Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 3740476, 18 pages
https://doi.org/10.1155/2021/3740476

Review Article

A Review of Big Data Resource Management: Using Smart Grid Systems as a Case Study

Muhammad Fawad Khan,[1] Muhammad Azam,[2] Muhammad Asghar Khan,[3] Fahad Algarni,[4] Mujaddad Ashfaq,[5] Ibtihaj Ahmad,[6] and Insaf Ullah[3]

[1]*School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea*
[2]*School of Energy and Power Engineering, Jiangsu University, Zhenjiang 212013, China*
[3]*Department of Electrical Engineering, Hamdard University, Islamabad, Pakistan*
[4]*College of Computing and Information Technology, University of Bisha, Saudi Arabia*
[5]*Department of Electrical Engineering, Foundation University Islamabad, Rawalpindi Campus, Pakistan*
[6]*Department of Computer Science Engineering, Northwestern Polytechnical University, Xi'an, China*

Correspondence should be addressed to Muhammad Asghar Khan; m.asghar@hamdard.edu.pk

Received 31 August 2021; Revised 2 October 2021; Accepted 12 October 2021; Published 22 October 2021

Academic Editor: Suleman Khan

Big data has recently been a prominent topic of research due to the exponential growth of data every year. This massive growth of data is causing problems with resource management. The available literature does not address this problem in depth. Therefore, in this article, we aim to cover the topic of resource management for big data in detail. We addressed resource management from the perspective of smart grids for a better understanding. This study includes a number of tools and methods, such as Hadoop and MapReduce. Large data sets created by smart grids or other data-generating sources may be handled using such tools and approaches. In this article, we also discussed resource management in terms of various vulnerabilities and security risks to data and information being transmitted or received, as well as big data analytics. In summary, our comprehensive study of big data in terms of data creation, processing, resource management, and analysis gives a full picture of big data.

1. Introduction

Over the past 20 years, data has been increasing tremendously in different fields. According to the International Data Corporation (IDC), the total copied and created data volume all over the world was 1.8 ZB (zettabytes), which has increased by approximately nine times within five years [1]. And there is a prediction that in the near future, this figure will double at least every other two years. Considering these statistics, one can well imagine about the drastic growth of big data and the issues related to it. Big data deals with huge data sets mostly in exabytes, zettabytes or yottabytes. Figure 1 can give a better estimate of these mega units that represent an enormous scale of volume. In Figure 1, these higher units of volume are converted to bytes in order to get a better estimation and clear picture of the huge volume of data sets in big data.

The enormous increase of data is generating resource management issues. Resource management is a technique which is used to utilize the resources in efficient way by improving the network throughput, capacity, robustness, and efficiency. Importance of management of resources in data applications is growing day by day.

Big data is linked with a huge amount of data sets. Most of the big data comprise a huge amount of unstructured data sets as compared to traditional data sets that require more real-time data analysis [2]. Additionally, big data help us in categorizing the data looking at different aspects considering the value of the data. It creates new opportunities for effectively organizing and managing such enormous data sets

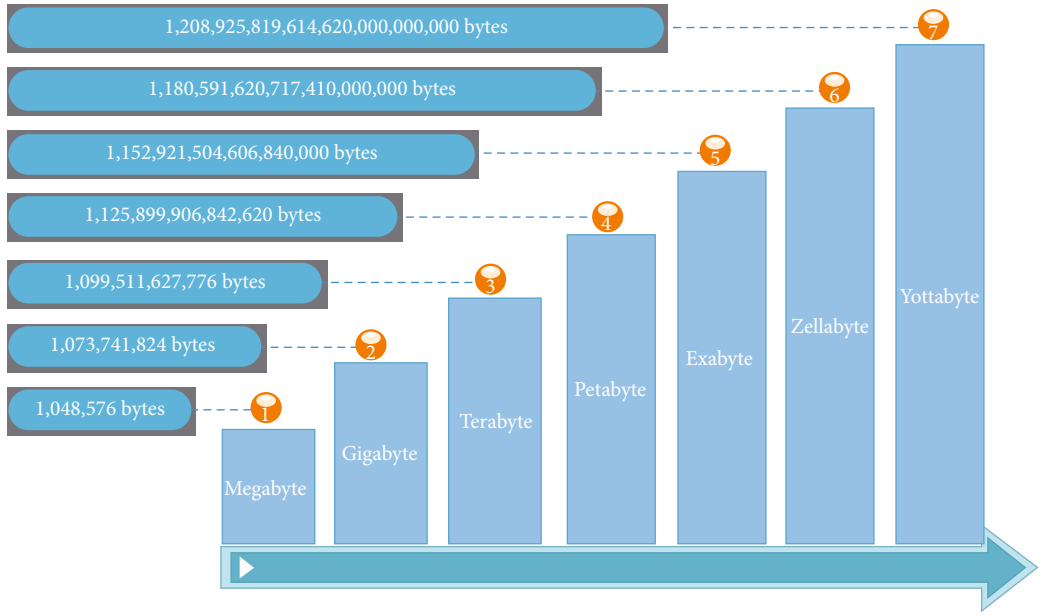

Figure 1: Bigger units of volume converted to bytes [4].

according to their value [1]. There are many applications of big data in the smart world. Everything is shifting from offline to online and cloud systems. Smart grid systems are a technological innovation whose adaptation has recently increased all over the globe. Two-way communication flow makes it more efficient, reliable, sustainable, and cost effective as compared to traditional power grid systems. But two-way data flow of a huge number of sensors is a big challenge for data scientists [3]. Due to limitation of resources, resource management is needed to get the effective results in every field. In many different fields of communication and IT, work on resource management has been done to a great extent. Different aspects of big data have been highlighted in the existing literature as shown in Table 1.

There are four contributions in this article. To begin, we will go through data generation sources. The second contribution is a consideration of the relevance of resource management in the context of smart grids, as well as the sources of big data. Third, we go through the strategies and tools that go into analyzing various cases. Finally, we discuss unresolved difficulties and obstacles in large data analysis in general.

1.1. Data-Generating Sources. Big data includes large data sets produced by different applications and devices. The umbrella of big data covers various fields; some of them are given as follows.

(i) *Black box data*: the black box of jets, airplanes, helicopters, etc. captures voices of the flight crew, performance information, and recordings of microphones of the aircraft

(ii) *Social media data*: social media websites like Facebook, LinkedIn, and Twitter hold the information of millions of people across the globe [5].

(iii) *Stock exchange data*: it also contributes in generating a huge number of data sets comprising the information regarding buying and selling of shares of different companies

(iv) *Smart grid data*: one step ahead of a typical power grid is the smart grid that also generates information at an enormous scale [6]

(v) *IoT*: the internetworking of devices, sensors, and applications works on the principle of information exchange among the devices and hence is a leading contributor towards big data [7]. Issues in the IoT-based smart grid are that it uses internet-based protocols and infrastructure of public communication which are more exposed to security threats [8].

(vi) *Search engine data*: search engines like Google, Yahoo, and Bing also create a huge amount of data

1.2. Why Big Data? Big data, an emerging and one of the most important technologies in the world of internet, IoT, mobile networks, wireless sensor networks (WSN), smart grid systems, medical and health monitoring systems, etc. Big data have several benefits:

(i) The limitation of fossil fuels and natural resources has raised the demand for efficient energy generation, distribution, and monitoring systems. In response to requirements, the smart grid is a

Table 1: Comparison with existing related surveys.

Ref	Research area	Year	Remarks	Issues identified	Possible solutions
[9]	Scope and privacy	2013	This paper presents scope and privacy concerns in big data.	Privacy and security issues	✗
[4]	Applications and challenges	2014	This comprehensive survey covers communication and business applications as well as challenges and technologies.	Volume of data	✓
[10]	Clustering algorithms	2014	This survey provides a number of algorithms related to clustering. Comparison of existing clustering algorithms has been included.	Limitations of data clustering algorithms	✗
[11]	Platforms for big data analytics	2015	This study offers a survey on available platforms for big data analytics. Pros and cons of each platform are explored.	Drawbacks of different data processing platforms	✗
[12]	Mining algorithm	2015	It presents brief introduction of data analytics and the mining algorithm to extract the useful information from big data.	Issues related to platform, framework, security, privacy, and data mining perspective have been highlighted.	✗
[13]	Parallel processing	2016	This survey paper presents an overview of parallel processing and highlighted the processing efficiency of different cases.	Novel data, processing model, energy efficiency, and large-scale machine learning	✗
[14]	Networking for big data	2017	This survey provides the introduction of networking in big data as well as networking features, challenges, and opportunities.	Big graph mining, dynamic representation, time evolution, security, privacy, and scheduling for big data related to networking perspective	✗
[15]	Modern computing paradigms	2017	New computing paradigms are discussed for big data in the IoT case and limitation of cloud computing for the IoT applications. Data base management systems based on NoSQL are investigated for different authorizers.	Storage, management, security, privacy, computation, and resource performance	✗
[16]	Big data issues in smart grids	2019	This article highlights issues related to big data analytics, technologies, and architectures in next-generation power systems.	System, data management, and analysis	✗
This article	Big data resource management in smart grids	2021	Big data in the domain of a smart grid is explored and the resource management for smart grid applications is discussed. Techniques, tools, and challenges are also elaborated.	Volume, data integration, storage and visualization from multiple sources, data backup, privacy, security, confidentiality, energy management, and quality	✓

technological advancement that is a solution to the energy crisis. The generated big data from the smart meter in terms of volume, variety, and velocity would be very much beneficial for efficient utilization of energy as well as for better energy planning [17]

(ii) Different companies of marketing agencies use big data resource management strategies in order to improve the response of their campaigns, promotions and other advertising mediums [18], and information of the social network like Facebook [5]

(iii) Hospitals are providing quick and better services using the information regarding the previous medical history of patients [19] and predicting the future health conditions using big data analysis in the domain of health care [20, 21]

1.3. 5 Vs of Big Data. Volume: the first "V" is the large volume of the data clusters of big data. The data is so large that it cannot be analyzed by any conventional methods. Five versions of big data are shown in Figure 2. The data is increasing at a very fast rate, and according to experts, 78% of the current data on social websites has been produced in 5 years since 2011 making it the largest data generated to date. Other examples include the following: Facebook produces 500 terabytes of data on a daily routine, according to a report of the IDC USA; the increase of data will be 400 times by now in 2021. The explosion of data which has been collected in e-commerce is 10 times more in quantity of an individual's data transaction [22].

