Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 3976302, 14 pages
https://doi.org/10.1155/2023/3976302

Review Article

Big Data Analytics, Processing Models, Taxonomy of Tools, V's, and Challenges: State-of-Art Review and Future Implications

Sandeep Dasari and Rajesh Kaluri

School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India

Correspondence should be addressed to Rajesh Kaluri; rajesh.kaluri@vit.ac.in

Received 6 June 2022; Revised 8 October 2022; Accepted 13 April 2023; Published 10 May 2023

Academic Editor: Zengpeng Li

In the current digital era, data is budding tremendously from various sources like banks, businesses, education, entertainment, etc. Due to its significant consequence, it became a prominent proceeding for numerous research areas like the semantic web, machine learning, computational intelligence, and data mining. For knowledge extraction, several corporate sectors depend on tweets, blogs, and social data to get adequate analysis. It helps them predict the customer's tastes and preferences, optimize the usage of resources. In some cases, the same data creates complications that lead to a problem named as big data. To solve this, so many researchers have given various solutions. Based on literature analysis formulated 6-s simulation towards big data, detailed information about characteristics, a taxonomy of tools, and discussed various processing paradigms. No one tool can truly fit for all solutions, so this paper helps to make decisions smoothly by providing enough information and discussing major privacy issues and future directions.

1. Introduction

Advancement in digital resources leads to the rapid evolution of massive collections of data, which include complex, sprouting data sets that are congregated from heterogeneous data sources. Such voluminous data is termed as big data. It was coined by a scientist named John R. Mashey in the 1990s. Every online activity creates some data, especially internet usage, which has become a basic need in COVID-19 pandemic time. Facebook receives more than 500 terabytes of data in terms of photos, videos, messages, comments, etc. [1]. On average each day, 2.5 quintillion bytes of data can be generated, and in today's world, 90% of the information was developed in the last two years [2]. This data is available in the form of structured, unstructured, and semistructured data types. Rapid evolution in data processing mechanisms makes big data to become a hotfoot in all science and engineering domains, including physical, biological, and biomedical sciences [3].

Before explorations in data analysis technologies, organizations could not be able to handle their archives efficiently by applying traditional techniques [4]. Commodity systems are restricted with storage, invariable tool management, performance, scalability, and flexibility [5]. In the case of privacy concerns, data will also get an effect. Especially that the privacy of data is very important in the case of maintaining sensitive credentials like bank details, legal information, medical records, biometric details of officials, etc. By adding machine-learning features, the scope of safeguarding the data can be increased [6]. It can be perfected by applying a concept called privacy preservation machine learning (PPML) [7]. Its processing mechanism is shown in Figure 1. Majorly, it includes differential privacy, homomorphic encryption, multiparty computation, federated learning, and ensemble techniques [8]. Each approach plays an effective role in solving various privacy concerns and reduces the vulnerability intensity exposed to an attacker. In the online era, preserving the privacy of end users on social media has been regarded as a bottleneck issue.

Ironically, as the data analytics introduced in this paper become more progressive, the risk of privacy is also growing. As such, many privacy-preserving solutions have been

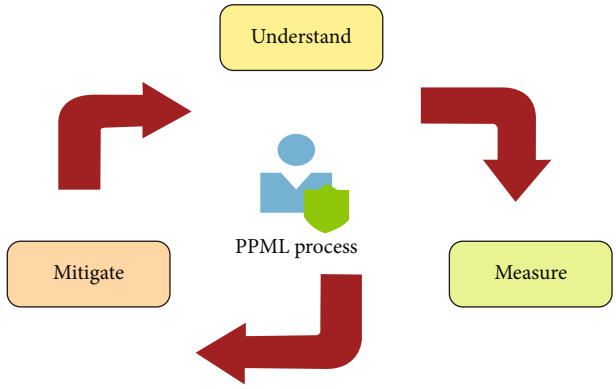

Figure 1: PPML process.

proposed to address these issues. There are two main approaches. The first one is k-anonymity, which is a property implemented by certain unknown data. Given the private data and a set of selected fields, the system has to make the practically useful data without finding the individuals who are the subjects of the data. The second one is differential privacy, which can provide an effective way to improve the quality of queries from statistical databases by reducing the chances of finding its records. The evolution of big data can be considered in 3 phases, as shown in Figure 2, each with its own set of features and capabilities.

Although big data has been accepted by so many organizations, research on big data in the cloud is in its early stages. Several concerns have not been fully addressed. Moreover, new challenging issues continue to emerge from various applications by organizations. Some of the prominent research challenges are as follows [9].

