han K. Towards sustainable smart cities: a review of trends, architectures, components, and open challenges in smart cities. *Sustain Cities Soc*. 2018;38:697-713.
173. Jain PK, Saravanan V, Pamula R. A hybrid CNN-LSTM: a deep learning approach for consumer sentiment analysis using qualitative user-generated contents. *Trans Asian Low-Resource Language Inform Process*. 2021;20(5):1-15.
174. Jain PK, Pamula R, Srivastava G. A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Comput Sci Rev*. 2021;41:100413.
175. Jain PK, Quamer W, Saravanan V, Pamula R. Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. *J Ambient Intell Humaniz Comput*. 2022;1-13.
176. Meena P, Pal MB, Jain PK, Pamula R. 6G communication networks: introduction, vision, challenges, and future directions. *Wirel Pers Commun*. 2022;125:1-27.
177. Ahmed U, Lin JCW, Srivastava G. Heterogeneous energy-aware load balancing for industry 4.0 and IoT environments. ACM Transactions on Management Information Systems (TMIS); 2022.
178. Djenouri Y, Srivastava G, Belhadi A, Lin JCW. Intelligent blockchain management for distributed knowledge graphs in IoT 5G environments. *Trans Emerg Telecommun Technol*. 2021;e4332.
179. Syu JH, Wu ME, Srivastava G, Chao CF, Lin JCW. An IoT-based hedge system for solar power generation. *IEEE Internet Things J*. 2021;8(13):10347-10355.
180. Lin JCW, Srivastava G, Zhang Y, Djenouri Y, Aloqaily M. Privacy-preserving multiobjective sanitization model in 6G IoT environments. *IEEE Internet Things J*. 2020;8(7):5340-5349.
181. Tan L, Yu K, Lin L, et al. Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space–air–ground integrated intelligent transportation system. *IEEE Trans Intell Transp Syst*. 2021;23(3):2830-2842.

How to cite this article: Mukherjee S, Gupta S, Rawlley O, Jain S. Leveraging big data analytics in 5G-enabled IoT and industrial IoT for the development of sustainable smart cities. *Trans Emerging Tel Tech*. 2022;33(12):e4618. doi: 10.1002/ett.4618