Variety: variety targets the type of data that we have. It may be a structured, semistructured, or unstructured data set. However, the majority of big data is unstructured that is randomly generated by multiple sources. Big data is not

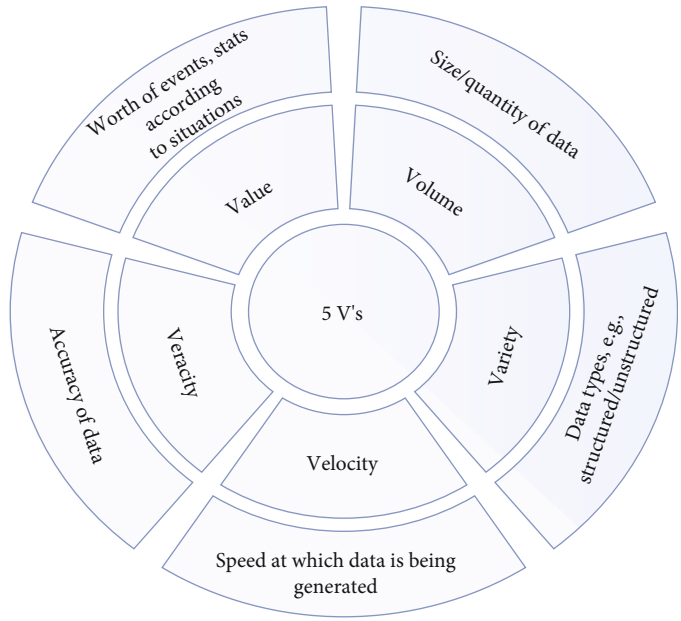

Figure 2: 5 Vs of big data.

just bits and pieces; it is much more than that. Big data includes audio, video, 3D data, and unstructured text, including log files and social media. The traditional data includes lower volumes of consistent and structured data.

Velocity: the third V is velocity, which deals with the pace of data that is being generated by different sources like machines, mobile networks, business processes, and human interaction with things like social media sites and internet banking. The information flow is continuous and massive as well. Handling this rate at which data is being generated provides a strong basis for valuable decisions. It leads toward rapid interpretation and strategic competitive advantages to help businesses and researchers from this real-time data.

(i) Clickstreams and ads capture a large amount of data, e.g., millions of events per second

(ii) It takes a fraction of seconds to reflect market changes for high frequency stock trading algorithms

(iii) Online gaming produces huge amount of data from millions of concurrent users producing multiple inputs per second

Veracity: uncertainty or inaccuracy of data can be dealt under the 4th V of big data that is termed as veracity [23]. Data veracity refers to the abnormality and noise in the data. It also deals with whether the stored data is meaningful to the analysis or not. As compared to velocity and volume, it is the observation of the community that veracity in data analysis is the supreme challenge.

Value: value is also very important when business models are considered. Data should be analyzed in accordance to the value of data to get the best result out of the analysis. Value is critical for business initial phases, because it is the matter of investing money and reducing the risks. Still, many companies are not using this application of big data in effective way. Better use of this application will not only be effective for revenue generation but also help to avoid fraud.

1.4. 5 Vs and Smart Grid. Smart grid incorporates conventional power systems with a bidirectional infrastructure that integrates electricity and information flow. The smart grid is a complex interconnected system that generates a diversified variety of data with huge volume, high velocity, and veracity. These 5 Vs are worthy of importance when we discuss automated electric grid systems [24].

1.5. Major Contribution. Our focus is to provide a comprehensive survey on big data for smart grid applications. The contributions of this work are summarized as follows:

(i) General overview of big data and smart grid systems

(ii) Big data-generating sources in the smart grid

(iii) Importance of resource management

(iv) Tools and techniques for the analysis of big data

(v) Research challenges

Most of the existing literature on big data and smart grid mainly focus on its applications, issues, tools, technologies, and techniques separately, but big data resource management in the context of the smart gird has not been explored so far. References [16, 25] targeted the issues related to SG, but the management perspective is missing in the literature. Different domains of big data are being targeted in various reviews, but a comprehensive survey is not present in the existing literature that covers a holistic picture of big data. This paper has been written in such a way that it clears the complete picture of big data for the beginner in this field. Unlike other available literature on the big data smart grid,

Table 2: Comparison with existing work.

Ref	Techniques/tools	Smart grid	Issues in smart grid	Resource management discussed	Challenges and opportunities discussed
[1]	×	×	×	×	✓
[3]	×	✓	×	✓	×
[8]	×	✓	✓	×	✓
[17]	×	✓	×	×	✓
[22]	✓	×	×	×	✓
[26]	×	✓	✓	×	✓
[27]	×	×	×	✓	✓
[28]	×	✓	×	✓	×
[29]	×	×	×	✓	✓
[30]	×	×	×	×	✓
[31]	✓	✓	×	×	✓
[32]	×	×	×	✓	✓
[33]	✓	✓	×	×	✓
[34]	✓	✓	✓	×	✓
[16]	✓	✓	✓	×	✓
[25]	✓	✓	✓	×	✓
This article	✓	✓	✓	✓	✓

it not only covers the main topic but gives a clear holistic picture of the importance of big data and resource management in smart grid networks. Table 2 represents a comparative analysis of this article with existing literature available. Uniqueness or real contribution of this article is clearly judged by Table 2.

1.6. Article Structure. The paper's organization is shown in Figure 3 and is described in the later sections. In Section 2, we discuss the smart grid systems and sources generating big data like sensor and information flow of different applications. The motivation for deploying resource management is presented in Section 2.1. In Sections 3 and 4, we have discussed different techniques and tools like Hadoop/MapReduce.

2. Smart Grid Systems

Smart grid power systems are new innovative power systems which will not only provide more electricity to meet the increasing demand but also improve reliability, efficiency, and quality. This system will allow other individuals to add their energy in the national grid which includes many energy sources like renewable energy resources (solar, biogas, wind, etc.) [35]. Traditional power distribution systems transport energy to the consumer side from a central power plant using transmission lines [24]. Major stakeholders of smart grid systems are the distribution, transmission, consumption, and communication networks. The communication network is actually the main portion that converts the conventional grid to a smarter one. There is a two-way communication between the distributor and the consumer in smart grid networks. This information exchange and continuous monitoring of energy enables efficient utilization of power in emerging smart grid networks [36, 37].

Smart grid systems enable the grid to observe and control the power parameters accurately. This system also offers to make decisions on time as well as allows us to integrate renewable systems. In this advancement of the grid system, communication technology plays a pivotal role as depicted in Figure 4. It establishes a strong link between the distributor and the consumer to make the network more efficient.

Reliable and efficient distribution of electricity is a basic requirement with essential energy production units. The power grid infrastructure was deployed in early ages and is now reaching full life. For more competition, there is a need for strong political and regulatory push, lower energy prices and more energy efficiency, and greater use of renewable energy like biomass, water, solar, and wind to keep the environment clean. The load demand has remained the same or has slightly increased in the previous years in industrial countries. Some of the developing countries show a rigorous increase in load demand. But now, load demand is increasing exponentially due to more industries and increasing population [38]. On the other side, aging equipment may lead to shortfall of electricity during peak hours. In different parts of the world, regulators are advising utilities to find the cost-effective solution for transmission and distribution of electricity. That is why new techniques like the smart grid (based on modern communication) are emerging to operate power systems which guarantee a secure, sustainable, and competitive energy supply. The important goals of an advanced electrical grid are to ensure an environment-friendly, transparent, and sustainable system. Utilization of renewable energy resources are worthy of importance in order to meet the above-mentioned goals of smart grid systems.