1.1. Data Integrity. An important feature of big data is to obtain integrity in security. Integrity means that data will be updated only through authorized persons. The expansion of cloud services provides users with a repository and mastery of their data in the data centers of the cloud. Such applications must check the data integrity, and another main challenge is to make sure about the rightness of the users' data. Users may not be physically able to get the data, but the cloud should provide flexibility for the user to scrutinize how the data can be maintained.

1.2. Data Transformation. Converting the data into a form that can be worthy of analysis is a major issue in big data. Due to the various data formats, big data can be transformed into an analysis workflow in two types. If the data is structured, it can be preprocessed before it is stored in relational databases. Later, the data can be retrieved to perform analysis. In case of unstructured data, it must be stored in distributed databases, such as HBase, before it is processed for analysis.

1.3. Data Quality. In the earlier days, data processing could be done on clean datasets from known and limited sources. Therefore, the outcomes are in an acceptable manner. However, with the inception of big data, data can be produced from various sources; all these resources are not well-known or verifiable. Fewer data quality has become a significant issue for so many service providers because data are frequently collected from huge sources [9]. For instance, large volumes of data are produced from smartphones, where inconsistent data formats can be produced as a resultant of heterogeneous sources.

1.4. Data Governance. It involves leveraging information by aligning the objectives of multiple functions, such as telecommunication carriers having access to huge troves of user's information in various forms like call details and marketing seeking to monetize this information by selling it to third parties. Moreover, big data plays an active role in providing opportunities to service providers by making users information more valuable.

2. Role of Data Analytics

Data will not be helpful or valuable until it is appropriately integrated and understood. Organizations are utilizing big data advances to explore the results of continuous information-driven choices. Various data analytics processing mechanisms help to stay ahead of business challenges by improving new designs and products, customer personalization, successful marketing, the generation of new profit options, and improved processing efficiency [10].

Real-time scenarios that spectacle the big data importance are as follows:

(i) By 2025, the big data analytics industry is likely to be generated more than $103 billion

(ii) The US economy loses up to $3.1 trillion per year due to poor data quality

(iii) Every individual in 2020 created 1.7 gigabytes in less than a second

(iv) Every day, internet users generate 2.5 quintillion bytes of data

(v) 97.2 percent of entrepreneurs are investing in artificial intelligence and big data

(vi) Netflix gets a profit of $1 billion each year on user retention thanks to big data

2.1. Working with Big Data Analytics. Data scientists and other statisticians collect data from different data sources. Data preprocessing can be performed on data such that it can be made accessible for processing systems to understand and analyze the massive volume of data. Valuable insights can be generated by applying machine learning algorithms [11]. The data analytics process includes four steps: data collection, cleaning, processing, and analysis [12]. The data necessary for analysis is based on a question or an experience. Depending on the demands that lead to analyzing the data to use as input to the study, specific population factors can be identified and information can be categorized [13].

Data collection is the practice of acquiring information on specified variables specified as data requirements. The main emphasis is on collecting data that is accurate and

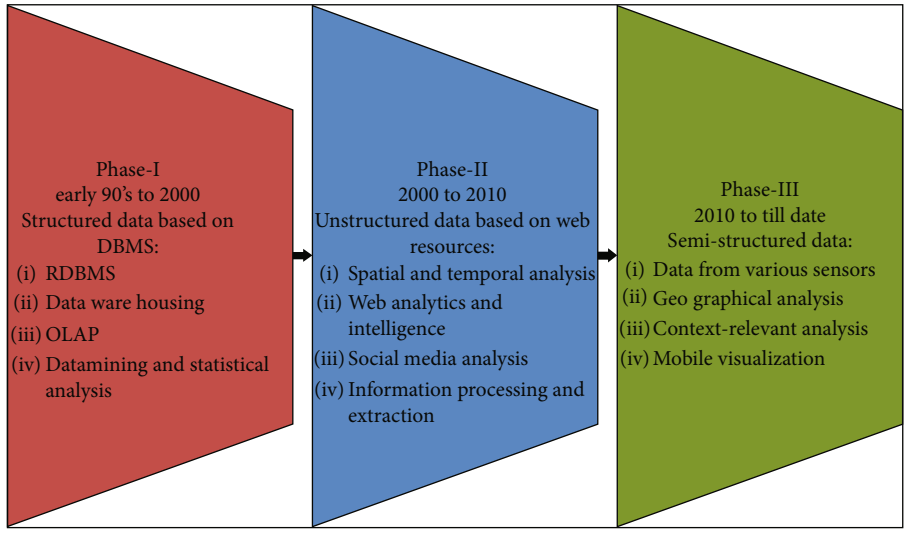

Figure 2: Evolution phases of big data technologies.

authentic. It does not guarantee that the data obtained is accurate. It gives a baseline against which to measure progress as well as a user's requirement [14]. Data is gathered from various sources, including organizational databases and web page information. The resulting data could be unstructured and contain irrelevant information, and it must undergo data processing and cleaning. For analysis, the obtained information must be transformed or structured, which includes reconstructing the data to match the needs of the various analysis tools [15].