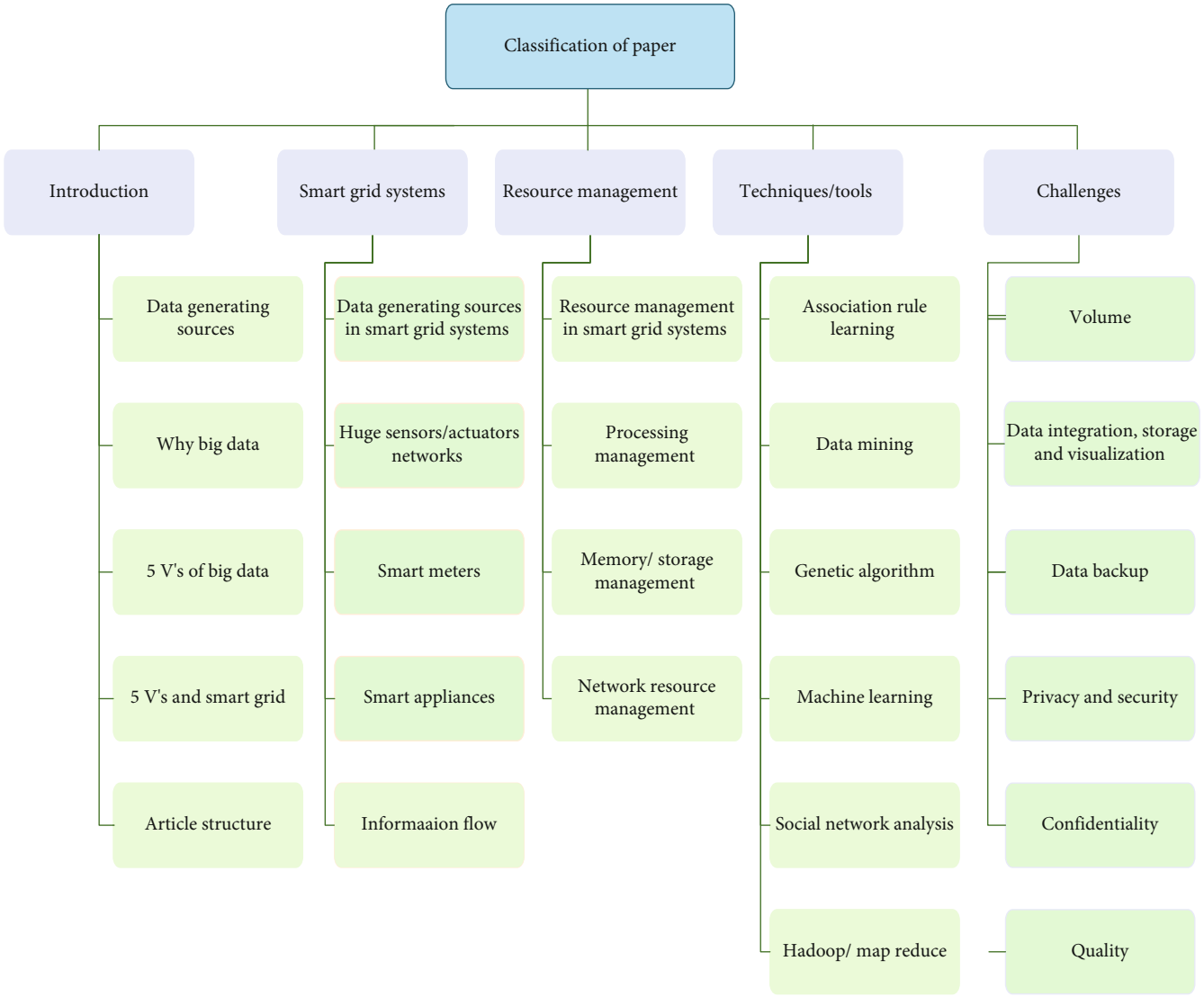

Figure 3: Organization of paper.

2.1. Types of Data-Generating Sources in Smart Grid Systems

2.1.1. Huge Sensor/Actuator Network. A smart monitoring system is actually a strong source that generates huge data sets. It is impossible to implement a smart monitoring infrastructure without using low-cost but intelligent devices. A new scheme of sensors termed as smart sensors has recently been introduced that fulfils the criteria discussed above, i.e., low cost, ultralow power, and more intelligence [39, 40]. The importance of smart sensors has been discussed in detail in [41], and a new type of sensor termed as "stick on" was investigated. These sensors do not even need physical contact with the utility asset for some applications. They have the capability to monitor different parameters of interest only by getting close to utility assets. Fang et al. have also discussed the self-powering smart sensors and challenges related to that domain.

2.1.2. Smart Meters. A smart grid comprises smart meters that play a very important role. A grid, by definition, is an electric system that includes electricity generation, transmission, distribution, and consumption. A traditional power grid system comprises a typical setup that supplies electricity to users and consumers by carrying that power from a few central generators [42]. One main advantage of the Smart Meter System is their simple operation of the overall process even if they are varied in technology and design. These intelligent meters gather information from end consumers every 15 minutes or once a day and transmit that valuable information to the data collector through the Local Area Network (LAN). Arif et al. [43] developed a smart meter based on GSM and ZigBee. These meters are capable enough to update the information of the service provider about the energy measurements. The service provider can use this information to notify their consumers via Short Message Service (SMS) or using the internet. A hardware architecture is presented in [44] which discussed the adapted communication protocol and monitored the energy using web-based application. Managing the energy in smart grid systems using a mobile application is investigated in [45] to improve

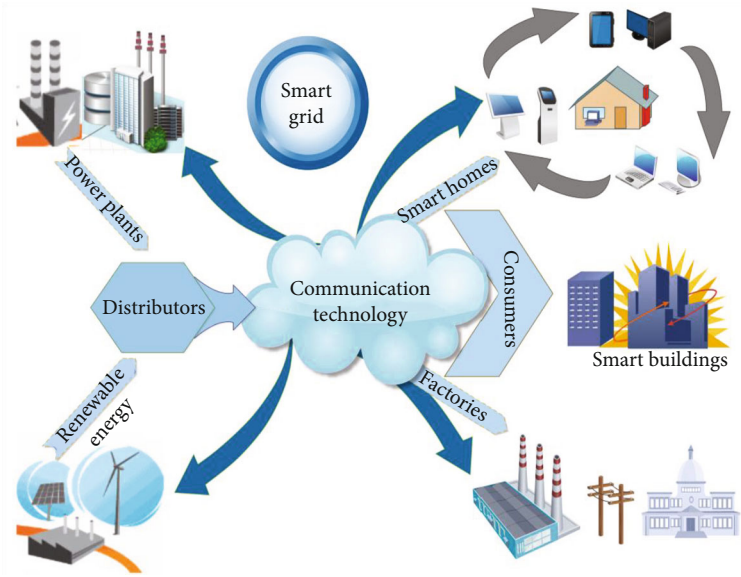

Figure 4: Smart grid look.

the availability and data exchange. Policies and security concerns vary from country to country which has been discussed in [46], and the smart grid development in different countries has been compared.

2.1.3. Smart Appliances. The smart grid system has done a lot in aligning the electricity demand and supply during peak hours by promoting small-scale renewable energy generation [47]. Talking about smart houses, the key element is the smart cards that are responsible for communication between the smart meter and appliances. These smart cards act as a communication link for the transference of information. The town server holds the connection of the number of such smart houses and is responsible for controlling the power provided by the service provider and the power generated by regenerative sources. A town server network has been discussed in [48] which manages the communication and the whole power consumption between the systems. A smart house architecture is presented in [49] which is proposed for a demand-responsive energy management system based on Information and Communication Technology (ICT).

2.1.4. Information Flow. In smart grid systems, communication or information plays a crucial role in making decisions. Normally, decisions are based on the collected information. In power systems, most of the time, information plays a very critical role. The grid is becoming smarter with the passage of time by the use of modern technologies which facilitate bidirectional information sharing between customers and the utility [50]. The smart grid consists of sensors, actuators, smart meters, control units, computers, etc. The information of all these sources flow from one point to another. Effective management systems are necessary to manage the information of these heterogeneous complex and bulk data networks. In [51], Suciu et al. examined the cyberphysical system (CPS) from an information flow perspective. A method is presented to analyze the leakage of information by using the advice tape concept in the field of algorithms.

3. Management Perspective

3.1. Resource Management. Resource management in every field is very important to optimize many parameters. In Figure 5, the variety of resource management processes are shown. Resource optimization is a supreme parameter to minimize the cost and improve efficiency. Normally, the resource is in the form of a spectrum which is sparse due to the exponential increase in the user devices. Resource management is an effective and efficient allocation of resources in any platform. User devices are increasing exponentially with the times and generating a lot of distributed data in various forms. Data handling is a big challenge for researchers. Without efficient management in big data applications, it is very hard to tackle such huge data. A huge research space is available for exploring resource management in big data.

In big data, resource management in the sense of memory and complexity is rarely explored. Different applications create 2.5 quintillion bytes of data every day [52]. The amazing thing is that 90 percent of the data in the world has been created in the last two years. This data includes all applications like sensors used to gather information about climate, posts to social media sites, cell phone GPS signals, digital pictures and videos, purchase transaction records, etc. Different aspects in big data like resource management, processing, analytics for social media, database technique packing algorithms, security, and privacy concerns are covered in [53]. Speed of information technology growth is increased from Moor's law at the beginning of the 21st century. Excessive data is creating more challenges in data science. On the other side, data science is extremely important to produce productivity in businesses which will create a lot of opportunities. Reference [4] discussed a closed-up

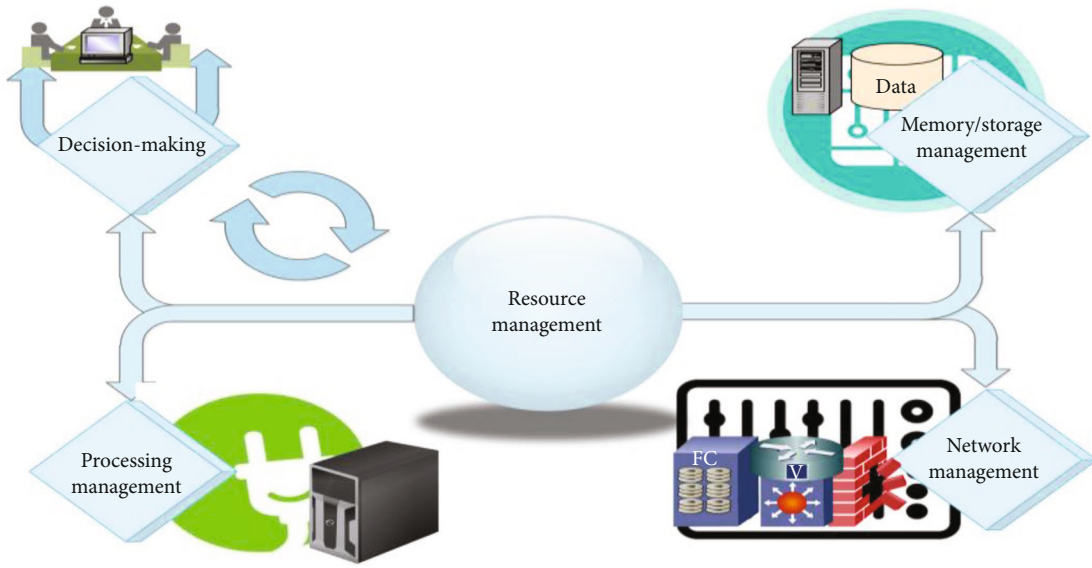

Figure 5: Resource management.