The data that has been extracted and arranged may be duplicated, incomplete, or result in false reports. To avoid this problem, data cleaning can be induced. Depending on the nature of the data, different forms are available. Specific results might be evaluated against reliable published amounts or established thresholds. While cleansing financial data, quantitative data methods can also be used to find outliers that are eliminated from further investigation [16]. After it had been processed, organized, and sanitized, the data would be ready for analysis. Various data analysis techniques are available to analyze, evaluate, and develop conclusions based on the criteria [17]. Data visualization is used to examine data graphically to obtain an effective analysis of the data. With the help of visual elements like line graphs, bar graphs, and area charts, these tools provide an acceptable way to see and understand patterns, outliers, and trends in data. Additionally, it is a better way for entrepreneurs to impart data to a nontechnical audience.

To find the relationships between statistical data models and variables, such as regression and correlation analysis can be utilized. These data-descriptive models aid in the simplification of evolution and communicational outcomes. The process may involve additional cleaning or data collection, and therefore, these activities are iterative. The following Figure 3 represents various analytic use cases involved in data.

2.1.1. Descriptive Analysis. It is a statistical approach that can be used to summarize historical data and find patterns and facilitate in the preparation of reports including financial

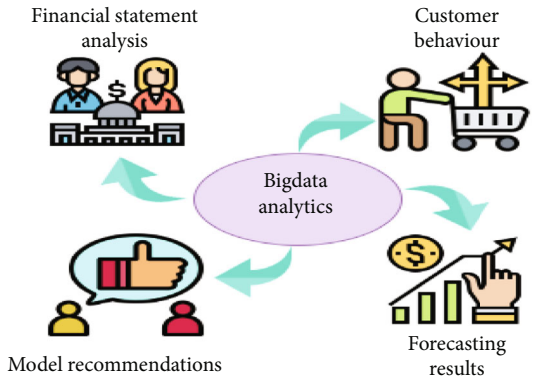

Figure 3: Different use cases of big data analytics.

statements, revenue, and marketing by summarizing data in an accessible way [15]. Data aggregation and data mining are two approaches involved in this analysis to discover historical data. Data is first collected and arranged by data aggregation to make the datasets more viable for analysts. Data mining involves a search of the data to identify patterns and meaning. Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment. Quick and easy reports which show how performance can be achieved, identifying performance issues and gaps earlier—before they become serious complications.

2.1.2. Diagnostic Analysis. To point out what created the issue, data mining, drill-down, and data recovery are examples of techniques [18]. Businesses use this because they give a thorough understanding of the issue. For companies that collect customer data, diagnostic analysis is the key to understanding why customers select particular orders. These insights will help improve products and the user experience.

2.1.3. Predictive Analysis. This type of analytics looks at both previous and present data to make forecasts for the future. It

examines existing data and forecasts future data mining, artificial intelligence, and machine learning trends [18, 19]. It speculates on customer and market trends; for online learning specifically, predictive analytics is often found incorporated. Personalizes the training needs of learners to find strengths, weaknesses, and gaps; required learning resources and training can be offered to improve individual needs.

2.1.4. Prescriptive Analysis. This type of analytics provides a solution to a specific issue. Both predictive and descriptive analysis can be used in perspective analytics [16]. AI and machine learning can be used for this type of analysis. It generates recommendations and makes decisions based on the computational findings of algorithmic models. Automated decisions or recommendations required specific, unique, and clear directions from those utilizing the analytical technique to analyze specific learners who required additional support, regardless of how many students or employees and identify0 outstanding learners to provide additional resources [20].

2.2. Big Data Can Influence the Decision-Making Process as Follows

2.2.1. Customer Service Metrics in Real Time. Providing services to the customer is one of the essential areas in which firms must deliver metrics nowadays. Firms exploit real-time data to provide customers with one-on-one customized services and solutions [21]. Big data helps provide users with personalized loyalty programs.

2.2.2. Improving Operational Efficiency. Currently, companies are trying to use data for automating operations, optimizing trading strategies, and getting overall corporate efficiency outcomes. For example, vehicles equipped with sensors collect data and transmit it to central servers for analysis [22]. Personal car owners are also notified about prior repairs or services, which resource the company in enhancing the performance of its automobiles [23].

2.2.3. Increased Efficiency without Investing More. Without investing more resources, one can envision the customers. Sprint, a telecommunications firm, employs big data analytics to analyze real-time information to assess network failures, maximize resources, and improve customer experience [24]. This can result in enhancing the brand's delivery rate.

Some of the remarkable benefits of using big data analytics are as follows:

(i) Data from various sources like Twitter and Facebook helps better customer retention, marketing strategy decisions, and business intelligence for better insights and predictions

(ii) Customer service issues can be resolved quickly with NLP (natural language processing) features to improve customer satisfaction, solve problems, etc.

(iii) Operational efficiency improves results and produces when large volumes of data are well-analyzed and used to update services, security issues, healthcare, IoT, etc. [11].

(iv) Errors can be identified in an early stage to reduce the risk to services

(v) Big data analytics lay a path for large data warehouses to integrate with multiple sources and advanced technologies and process the data

3. Related Work

Budding big data applications has got progressively significant over the past couple of years. Previously proposed systems help big data applications assist to store, investigate, and measure information. Adopting the right combination of technology is also vital to obtaining the best results [25]. Parallel processing became an arising multidisciplinary area of research that is progressively drawing the consideration of scientists and professionals from different areas, including technology, intellectual, and sociologies [4]. In this paper, the authors contributed their work on existing frameworks and challenges of big data with assessment evaluation which provides a comparative study of various accessible frameworks and workloads. Existing information handling innovations are not appropriate for adopting the extraordinary measures of produced information like big data sets and conventional information strategies; they show moderate responsiveness and lack of versatility, execution, and precision. So many data collection challenges remain to be addressed. It may lead to a compromise of information security to a great extent [26].

The combination of AI and information management approaches offers better outcomes. However, many explorations work center on big data. Hadoop, Spark, Storm, and Flink frameworks, Flume, Kafka, GFS, HDFS, Smaza, and spark streaming technologies effectively implement data [5]. Provided elaborate focus on case studies on mLib, Spark, Flink, and Mahout [6]. Deep learning helps to handle a vast volume of data, which is common in big data, and gives acceptable results with complex datasets [7, 27, 28]. Boundless numbers of challenges, tools, and technology considerations are provided in [14, 16, 29, 30].

In [31], the authors worked on dimensions in big data and data analysis and then focused attention on the issue of inconsistencies and their impact on data analysis. They introduced four classifications of inconsistencies, frameworks, and tools in big data and pointed out the efficiency of inconsistency-induced learning as a tool for data analysis.

In [32], the authors discussed the infrastructure of big data service architecture, which includes collecting and storing data. Practical application scenarios of Flume, Kafka, GFS, HDFS, and different data processing modes are focused. They introduced mobile big data development, the commonly applied data analysis techniques. Three typical mobile big data analysis applications, namely, wireless channel modeling, human offline and online behavior analysis, and speech recognition in the Internet of Vehicles, are introduced. To obtain the minimum consistent subset, the concepts of edited nearest neighbor, traditional condensed

nearest neighbor, and reduced nearest neighbor were used [33].

In [34], the authors spotted light on complications faced by data scientists in integrating and implementing a high volume of medical data acquired from multiple platforms. Provided information about the medical field, digitization of information, bioinformatics tools, and suggested improvements required in the e-health sector. Table 1 describes key findings and the scope of research contributions in big data and machine learning.

4. Stimulation towards Big Data

The emerging goal of massive data analysis is to update commodity techniques such as rule-based systems, model mining, and decision trees. Other data mining techniques to develop business rules on massive data sets effectively can be realized either by improving algorithms that use the storage of distributed data, computation in memory, or by using the computation in clusters for parallelism [48]. Previously, this was done using grid computing, which has been taken by cloud computing in the present day [49].

Traditional databases use code locality to process the data, but it is hard to import or export. Meanwhile, there may be a chance that data gets manipulated. Processing the data within time is also a big deal [22]. Cloud techniques help to work effectively, revolutionizing the IT industry by adding elasticity to the way it was consumed and enabling organizations to pay only for the resources and services they use. It is necessary to integrate big data issues with cloud technologies to produce effective results in the current scenario [50].

To handle large records, statistics integration usually needs to extract vast quantities of records from massive sources [51]. These dispensed statistics are desired to gather with the aid of a suitable device or software program, and information storage control schemes need to be provided for these massive facts within the sequential processing steps. The forms of big data majorly involved are static batch data and dynamic stream data [52]. Batch data processing is stored in a static format, and stream data is a continuous real-time data instance sequence. The streaming data cannot be stored fully, and various elements can be removed after processing.