view about big data including opportunities, applications, and challenges. Chen et al. have also discussed techniques and technologies used to deal with big data problems. In [1], Chen et al. explained in detail about the background of big data, related technologies, data storage, applications, issues, and practical applications. In [54], different aspects of resource management for big data platforms have been examined. Pop et al. also discussed the importance of resource management for smart cities. The smart grid is the part of smart cities which will enable the use of energy in more efficient ways. Resource management for big data applications is an open issue for current development in the era of the smart world. Predictive resource planning and allocation discussed in [55] is energy saving and will ultimately save on costs. Won et al. in [56] investigated the advance resource management for multiple tenants using access control to share the computing resources. In this environment, multiple tenants having different demands can share computing resources like data, storage, network, memory, and Central Processing Unit (CPU). Researchers claimed that multitenancy reduces cost and offers highly effective saving computing resources to acquire a similar environment for data management and processing. The novel approach is used to support the multitenancy features for Hadoop. It is understood that Hadoop is a large-scale distributed system which is commonly used for the processing of data. Resource management in big data is rarely covered in the community although popular literatures and academia have many examples on initiatives of big data. Senior managers are hesitant to commit resources on data sciences on a sustainable basis. Reference [57] covered the theme of improving organizational resource management and created a concentration to attain a positive capability with initiatives of big data. The relationship between dynamic capabilities and big data is of great significance because processes of data need to be developed step by step as organizations want new insights from big data. The smart grid is a growing technology in the power system which also needs data or information management to efficiently utilize the resources which is investigated in [3]. They discussed how to manage different types of front-end intelligent devices like smart meters and power assets.

Resource management in smart grid systems: the exponential increase in the data has prompted many challenges to develop systems to manage the resources. It is required to manage resources to analyze the huge amount of data efficiently. It is impossible to manage the big data in traditional ways. There is a need to manage the resource like processing, memory, and network resources so that we could be able to process the complex data systems in comfortable ways in emerging smart grid networks.

3.2. Processing Management. Processing in big data plays a pivotal role to analyze the data to extract the required results. Big data-processing techniques process data sets of terabytes or even more than that. Processing is further divided into distributed and parallel processing using traditional application frameworks like OpenMP and MPI which are still playing an important role. Newly investigated big data processing and cloud computing frameworks like Spark, Hadoop, and Storm are becoming popular. But in the parallel application framework, it requires a physical cluster to run the system efficiently. A resource-sharing approach using a cluster as a service for a private cloud has been discussed in [58]. The ClaaS model is proposed to make the implementation simple. Authors claimed that it is an effective model for sharing a cluster between several frameworks. Parallel processing systems like batch, graph, stream, and machine learning techniques have been discussed in [13] in which optimization and extensions for the MapReduce platform are also discussed. There are many platforms that are developed for processing purpose, but

[59] developed a pipeline structure for the heterogeneous execution environment to integrate data jobs. To integrate data jobs in a heterogeneous execution environment, developers need to write a long glue code to get data sets into and out of those jobs. For the data pipelining and integration support, some frameworks are also proposed such as Crunch, Cascading, and Pig, but these frameworks are built on top of a single data-processing execution environment. Resource management in big data regarding business purpose is an important aspect that either these initiatives are helping managers to grow their business and make it more profitable or not. It is invested in [57] which classifies the organizational resource management in three aspects. First is to establish a business process archetype and second to create a dynamic capability and identify the drawbacks of the resource-based theory. Lessons are learned, and the implications for business research and practice are sorted out. An applied example to apply the data techniques to smart cities had been investigated in [60], and an IoT-based architecture was proposed. Some services implemented in the smart campus of Murcia University and some services are focused on tram service scenarios where thousands of transaction data are considered.

3.3. Memory/Storage Management. Growing memory capacity has accelerated the development in memory of big data processing and management [27]. Real-time data analytics require intelligent memory or storage systems which have the least latency to read or write the data. Initially, the need of this type of performance was encountered by well-known global companies like Google, Amazon, and Facebook, but now, it is becoming an obstacle for other organizations which are looking for a meaningful real-time service like social gaming, advertising, and real-time bidding. To meet the requirements in real time for analysis of large data sets in milliseconds requires RAM memory. Bandwidth, capacity, and memory storage have been doubling after every three years while its price is dropping by a factor of ten every 5 years. Noteworthy advances have been observed in nonvolatile memory (NVM) e.g., SSD. Hardware technology advancement in recent times has generated interest in hosting the whole database as well as overturned many earlier works [61] in memory to provide real-time analytics [62, 63] and faster access. Comprehensive memory management and some key factors to achieve efficient memory data management and processing are investigated in [27]. There are privacy and security implications [64] of pervasive memory augmentation which effect what and how humans radically change the scale and nature of external cues. The presence of ubiquitous displays in personal devices and environment provides new opportunities for showing memory cues to trigger recall.

3.4. Network Resource Management. Recent findings show that human behavior is highly predictable [65]. Improving the performance in wireless systems by exploiting the predicted information draws an attention which is known as anticipatory, context aware, and predictive resource allocation in the literature [66, 67]. Context awareness is not a new concept in the computing science. Context is any information that can be used to characterize the situation of an entity. Entity can be anything like a place, object, or person that is contemplate relevant to the interaction between an application and a user. Energy-efficient predictive resource allocation and planning is presented in [55] based on predictive analysis and results that show that the proposed policy can drastically reduce the energy consumed by the BSs. Won et al. [56] introduced an advanced resource management (e.g., network, memory, storage, and data) with an access control in a multitenant environment. Multitenancy facilitates management of multiple users who use similar systems. Using this concept, the system is able to permit multiple users to maintain and develop their own environment; otherwise, an application is required to provide their products by manual anthropology to address the requirements of each company. The manual approach resulted in an excessive maintenance cost because it desires the management of each company separately. Hadoop Apache was used as a base platform for providing features of multitenancy.

4. Techniques

There are different techniques defined by the researcher to analyze big data. Researchers are continuously working on the development of new techniques as well as focusing on the improvement of existing ones. Some of the most common and regularly used techniques for the analysis of big data are

(1) Association rule learning (ARL)

(2) Data mining

(3) Genetic algorithm

(4) Social network analysis

(5) Classification tree analysis

4.1. Association Rule Learning. With the evolution of data generation, new methods of data analysis are needed to carry out in-depth analysis of the clusters. One such rule is ARL. It is a method related to rule-based machine learning and used to discover interesting relations between variables in large databases. With the increase in the quantity of big data, ARL is being implemented across the globe in a number of fields to study relationships between variables to sort data according to desired variables. Its applications range from consumer markets to modern-day communication.

With the increase in the internet users, the data generation across the different levels of the World Wide Web has sky rocketed, but the internet still works on criteria and the protocols of the past. The internet cannot keep up with the recent increase in the data generation and storing. The modern-day data supports a large number of elements which can be used to create semantics to understand the trends of data. Reference [68] introduces the design of a human mind-based semantics to improve

internet decision-making associated with an analysis technique for accurate reasoning about the internet and to compare the current algorithms. As the data volume has increased, the complexity of the mobile network, furthermore, the user-based data generation and complex interrelations, has also grown. Reference [69] uses ARL to produce a deep network analyzer (DNA) for anomaly detection in order to make further improvement in the network and make an accurate gain prediction to address a wide range of problems faced by internet service providers (ISP). With the increased span of the internet and complexity of the networks, the dark network has also increased. Since its dawn, the dark net has been the corner stone of illegal activity across the global networks resulting in huge loss of value from patents of companies to breaching of the defense networks of countries. This new dawn of internet volume explosion has made it easier for cyberattacks on critical infrastructure. ARL can be used for the analysis of the data clusters of the dark net, predicting relationships between a number of factors such as malusers and beneficiaries of such activities.

Talking about the smart grid, it is understood that it uses a vast network of smart sensors. These sensors are responsible of generating a huge volume of data that must be categorized in a mathematical or scientific way to make this advanced network more efficient. References [3, 28] explored numerous applications as well as the techniques used for managing the big data of smart grids. Reference [70] uses ARL-based learning to draw a pattern between malware and cyberattack activities and draw a rule-based diagram to point infected machines and routes as well as probing the dark net. Similarly, cloud computing has taken the main stage after the recent internet revolution [71]. The increased data and information flow through a wireless medium between intermediate devices in smart grids has also increased the concern of its privacy and security detail. Reference [72] introduces an ARL-based analysis technique for sifting through mined data in order to prepare routines for improving privacy and security along with guaranteed result from the data mining operations.

The recent increase in the large amount of data generation has been problematic for a number of reasons. It takes a toll on services to store and make it accessible; furthermore, to sift through the data, to find relationships, and to make it efficient for the user are a challenge. The data sorting is very important from a number of views like market investment to national security concerns. ARL helps to narrow down the study criteria to a limited pool of variables making it easy to analyze large smart grid data clusters from the point of view of the concerned. Reference [73] focuses on the discovery of these routines in the big data stacks and a post analysis of these rules to arrange them in a better fashion for a more efficient process. Similarly, [74] addresses the problem by an algorithm to highlight the mined rules by assigning them weights in binary digits. Reference [75] uses a number of ARL-associated trees to improve the efficiency of the mined rules in order to keep up with the rapidly increasing data clusters. References [76, 77] proposed a data mining algorithm based upon MapReduce to sift through the data and produce a more efficient rule-based tree for decision making. Reference [78] is the implementation of the ARL mining on the data in order to improve the failure rate of products and predict the market variables concerning it by studying the trends across the social networks. Also, with the evolution of the internet, it also offers a number of concerned parties a chance to evaluate the customer psyche. Reference [79] is used to mine according to ARL the activities of the users related to their user's concerning factors like when, where, and what for in order to get a better understanding of the response of the customer community.

Increase in the development of a smarter network in service sectors and usage of internet services in many areas of application have further increased the ease of access and maintenance for a number of complex networks. One such example is the new power networks. Power networks are also very important and help in the distribution of the electricity to domestic and industrial use making them an essential part of modern-day life. A failure can lead to disaster. Reference [80] uses a number of algorithms to rule mine the data from power station differential equations. The mined data projects the values regarding the system helping in their maintenance as well as further development of the networks.

4.2. Data Mining. Data mining software permits users to make the analysis of data from different dimensions, summarize it, and categorize the relationships identified. The concept of data mining is gaining popularity in the modern era of information and technology. In the information economy, data is being downloaded, uploaded, and extrapolated. So data mining is the incorporation of mathematical methods and algorithms including classification to extract patterns regarding desired data.

A dynamic power grid not only focuses on energy storage but is also concerned about the value of information [81]. According to IEA (International Energy Agency), out of our total final consumption of energy, 32% is consumed by residential and commercial buildings [82]. So it requires more intelligent strategies for processing and analyzing the big data related to the smart grid and residential and commercial buildings. The data mining technique can be used for categorizing smart data into useful information.

Data mining is also used for a number of purposes in the daily civil services ranging from engineering to finance. In Public Structural Development (PSB), data is collected from various aspects and sensors, providing information such as the structural integrity of various structures forming recognition-based patterns on the statistics and is known as structural health monitoring (SHM). The data does not provide the parameters like acceleration and displacement velocity but actually provides the change in the parameters of the structure; a number of mathematical models compute and provide the output according to [83]. Signal sorting of radar communication is an important factor in modern warfare electronics. Modern radars are highly advanced and provide a number of challenges like using multiple emitters eliminating conventional algorithms and producing a ton

of data. A number of developed sorting methods are discussed in [84] in order to sift through the data.

In finance, data mining plays an important role in operation planning. In large-scale operations, business, routines, and processes, it is very important to decompose itself into smaller multiple units for the multiple objective optimizations of the entity also known as role-based access control (RBAC). It uses data mining techniques to discover rules from user permissions from access lists. Reference [85] uses the said technique to further optimize the entity in discussion using RBAC and edge concentration. On the other hand, customer database and trend understanding is very important for service providers. It involves a large database of customers and their daily activates, practices, and behavioral traits. Reference [86] involves the utilization of the multivariate data collected from ends like phones and services. The process involves receiving data and updates from a large number of devices and nodes, sorting and production of desired characteristics and trends. As technologies continue to improve in use and experience, similarly, Facebook-like applications have attracted large user bases linking the virtual space with the real world. In [87], the authors used data mined from the geosocial networks to understand traits and response of the users to provide better statistical analysis for concerned parties providing peoples opinion regarding decisions.

Tax evasion is a common felony practiced at a large scale. Due to the high data volume, it is impossible to detect such a large amount of tax theft, so the data must be sorted and analyzed and the results extrapolated as [88] used the color network-based model (CNBM) for the construction of a pattern tree providing a link between tax evasion techniques and behavior trends. Similarly, electric power is a basic necessity and very important to the modern-day life sustenance. A large number of energy frauds are committed around the world. Reference [89] introduces a technique involving data mining through the advanced metering infrastructure (AMI) plotting the data to provide a number of plausible suspects without including field inspections by constructing a cluster using homogeneous data and constructing prototypes.

4.3. Classification Tree Analysis. Classification tree analyses are used to generate the prediction regarding the membership of cases or objects in multiple classes using one or more predictor variables through the help of categorical measurements. Classification tree analysis is one of the main techniques used in data mining.

A decision tree helps the routine in classifying the best option out of the members and their classes presented in the tree which helps in sifting through a large amount of data. For having accurate classes and objects, the training data provided to the tree must be closely related to the analysis data for a comprehensive decision. Reference [89] involves multiple methods and approaches to improve the accuracy of the sample or the training data using multiple attributes of the data. In big data, sometimes, it is needed to compute numerical and mixed data which has to be made discrete, as many of the convention methods and algorithms are not suitable for big data computation. Reference [90] involves the development of an algorithm to perform discretization and to be further structured into fragments to contain one data each in each object of the classification tree.

With the emergence of high-speed internet to the masses, cloud services are used across the globe from domestic to industrial use. The number of cloud users has reached a peak above any other service comparable like email and social networks. From storage of personal items to office use, the cloud has replaced a number of services but is also giving rise to such things as security and privacy which increases the data multifold. As the user database is increasing, so is the need of betterment in the current cloud structure and implementation. Reference [91] proposes a number of implementations in order to answer the challenges faced by cloud computation.

Data stream is a constant influx of infinite data continuously in a nonstationary manner. In the stream, such algorithms are placed that are learning so that they can overcome the limitations of time and hardware. Reference [92] involves an algorithm to deal with the issue of

(1) What and how data to present

(2) The display of the recent data by manipulating the nodes of the classification tree; [93] uses MapReduce in collaboration with tree analysis to sift through the data

4.4. Genetic Algorithm (GA). GA are an adaptive and stage-dependent search algorithm. It is based on the evolutionary ideas of genetics and natural selection process. GA is an intelligent optimization algorithm which uses sifting and sorting of a random search. Genetic algorithms (GAs) are designed in such a way so that they can simulate processes in natural systems necessary for evolution. It is obtained solving both limited and nonlimited optimization sets that grow taking the best characteristics like biological evolution, each time replacing the previous with the next.

GA is used in a number of applications along with big data tools like Hadoop. Hadoop is a cluster which consists of thousands of servers and tens of thousands of CPUs which queue up a number of jobs which require multimode scheduling using software. It is needed to improve their efficiency which has been done in depth in [94, 95] keeping in mind real-time system status.

With the increase in the amount of data-generating sources, a better system of mining the data from multiple sources is required. One such source is the modern communication system, which is complex due to the user base. Reference [96] performs an analysis on the problem through classification and regression analysis by studying the big data clusters to detect anomalies. With introduction of cloud computing, it is needed to efficiently mine data for which the big data cloud is used with the help of a number of tools like Hadoop and MapReduce. So it is important to optimize the work of the routine. Reference [97] does in-depth analysis of the process of optimization. With the introduction of cloud computing and software as a service, a number

of web services continue to increase, resulting in increased business value and routines. Reference [98] is based on an analysis on improving the scheduling criteria using MapReduce.

As optimization techniques are used in multiple fields, for instance, big data is generated in the medical field with the help of the high-res scanning technology available. In [99], genetic analysis is used to propose a distinct classification analysis for a number of variables than build a table or tree which could help to develop values for a number of conditions like diseases and infections.

4.5. Social Network Analysis (SNA). SNA actually measures the relationships and flows between groups, people, organizations, URLs, computers, and other connected information/knowledge entities.

Nowadays, the internet data has increased multifold and is changing at an alarming rate. The analyses of mined data with algorithms and traditional methods are costly for the large amount of data. So [100] performs the big data tools to map a tree of the traffic nodes on the internet regarding social websites. The internet with its complexity is a collection of large databases with multiple modes. This ranges from text to pictures, videos, etc. However, there is no sorting process for the multitudes of data mined from the internet. So [101] does depth analysis to provide a better approach to the sorting of the data types in order to make the process of data mining more accurate.

Microblogging sites like Facebook, LinkedIn, and Twitter are very important to modern-day communication and play an important role in the daily lives of a large customer database. References [102, 103] use a number of analysis techniques to separate conversation and posts regarding certain data types and construct a tree about relationships interacting with the desired data which in turn can be used for a number of purposes and by many concerned parties for an advertisement-like process. The mined data is used to construct a tree based on the interest of users and recommending them items using a recommender system based on the user activities. On the other hand, with the increase of the microblogging and social websites, social events have been arranged and invitations are sent out via the internet. People attend and get awareness to it, and it further includes the coverage of the event by the people as well. References [104, 105] make the use of algorithms to plot a tree from the information on these sites to detect events and related social activities. With the growth of the internet, social websites have also increased the number of social event coverages on the internet. These events produce multimode data such as videos and images that can be used by concerned parties. Reference [106] proposes an algorithm for multimode tracking of the event for getting a varied form of data.

With the large number of data appearing on the internet, it is also needed to compute and sort the mined data in order to further maximize the use. Reference [107] uses in-depth analysis to compute and arrange the data to produce and maximize the results for use in a number of fields like business and media. A lot of reviews and feedback are found on the social websites. Innovation diffusion deals with the response a new product receives in the customer user base like [108] uses a number of algorithms to do in-depth analysis of the data mined through social websites and networks for a concerned product or party.

5. Tools of Big Data

The survey covers two universal tools used for the analysis of big data generated by the smart grid, social media, IoT, stock exchange, etc.