4.1. 6S: A Beeline towards Big Data. The following points present why all major sectors are adopting big data technologies.

(1) *Supervision: better data management*. The majority of data processing platforms and business intelligence tools let data scientists sit in one place and drive the data analysis, helping to perform various types of operations without technical complications. This includes organizing, collecting, and storing. Spending on analytics to gain competitive intelligence on future market conditions, to target customers more successfully, and to optimize operations and supply chains generates operating profit increases profits

(2) *Suppleness: better speed, capacity, and flexibility*. Big data services can be provided to utilize substantially large data sets which can provide the necessary storage and computing power to change data according to requirement

(3) *Sageness: better visualization*. Data visualization tools support observing data in pictorial representations like graphs which can be easy to understand. In real-time scenarios also visualization tools can process voluminous data quickly. Managers can use big data to understand more about their businesses insights and transform the generated knowledge into effective decisions to enhance company performance

(4) *Scope: better opportunities*. More consumers understand the competitive benefit of becoming a data-driven company as big data analytics technologies improve. Nowadays, marketers majorly focus on sentiment analysis, where they can collect data on how customers think about certain products and services by analyzing consumer responses on social media sites like Facebook and Twitter

(5) *Statistics: better data analysis methods*. The data are not just figures in a database anymore. Text, audio, and video files can also provide valuable insight; good tools can even recognize specific models according to predefined criteria. This happens in large part through the use of natural language processing tools, which can be essential for text mining, sentiment, and clinical analysis. The healthcare sector uses big data to improve patient care and to find better ways to manage resources and personnel

(6) *Surety: better risk analysis*. Risk is a facet of almost every business decision. There is no way to avoid risk, especially when a company is looking for growth, diversifying products, or trying to achieve new targets. In addition to financial markets, big data risk management can be applied in healthcare, retail, manufacturing, and e-commerce and can be applied to a wide variety of corporate threats, such as regulatory risk [53]. Financial institutions can fastly find that big data analysis is expert at identifying fraud before it becomes extensive, preventing further damage to the organization

5. Big Data Applications

5.1. Recommendations to Customers. Customer data are collected in various no. of ways, including what websites they browse, where they reside, when they approach customer support, and if they communicate with their brand on social media. It is a massive amount of seemingly unrelated data, but organizations that can correctly mine it may provide a more tailored experience [39]. Companies must offer the appropriate products to the targeted consumer on the right

Table 1: Major contributions in big data and machine learning research.

References	Year	Scope of analysis	Key findings
Islam et al. [35]	2022	Detection of distributed DDOS attacks on financial organizations. Multiple classification models are used for the prediction of DDOS attacks. SVM, KNN, and RF algorithms are used.	SVM produces the best results, compared with other algorithms.
Najar and Naik [36]	2022	Detection of DDoS attack packets using K-nearest neighbor, random forest, support vector machine, and multilayer perceptions.	Random forest shows better performance compared with other algorithms.
Ananthu et al. [37]	2021	Recognizing fraud transactions by analyzing the transaction records and integrating big data with machine algorithms for accurate and fast detection.	Performed comparison of RF, logistic regression, and decision tree classifier; RF produces better results.
Wang et al. [38]	2020	Big data architecture, data collection and storage processing using ETL, different types of NoSQL databases with their merits, demerits, scenarios, and various processing strategy modes are discussed.	Investigated service, present architecture, and cloud services in big data.
Mahmud et al. [39]	2020	Discussed problems in sampling and partitioning of data analysis, record and block-level samplings, and three classical horizontal partitioning schemes.	Sampling-based approximation projects are considered for analysis.
Shoumy et al. [40]	2020	Applications, trends, and state of art technical analysis and their performances.	Building a comprehensive multivariant database for qualitative analysis.
Hasliza et al. [29]	2020	Challenges and issues faced in the consolidation of data and implementation in the public sector.	Consequences in data due to lack of management support, policies, human errors, and improper maintenance.
Ketu et al. [23]	2020	Illustrated analysis on Hadoop and Spark, evaluation can be done based on functional principle, performance, compatibility, failure tolerance, cost, flexible use, data processing, and security.	In-memory computations of spark are more effective than on-disk Hadoop computations.
Kshirsagar and Kumar [41]	2020	Feature reduction method based on correlation, information gain ratio, and ReliefF.	Accuracy and feature section procedure was improved by implementing PART classifier.
Hiriri et al. [34]	2019	Presented a background study of V's, theories and techniques involved in big data, and comparison of uncertain types respective of data analysis.	Discussed state of art analytical techniques, the impact of uncertainty, and open issues focused on data related to financial sector decision processing.
Rao et al. [42]	2019	Tools, technologies, and functionalities involved in big data, parameters to be considered for Hadoop query processing, HACE theorem, and various data sources for data analysis.	Considered real-time scenarios on distributed ML tools presented core features of recent developments in large-scale graph processing and tools
Bindra and Sood [43]	2019	Detection of DDoS attacks by applying machine learning models.	Random forest is the best choice for identifying DDoS attacks.
Inoubli et al. [3]	2018	Investigated challenges in big data and provided a glance of various frameworks, presented an experimental analysis, and a comparative study of the most in demand frameworks with various batch and iterative workloads.	A comparison study on frameworks, processing models, and best practices.
Kolajo et al. [44]	2018	Focused on fraud detection systems, analyzing massive streams of credit card transactions, addressing verification latency, class imbalance, and concept drift.	Impact of concept drift and class unbalance in a real-world data set consisting of 75 million transactions.
Chen et al. [45]	2018	Classification of DDoS attack.	Classification model built on spark framework to achieve performance and accuracy.
Qiu et al. [46]	2016	Provided promising learning methods such as transfer learning, active learning, representation learning, deep learning, distributed learning, parallel learning, and kernel-based learning.	The connection of modern signal processing technologies with machine learning was analyzed and provided open issues and new research trends.
Landset et al. [47]	2015	Discussed different processing paradigms and comparison of engines including MapReduce, Spark, Flink, and Storm.	Flink gives the results with a combination of batch and streaming models.