(1) Hadoop

(2) MapReduce

With all the Vs of big data, conventional means are not enough to tackle the problems and challenges presented by big data and its handling. So for a better analysis method, a number of tools were developed on the techniques mentioned above to work on big data handling. The survey will include Hadoop and its algorithm of MapReduce.

5.1. Hadoop and Its Importance. Hadoop was developed as an Apache top level project. It was an open-source implementation of frameworks which provided qualities like reliable, scalable, and distributed computing and data storage. It is a flexible and highly available architecture [109]. The following were goals of the Hadoop Apache Project.

(i) Facilitate the processing and storage of large and rapidly growing data sets, e.g., unstructured and structured data

(ii) Simple programming models

(iii) High availability and scalability

(iv) Use commodity hardware with little redundancy

(v) Fault can be tolerated

(vi) Move computation rather than data

In 2003, Hadoop was bought and implemented by Google, and in 2004, the Hadoop MapReduce Algorithm was developed. Hadoop has the following three important features.

(1) Hadoop is based around analyzing big volume data in large amounts that are further analyzed by breaking it according to one of the analysis techniques like ARL and CTA. One application is the analysis of large data clusters provided by RFID sensors in a large number of applications such as the Geographic Information System and earth observation with the help of ARL and genetic analysis to filter out the required data from the cluster

(2) Hadoop is also being used in the analysis of large number of servers ranging from cloud servers to app-related services. These servers get a constant

influx of data from a large number of devices like smart phones having a large array of sensors providing a continuous stream of data. In order to sift through the data, ARL is used with Hadoop to develop relationships and links but with the data in the clusters. Furthermore, the analysis of the mobile networks also yields a large amount of mined data which cannot be handled via conventional means. So Hadoop is used to sift through garbage data and get the required relationships

(3) With such volumes of data, a large amount of text is also generated. CTA is used with Hadoop to analyze the relationships in the text to arrange them

5.2. MapReduce. It is important to differentiate between MapReduce and an implementation of MapReduce, in order to fully understand the capabilities of Hadoop MapReduce. It is an implementation of the algorithm maintained and developed by the Hadoop Apache Project. If you take MapReduce as an engine, then it is an efficient engine which takes data as fuel converting it into energy in a quick and efficient manner.

5.3. Advantage. The major advantage is that data processing over multiple computing nodes is made easier using MapReduce.

5.4. Working. It can be implemented in three stages, namely, map stage, shuffle stage, and reduce stage.

(i) *Map stage*: the mapper's job is to process the input data and create small chunks of data, and that is stored in the Hadoop file system (HDFS) in the form of a file or directory. Then, line by line, the input file is passed to the mapper function

(ii) *Reduce stage*: shuffling and reducing both combine to form the reduce stage. The data that came from the mapper is then processed by the reducer. It gives a new version of the output after processing, which will be stored in the HDFS

MapReduce is based around the analysis of the large amount data inputs to make it very applicable on the modern data and communication network of smart power grids. With the complexity of the modern network, it has to be analyzed for anomalies and dark net trenches which target the data. Furthermore, it is used to analyze the network to maintain and upkeep the internet speed. It also analyzes the cloud network relationships with the internet tracking the big data associations with the cloud network. MapReduce is also used in the database analysis to analyze large data clusters of XML, structured query language (SQL) based on CTA or genetic analysis. It helps a lot in financial and administrative sectors, tracing and locating relationships between data. It had already helped a lot in the federal tax audit for tracing culprits and identity theft. It is also used in the power and domestic services from computing power network algorithms of a city to the traffic patterns in certain hours of the city workload.

6. Challenges and Open Issues of Big Data

With the constant evolution of the internet and its related data-generating sources, the volume of big data is increasing at an alarming rate making it necessary for the developers and researchers to keep coming up with new means and analyses to handle big data. It also involves the development of new technologies to look after the hardware prospect of big data computation. So out of the multifold challenges, the following were surveyed.

(i) Volume

(ii) Data integration, storage, and visualization from multiple sources

(iii) Data backup

(iv) Privacy and security

(v) Confidentiality

(vi) Energy management

(vii) Quality

6.1. Volume. The volume of the big data in a smart grid is increasing daily. With the increase in the complexity of the data-generating sources, it is impossible for the conventional data manipulation and sources to deal with big data. By entering smart devices into the mix, the big data clusters are also increasing with higher velocity than the previous 5 years of big data. So using ARL and CTA in collaboration with MapReduce, new developments are being done in new protocols [110] to handle the flood of data across the cloud servers. Reference [111] involves the development of new internet protocols for 5G based on the data accumulated from the study of 3G and 4G internet. With the emerging trends, there is a need of a proper big data computing architecture which is proposed in [112] for smart grid analysis. This communication architecture involves resources of data generation, storage, transmission, and analysis of data. Similarly, [29, 111] established development in the analysis of large RFID and sensor array networks.

So, volume will always remain one of the big challenges in big data as any restriction or limitation on increasing the size of the data cannot be made. Proper data compression methods and continuous research and improvement in big data-handling tools and techniques are the only way to tackle this regularly increasing flood of volume.

6.2. Data Integration, Storage, and Visualization from Multiple Sources. Conventional data analysis mostly deals with data generated from a single point. For the case of power grids, data is being generated by distributed grid stations in different areas. It is difficult to store, process, correlate, and visualize data from multiple sources at the same time. HDFS is no doubt a reasonable storage file system, but it needs to be tailored when the data is collected in different representations and formats [25].

6.3. Data Backup. Maintaining the backup of collected data is important, but it is very challenging to implement. There are always limited resources for storage and processing of data, and data is being generated at unprecedented scales. There is a need for specifying a life cycle of data. Backup data should be discarded from the storage after completion of the life cycle. The data life cycle management system is itself an open issue because it is very difficult to decide which data shall be discarded without defining a standard principle for removal of stored data from the memory [1].

6.4. Privacy and Security. The bulk of information flow and advancement in technology have made living easy, but this advancement in the conventional grid system has serious security issues. Ensuring privacy and securing end-to-end communication in big data are a real challenge for researchers. Considering these security threats, [17] discussed some new findings regarding privacy and security of big data. Internet-based protocols and public communication infrastructure are used in the smart grid which is the cause of arising vulnerabilities that are discussed in [8] in detail. ICN (information-centric networking) is also a strong network architecture for smart grid systems with self-security and congestion control enabled. Reference [112] applies the ICN approach on advanced metering infrastructure to tackle the vulnerabilities regarding data security. New protocols are discussed in [111] to protect large data clusters of XML and SQL from cyberattacks and nonassociated data mining, while [113] deals with the new protocol development of cloud computation based around ARL and CTA. There is a two-way communication between the supplier and the consumer in the smart grid network. Bill payments and transactional data generally include confidential information of the customers. This personal information of the consumer is under serious threat and is one of the most important areas that must be monitored and improved on regular basis.

6.5. Blockchain. Blockchain technology is considered the most famous technology based on its high-level data transparency and security. This technology helps to meet the system requirements of smart grids effectively. A blockchain comprises a series of blocks that helps to keep the records of the data in different hash functions with the timestamps. This is beneficial as the data cannot be altered or tampered with. Since the data cannot be changed, data manipulation is impossible, thus protecting the data and reducing the chances of cybercriminals attacking the data.

6.6. Energy Management. Efficient utilization of energy is among the most focused topics of discussion all over the globe since the 20th century. The increasing demand of devices and computing systems for storage, processing, and transference of big data has also increased the energy consumption. Therefore, a concrete mechanism for power consumption control and management is worthy of importance for a clean environment and economic stability.

6.7. Quality. The quality of the data mined from large data clusters is a crucial factor in a number of applications of big data analysis. With the ever-increasing data volume and variety, it is necessary to develop algorithms to highlight the relationships between large data clusters. In [114], a variety of data clusters are investigated to improve the mining efficiency of ARL- and SNA-based network algorithms and are applicable in the e-commerce industry. CTA-based decision-making data mining from RFID networks with more accuracy has been discovered in [111].

Data generated from multiple sources, real-time processing, storage, and management of bulky data sets in different modalities and representation, and real modelling are some of the important reasons that restrict giving a fix or one-time solution plan for implementation of big data analytical systems [16]. For the above-mentioned challenges, various data scientists have suggested different solutions that have been cited in the paper. Despite the evolution of data science, this huge amount of collected data is still prone to real threats like cyberattacks, information leakage, personal privacy, and security threats. This is a vast domain that requires advanced solutions and regular improvements with the evolution of big data technologies.

7. Conclusion

Based on empirical data, discussion, and literature, it can be concluded that resource management for big data applications emphasizes effective information utilization and analysis. Communities can use smart grid technology to exchange energy in order to meet demand. This paper also addresses the concept of resource management for smart grid applications. It explains what this contemporary idea is and what its features are. The management of various resources, like memory and processors, has also been explored. In a nutshell, resource management is critical in this era of limited resources. Despite the fact that prior surveys have revealed a number of research gaps, there is still a lack of discussion on big data resource management and its recent problems. Our study not only goes over the tools and techniques used in big data analysis in great depth but also covers over the most recent challenges in this field. The growing volume, as well as security and privacy concerns, is underlined. This study provides a comprehensive overview of big data while also revealing unresolved challenges for researchers in the field.

Abbreviations

RES: Renewable energy resources
SG: Smart grid
ICT: Information and communication technology
GPS: Global positioning system
CPS: Cyberphysical system
NVM: Nonvolatile memory
ARL: Association rule learning
GA: Genetic algorithm
DNA: Deep network analyzer
ISP: Internet service provider
IEA: International Energy Agency
PSB: Public structure development

SHM: Structural health monitoring
RBAC: Role-based access control
CNBM: Color network-based model
AMI: Advanced metering infrastructure
ML: Machine learning
SNA: Social network analysis
CTA: Classification tree analysis
RFID: Radio frequency identification
HDFS: Hadoop distributed file system
ICN: Information-centric networking
SQL: Structured query language.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[2] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, "Big data analytics in mobile cellular networks," *IEEE access*, vol. 4, pp. 1985–1996, 2016.

[3] J. Baek, Q. H. Vu, J. K. Liu, X. Huang, and Y. Xiang, "A secure cloud computing based framework for big data information management of smart grid," *IEEE transactions on cloud computing*, vol. 3, no. 2, pp. 233–244, 2015.

[4] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[5] S. Peng, G. Wang, and D. Xie, "Social influence analysis in social networking big data: opportunities and challenges," *IEEE Network*, vol. 31, no. 1, pp. 11–17, 2017.

[6] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 674–686, 2017.

[7] H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, "IoT-based big data storage systems in cloud computing: perspectives and challenges," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75–87, 2017.

[8] W.-L. Chin, W. Li, and H.-H. Chen, "Energy big data security threats in IoT-based smart grid communications," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 70–75, 2017.

[9] S. Sagiroglu and D. Sinanc, "Big data: a review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 42–47, San Diego, CA, USA, 2013.

[10] A. Fahad, N. Alshatri, Z. Tari et al., "A survey of clustering algorithms for big data: taxonomy and empirical analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267–279, 2014.

[11] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 8, 2015.

[12] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data*, vol. 2, no. 1, p. 21, 2015.

[13] Y. Zhang, T. Cao, S. Li et al., "Parallel processing systems for big data: a survey," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2114–2136, 2016.

[14] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: a survey," *IEEECommunications Surveys & Tutorials*, vol. 19, no. 1, pp. 531–549, 2017.

[15] Y. Zhang, J. Ren, J. Liu, C. Xu, H. Guo, and Y. Liu, "A survey on emerging computing paradigms for big data," *Chinese Journal of Electronics*, vol. 26, no. 1, pp. 1–12, 2017.

[16] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: a survey," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4158–4168, 2019.

[17] J. Hu and A. V. Vasilakos, "Energy big data analytics and security: challenges and opportunities," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2423–2436, 2016.

[18] D. C. Marinescu, A. Paya, and J. P. Morrison, "A cloud reservation system for big data applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 3, pp. 606–618, 2017.

[19] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, vol. 11, no. 1, pp. 88–95, 2017.

[20] P. K. Sahoo, S. K. Mohapatra, and S.-L. Wu, "Analyzing healthcare big data with prediction for future health condition," *IEEE Access*, vol. 4, pp. 9786–9799, 2016.

[21] H. Attaullah, T. Kanwal, A. Anjum et al., "Fuzzy logic-based privacy-aware dynamic release of IoT-enabled healthcare data," *IEEE Internet of Things Journal*, 2021.

[22] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of "big data" on cloud computing: review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.

[23] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, "The anatomy of big data computing," *Software: Practice and Experience*, vol. 46, no. 1, pp. 79–105, 2016.

[24] J. B. Ekanayake, N. Jenkins, K. Liyanage, J. Wu, and A. Yokoyama, *Smart Grid: Technology and Applications*, John Wiley & Sons, 2012.

[25] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid - a review," *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 1099–1107, 2017.

[26] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: a survey," *IEEE Access*, vol. 4, pp. 3844–3861, 2016.

[27] H. Zhang, G. Chen, B. C. Ooi, K.-L. Tan, and M. Zhang, "In-memory big data management and processing: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 7, pp. 1920–1948, 2015.

[28] M. Jaradat, M. Jarrah, A. Bousselham, Y. Jararweh, and M. Al-Ayyoub, "The internet of energy: smart sensor networks and big data management for smart grid," *Procedia Computer Science*, vol. 56, pp. 592–597, 2015.

[29] Y. Mengke, Z. Xiaoguang, Z. Jianqiu, and X. Jianjian, "Challenges and solutions of information security issues in the age of big data," *China Communications*, vol. 13, no. 3, pp. 193–202, 2016.

[30] A. Cuzzocrea, "Privacy and security of big data: current challenges and future research perspectives," in *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data - PSBD '14*, pp. 45–47, 2014.

[31] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: a review," *IEEE Access*, vol. 7, pp. 13960–13988, 2019.

[32] A. Siddiqa, I. A. T. Hashem, I. Yaqoob et al., "A survey of big data management: taxonomy and state-of-the-art," *Journal of Network and Computer Applications*, vol. 71, pp. 151–166, 2016.

[33] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, 2015.

[34] B. P. Bhattarai, S. Paudyal, Y. Luo et al., "Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions," *IET Smart Grid*, vol. 2, no. 2, pp. 141–154, 2019.

[35] B. Speer, M. Miller, W. Schaffer et al., *The role of smart grids in integrating renewable energy*, Tech. Rep., National Renewable Energy Laboratory (NREL), Golden, CO (United States), 2015.

[36] S. Bruno, S. Lamonaca, M. La Scala, G. Rotondo, and U. Stecchi, "Load control through smart-metering on distribution networks," in *2009 IEEE Bucharest PowerTech*, pp. 1–8, Bucharest, Romania, 2009.

[37] V. C. Gungor, D. Sahin, T. Kocak et al., "Smart grid technologies: communication technologies and standards," *IEEE transactions on Industrial informatics*, vol. 7, no. 4, pp. 529–539, 2011.

[38] G. A. Boyd and J. X. Pang, "Estimating the linkage between energy efficiency and productivity," *Energy Policy*, vol. 28, no. 5, pp. 289–296, 2000.

[39] M. Batty, "Big data, smart cities and city planning," *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, 2013.

[40] M. Batty, *Smart cities, big data*, 2012.

[41] R. Moghe, F. C. Lambert, and D. Divan, "Smart stick-on sensors for the smart grid," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 241–252, 2012.

[42] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid — the new and improved power grid: a survey," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[43] A. Arif, M. Al-Hussain, N. Al-Mutairi, E. Al-Ammar, Y. Khan, and N. Malik, "Experimental study and design of smart energy meter for the smart grid," in *2013 International Renewable and Sustainable Energy Conference (IRSEC)*, pp. 515–520, Ouarzazate, Morocco, 2013.

[44] F. Clarizia, D. Gallo, C. Landi, M. Luiso, and R. Rinaldi, "Smart meter systems for smart grid management," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–6, Taipei, Taiwan, 2016.

[45] C. De Capua, G. Lipari, M. Lugara, and R. Morello, "A smart energy meter for power grids," in *Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pp. 878–883, Montevideo, Uruguay, 2014.

[46] J. Zheng, D. W. Gao, and L. Lin, "Smart meters in smart grid: an overview," in *2013 IEEE Green Technologies Conference (GreenTech)*, pp. 57–64, Denver, CO, USA, 2013.

[47] F. Qayyum, M. Naeem, A. S. Khwaja, A. Anpalagan, L. Guan, and B. Venkatesh, "Appliance scheduling optimization in smart home networks," *IEEE Access*, vol. 3, pp. 2176–2190, 2015.

[48] Z. Ahmed, A. Faroqi, and R. M. Navid-ur Rehman, "Implementation of smart system based on smart grid smart meter and smart appliances," in *Iranian Conference on Smart Grids*, pp. 1–4, Tehran, Iran, 2012.

[49] A. M. Carreiro, C. H. Antunes, and H. M. Jorge, "Energy smart house architecture for a smart grid," in *2012 IEEE International Symposium on Sustainable Systems and Technology (ISSST)*, p. 1, May 2012.

[50] G. S. Aleena, P. Sivraj, and K. K. Sasi, "Resource management on smart micro grid by embedded networking," *Procedia Technology*, vol. 21, pp. 468–473, 2015.

[51] G. Suciu, V. A. Poenaru, C. G. Cernat, G. Todoran, and T. L. Militaru, "ERP and e-business application deployment in open source distributed cloud systems," in *The Eleventh International Conference on Informatics in Economy IE*, pp. 12–17, 2012.

[52] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[53] R. Buyya, R. N. Calheiros, and A. V. Dastjerdi, *Big Data: Principles and Paradigms*, Morgan Kaufmann, 2016.

[54] F. Pop, J. K. Lodziej, and B. Di Martino, *Resource Management for Big Data Platforms*, Springer, 2016.

[55] C. Yao, C. Yang, and Z. Xiong, "Energy-saving predictive resource planning and allocation," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 5078–5095, 2016.

[56] H. Won, M. C. Nguyen, M.-S. Gil, and Y.-S. Moon, "Advanced resource management with access control for multitenant Hadoop," *Journal of Communications and Networks*, vol. 17, no. 6, pp. 592–601, 2015.

[57] A. Braganza, L. Brooks, D. Nepelski, M. Ali, and R. Moro, "Resource management in big data initiatives: processes and dynamic capabilities," *Journal of Business Research*, vol. 70, pp. 328–337, 2017.

[58] D. Cao, P. Liu, W. Cui, Y. Zhong, and B. An, "Cluster as a service: a resource sharing approach for private cloud," *Tsinghua Science and Technology*, vol. 21, no. 6, pp. 610–619, 2016.

[59] D. Wu, L. Zhu, X. Xu, S. Sakr, D. Sun, and Q. Lu, "Building pipelines for heterogeneous execution environments for big data processing," *IEEE Software*, vol. 33, no. 2, pp. 60–67, 2016.

[60] M. V. Moreno, F. Terroso-Saenz, A. Gonzalez-Vidal et al., "Applicability of big data techniques to smart cities deployments," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 800–809, 2017.