channel to accurately forecast the future and maintain resources in terms of human capital management [54].

5.2. Medical Field. Healthcare analysis aims to assist doctors in making data-driven decisions in proper time and improving patient care [55, 56]. This is more effective in the scenario of one who has a long medical history and is suffering from several health issues [57]. New AI systems and technologies would also be able to anticipate who is at risk of diabetes, allowing for additional screenings or weight control to be recommended [58].

5.3. Mobile Communications. Mobile service providers can inspect network speed and control the whole network using network analytics, which is a huge concern in telecom [59]. This enables network faults to be resolved in a matter of minutes while also improving service quality and customer satisfaction. With the proliferation of smartphones, location-based support services can be supplied to customers upon demand, based on analysis of real-time location and behavioral data [60]. This could increase the number of people who use mobile services [40].

5.4. Economic Firms. Financial analytics provides compelling chances to improve predictive modeling and better prediction of rates of return and investment effects. More precise forecasts and the capacity to lessen the inborn perils of monetary exchange result from admittance to large information and more noteworthy algorithmic information. Companies are attempting to comprehend client needs and preferences to guess future behaviors, grow in sales leads, utilize new channels and technology, meet consumer needs, and enhance customer satisfaction [61–64].

5.5. Customer Practice. The marketing functions of social media are constantly being investigated and employed nowadays, and their economic value is more apparent. New business trends in social media are moderately obtaining acceptance and popularity among consumers. Entrepreneurs can master even more extensive personal information about consumers through social media, which allows them to correctly assess their personal preferences, activities, and other information, allowing them to effectively address deep needs and access potential demands [60].

5.6. Business Marketing. The more information a company gathers, the more opportunities it has to make a pave in marketing, provide better services, increase consumer interaction, boost its brands, and reach the right people. Advertising is also an important aspect of marketing. In fact, marketing can be grown by advertising products in numerous channels [32]. Therefore, if a company adopts a practical approach to the data it collects from various channels, it can change the entire marketing strategy through data analysis.

5.7. Law Enforcement. Using historical data, such as kind of crime, location, scheme, social media data, drone, and smartphone tracking, law enforcement officers attempt to forecast the next crime site [65]. The police agency could determine the spatial association between crime sites and environmental factors.

5.8. New Product Development. The process of introducing a new product involves a lot of trial and error. Big data eliminates guesswork and aids in the development of perfectly suitable through effective management [66]. It makes to achieve product optimization, time and cost reduction, and service offering that would lead to less supervision and latency reduction; therefore, the minimum amount of resources is required [67].

5.9. Banking. In service delivery and operations, the banking sector has progressed by leaps and bounds over the past few decades [68]. Remarkably, most banks have not been able to utilize the data stored in their systems. Through proper data analysis, bank scans achieve fraud detection, prevention, customer segmentation, risk management, and recognized product offerings [69].

5.10. Education. In education, the pedagogical judgments made by a committee to assess a student's knowledge of the content or structuring of a course may have more impact on student learning and graduation rates [70]. This increases learning efficiency and not only enhances the student's experience but it also helps to evade some of the educational system's needs [71]. Data analytics opens up new possibilities for improving education by assisting instructors and students in making better decisions earlier in the learning process. Developments in applying data science to drive process innovation are rapidly expanding.

5.11. Emotional Analysis. The magnification of various social association networks produces huge data related to emotional analysis. It is a procedure of measuring peer emotions, thoughts, and results used to draw out effective information [71]. Machine learning algorithms play a meaningful role in this process; algorithms like SVM, K-nearest neighbor, genetic algorithm, ANN, and random forests can be used to facilitate this analysis [72].

6. Big Data Challenges

Big data originated with autonomous sources is heterogeneous, large-volume, distributed, and decentralized control and seeks to explore complex and evolving relationships among data [73].

Innovative data analysis tools and techniques must be required to meet large-scale data set analysis challenges and targets [74]. Therefore, we require a potential framework that meets processing speed and scalable storage systems for large-scale datasets. A lot of good research has been done to overcome these issues [75, 76]. Key challenges are categorized as data characteristics V's, process, and management challenges.

6.1. Data Characteristics: V. Big data producing data on a large scale poses three significant problems. They are data volume, velocity, and a variety of data. These are referred to as the 3 V-model of big data. Further, the model has been

Table 2: The V models.

S no.	Model name	V-full forms
1.	3 V's	Volume, velocity, variety
2.	4 V's	3 V's+veracity.
3.	5 V's	4 V's+value.
4.	7 V's	5 V's+variability and visualization.
5.	10 V's	7 V's+validity, vulnerability, and volatility.
6.	13 V's	10 V's+vocabulary, vagueness, venue.
7.	42 V's	13 V's+vane, vanilla, vantage, variability, varifocal, varmint, varnish, vaticination, vault, veer, veil, versed, vexed, victual, viral, virtuosity, viscosity, vivify, voice, voodoo, voyage, vulpine, valor, verdict, version control, vibrant, vet, vastness, and visibility.
8.	51 V's	42 V's+verification, verbosity, versatility, voluntariness, virtualization, violation, vitality, verve, and venturesomeness.

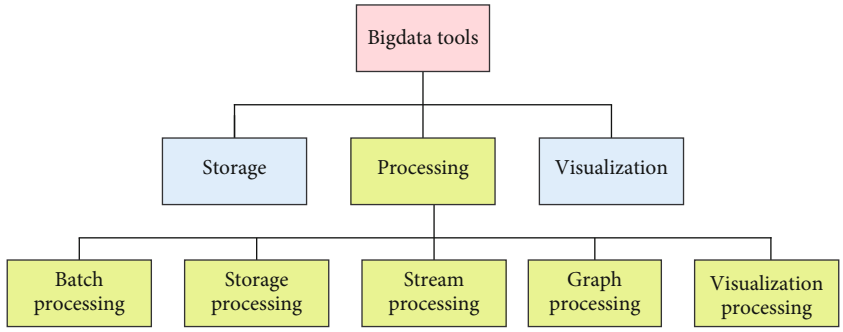

Figure 4: Big data processing tools classifications.

extended as 4 V's, 5 V's, 7 V'S, 10 V's, 13 V's, 42 V's, 44 V's, and 51 V's [77, 78], as shown in Table 2.

6.2. Process Challenges. These are related to processing and analyzing large datasets. It creates a major challenge, as the data is available in various forms and changing it into a suitable form for analysis is a challenging task. It involves data acquisition and storage, preprocessing, data analysis, modeling, and data visualization.

6.3. Management Challenges. These are related to the challenges encountered by an organization which are related to the privacy, security, and governance of data. These are also faced due to a lack of skilled experts who are well known for the trending tools and techniques. To place the appropriate method for dealing with each phase of data. Security is always a major concern as data is highly confidential such as financial and military data.

7. Tools for Big Data Analysis

The computational tools for big data are used to process data at different levels, which will help integrate and analyze various processing mechanisms. Big data software tools extract information from many data sources and process it. In conventional databases, tracking this data is quite challenging. As a consequence, we can employ tools to manage data [74]. Most of the big data technology logos include jungle animals. Logos are intended to signify something and often have a backstory. Big data can be represented with an elephant image that expresses that it is giant, intense, and complex to handle. Some more examples include pig, hive, and zookeeper. The evolution of tools required latency, throughput, fault tolerance, usability, resource expense, and scalability. Major classifications of processing tools are shown in Figure 4.