[61] R. B. Hagmann, "A crash recovery scheme for a memory-resident database system,," *IEEE Transactions on Computers*, vol. 35, no. 9, pp. 839–843, 1986.

[62] M. Zaharia, M. Chowdhury, T. Das et al., "Resilient distributed datasets: a fault-tolerant abstraction for in memory cluster computing," in *9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI}*, pp. 15–28, 2012.

[63] D. Loghin, B. M. Tudor, H. Zhang, B. C. Ooi, and Y. M. Teo, "A performance study of big data on small nodes," *Proceedings of the VLDB Endowment*, vol. 8, no. 7, pp. 762–773, 2015.

[64] N. Davies, A. Friday, S. Clinch et al., "Security and privacy implications of pervasive memory augmentation," *IEEE Pervasive Computing*, vol. 14, no. 1, pp. 44–53, 2015.

[65] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[66] H. Abou-Zeid and H. S. Hassanein, "Predictive green wireless access: exploiting mobility and application information," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 92–99, 2013.

[67] A. Nadembega, A. Hafid, and T. Taleb, "Mobility-prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2561–2576, 2015.

[68] B. Mokhtar and M. Eltoweissy, "Big data and semantics management system for computer networks," *Ad Hoc Networks*, vol. 57, pp. 32–51, 2017.

[69] K. Yang, R. Liu, Y. Sun, J. Yang, and X. Chen, "Deep network analyzer (DNA): a big data analytics platform for cellular networks," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2019–2027, 2017.

[70] T. Ban, M. Eto, S. Guo, D. Inoue, K. Nakao, and R. Huang, "A study on association rule mining of darknet big data," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Killarney, Ireland, 2015.

[71] A. Mohamed, M. Hamdan, S. Khan et al., "Software-defined networks for resource allocation in cloud computing: a survey," *Computer Networks*, vol. 195, article 108151, 2021.

[72] C. Huang and R. Lu, "EFPA: efficient and flexible privacy-preserving mining of association rule in cloud," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1–6, Shenzhen, China, 2015.

[73] M. Dai and Y.-L. Huang, "Organizing the discovered association rules based on general-specific (GS) hierarchical patterns," in *2005 International Conference on Machine Learning and Cybernetics*, pp. 2206–2211, Guangzhou, China, 2005.

[74] W. S. Seol, H. W. Jeong, B. Lee, and H. Y. Youn, "Reduction of association rules for big data sets in socially-aware computing," in *2013 IEEE 16th International Conference on Computational Science and Engineering*, pp. 949–956, Sydney, NSW, Australia, 2013.

[75] X. Zhou and Y. Huang, "An improved parallel association rules algorithm based on MapReduce framework for big data," in *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 284–288, Xiamen, China, 2014.

[76] P. Ducange, F. Marcelloni, and A. Segatori, "A MapReduce-based fuzzy associative classifier for big data," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, Istanbul, Turkey, 2015.

[77] H.-Y. Chang, Z.-H. Hong, T.-L. Lin, W.-K. Chang, and Y.-Y. Lin, "IPARBC: an improved parallel association rule based on MapReduce framework," in *2016 International Conference on Networking and Network Applications (NaNA)*, pp. 370–374, Hakodate, Japan, 2016.

[78] Z. He, Y. He, and L. Wang, "Root causes identification approach based on association rule mining for product infant failure," in *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, pp. 624–628, Hangzhou, China, 2015.

[79] C. Zhou, H. Jiang, Y. Chen, L. Wu, and S. Yi, "User interest acquisition by adding home and work related contexts on mobile big data analysis," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 201–206, San Francisco, CA, USA, 2016.

[80] G. Sheng, H. Hou, X. Jiang, and Y. Chen, "A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 695–702, 2018.

[81] K. K. Zame, C. A. Brehm, A. T. Nitica, C. L. Richard, and G. D. Schweitzer III, "Smart grid and energy storage: policy recommendations," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1646–1654, 2018.

[82] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gomez-Romero, and M. J. Martın-Bautista, "Data science for building energy management: a review," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 598–609, 2017.

[83] X. Li, W. Yu, and S. Villegas, "Structural health monitoring of building structures with online data mining methods," *IEEE Systems Journal*, vol. 10, no. 3, pp. 1291–1300, 2016.

[84] J. Wan, P. Nan, Q. Guo, and Q. Wang, "Multi-mode radar signal sorting by means of spatial data mining," *Journal of Communications and Networks*, vol. 18, no. 5, pp. 725–734, 2016.

[85] L. Dong, K. Wu, and G. Tang, "A data-centric approach to quality estimation of role mining results," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2678–2692, 2016.

[86] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, and M. Pazzani, "Scalable daily human behavioral pattern mining from multivariate temporal data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 3098–3112, 2016.

[87] J.-D. Zhang and C.-Y. Chow, "CRATS: an LDA-based model for jointly mining latent communities, regions, activities, topics, and sentiments from geosocial network data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2895–2909, 2016.

[88] F. Tian, T. Lan, K.-M. Chao et al., "Mining suspicious tax evasion groups in big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2651–2664, 2016.

[89] S. Y. Han, J. No, J.-H. Shin, and Y. Joo, "Conditional abnormality detection based on AMI data mining," *IET Generation, Transmission & Distribution*, vol. 10, no. 12, pp. 3010–3016, 2016.

[90] Y. Zhang and Y.-M. Cheung, "Discretizing numerical attributes in decision tree for big data analysis," in *2014 IEEE International Conference on Data Mining Workshop*, pp. 1150–1157, Shenzhen, China, 2014.

[91] C. Liu, R. Ranjan, C. Yang, X. Zhang, L. Wang, and J. Chen, "MUR-DPA: top-down levelled multi-replica Merkle hash tree based secure public auditing for dynamic big data storage on cloud," *IEEE Transactions on Computers*, vol. 64, no. 9, pp. 2609–2622, 2015.

[92] N.-Q. Doan, M. Ghesmoune, H. Azzag, and M. Lebbah, "Growing hierarchical trees for data stream clustering and visualization," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Killarney, Ireland, 2015.

[93] F. Yuan, F. Lian, X. Xu, and Z. Ji, "Decision tree algorithm optimization research based on MapReduce," in *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 1010–1013, Beijing, China, 2015.

[94] X. Huang, H. Zhou, and W. Wu, "Hadoop job scheduling based on mixed ant-genetic algorithm," in *2015 International*

Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 226–229, Xi'an, China, 2015.

[95] Q. Lu, S. Li, and W. Zhang, "Genetic algorithm based job scheduling for big data analytics," in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, pp. 33–38, 2015.

[96] M. De Sanctis, I. Bisio, and G. Araniti, "Data mining algorithms for communication networks control: concepts, survey and guidelines," *IEEE Network*, vol. 30, no. 1, pp. 24–29, 2016.

[97] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, "Genetic algorithm based data-aware group scheduling for big data clouds," in *2014 IEEE/ACM International Symposium on Big Data Computing*, pp. 96–104, London, UK, 2014.

[98] Y. Zhang, Z. Jing, and Y. Zhang, "MR-IDPSO: a novel algorithm for large-scale dynamic service composition," *Tsinghua Science and Technology*, vol. 20, no. 6, pp. 602–612, 2015.

[99] M. O. Ulfarsson, F. Palsson, J. Sigurdsson, and J. R. Sveinsson, "Classification of big data with application to imaging genetics," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2137–2154, 2016.

[100] P. Agarwal, R. Ahmed, and T. Ahmad, "Identification and ranking of key persons in a social networking website using Hadoop & big data analytics," in *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, p. 65, 2016.

[101] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: a set theoretical approach to big data analytics," *IEEE Access*, vol. 4, pp. 2542–2571, 2016.

[102] G. Ghosh, S. Banerjee, and N. Y. Yen, "State transition in communication under social network: an analysis using fuzzy logic and density based clustering towards big data paradigm," *Future Generation Computer Systems*, vol. 65, pp. 207–220, 2016.

[103] M. Narayanan and A. K. Cherukuri, "A study and analysis of recommendation systems for location-based social network (LBSN) with big data," *IIMB Management Review*, vol. 28, no. 1, pp. 25–30, 2016.

[104] M. Zaharieva, M. Del Fabro, and M. Zeppelzauer, "Cross-platform social event detection," *IEEE MultiMedia*, vol. 22, no. 3, pp. 14–25, 2015.

[105] O. Liu, K. Man, W. Chong, and C. Chan, "Social network analysis using big data," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pp. 6-7, 2016.

[106] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[107] E. Cambria, N. Howard, Y. Xia, and T.-S. Chua, "Computational intelligence for big social data analysis [guest editorial]," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 8-9, 2016.

[108] J. Zhang, F. Xia, Z. Ning et al., "A hybrid mechanism for innovation diffusion in social networks," *IEEE Access*, vol. 4, pp. 5408–5416, 2016.

[109] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Network*, vol. 28, no. 4, pp. 32–39, 2014.

[110] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower son with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[111] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: big data toward green applications," *IEEE Systems Journal*, vol. 10, no. 3, pp. 888–900, 2016.

[112] K. Yu, M. Arifuzzaman, Z. Wen, D. Zhang, and T. Sato, "A key management scheme for secure communications of information centric advanced metering infrastructure in smart grid," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2072–2085, 2015.

[113] C. Hewitt, "Orgs for scalable, robust, privacy-friendly client cloud computing," *IEEEinternet computing*, vol. 12, no. 5, pp. 96–99, 2008.

[114] Z. Han, M. Bennis, D. Wang, T. Kwon, and S. Cui, "Special issue on big data networking-challenges and applications," *Journal of Communications and Networks*, vol. 17, no. 6, pp. 545–548, 2015.