Before choosing a tool, one has to check the following aspects:

(1) License cost

(2) Customer service quality

(3) The expense of teaching staff

(4) The tool support and updates the policy of the vendor's software requirements

(5) Company assessment

7.1. Taxonomy of Tools

7.1.1. Storage Processing. The quick expansion of information requires more stringent data storage and management requirements [79]. The management and storage of large-scale data sets while maintaining data access dependability and availability are said to as big data storage. Major concepts include substantial storage systems, distributed storage solutions, and effective data storage mechanisms. On the one hand, the storage architecture should provide dedicated

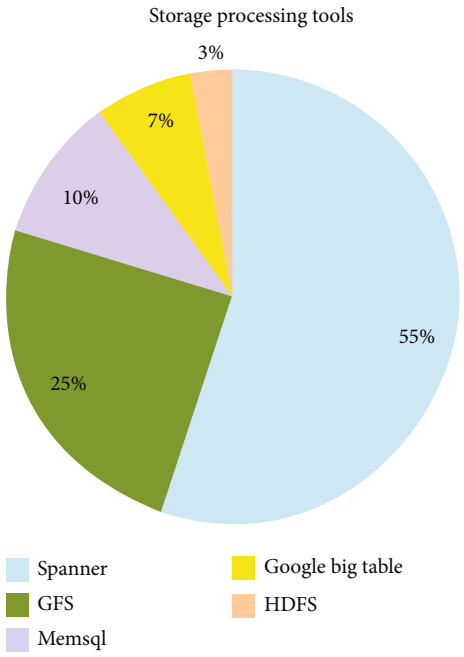

Figure 5: Storage processing tool statistics.

space for information storage; on the other hand, it needs to provide a robust, accessible interface for queries and understanding large amounts of data [80]. Data storage devices are traditionally used as auxiliary server equipment to store, manage, search, and analyze data using structured RDBMS. With the rapid improvement of data, data storage devices are getting more significant, and numerous internet companies are pursuing large storage capacities to stay competitive. As a result, there is an additional demand for data storage in research. For storage processing analysis, tools considered are GFS, HDFS, spanner, memSQL, Google Bigtable, Sqoop, and Flume, and among them, Figure 5 presents the popular storage processing tools statistics.

7.1.2. Batch Processing. Batch processing is essential for companies to manage massive amounts of data effectively. Particularly well suited to working frequent repetitive chores like accounting, the fundamentals of batch processing are the same in every business and for every project [70]. It has become prominent due to its numerous benefits in the field of enterprise data management. Batch processing has several advantages for businesses. Efficiency, when computing or other tools are readily available, helps process jobs. Companies can schedule batch operations for jobs that are not as urgent and prioritize time-sensitive jobs [81]. It can also be processed in the background to lower the processor burden. Compared to stream processing, batch processing is a less sophisticated system that does not require particular hardware or system support for data entry [44]. It requires minimal maintenance. For batch processing analysis, consider the following tools: Hadoop, Dryad, Mahout, Jaspersoft, Pentaho, Skytree Server, Tableau, Karmasphere, Talend, and MapReduce, and among them, Figure 6 presents the popular batch processing tools statistics.

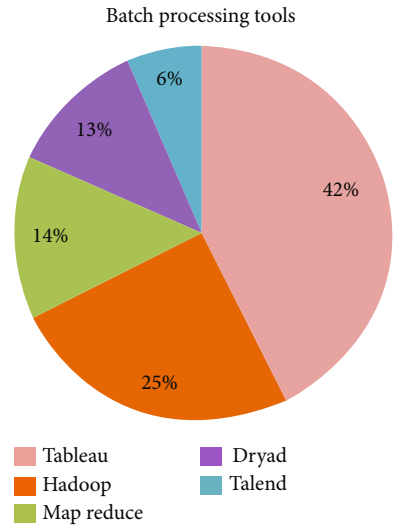

Figure 6: Batch processing tool statistics.

7.1.3. Stream Processing. The majority of data is now created in the structure of a stream. Batch data is just a snapshot of low-level data taken at a specific point in time [44]. In this perspective, data is delivered rapidly, one instance at a time, and algorithms must analyze it in a single pass while adhering to very severe space and time limits [82]. Streamlining algorithms apply probabilistic assurance to provide quick estimated results [83]. Settings in the streaming paradigm have been developed and abstracted by researchers.

On the one hand, MapReduce is unsuitable for expressing streaming algorithms [84]. On the other hand, conventional online algorithms are constrained by the memory and bandwidth of a single console. DSPEs (distributed stream processing engines) are a new breed of MapReduce-inspired technologies that aim to solve this problem [85]. These engines enable parallel computing to be expressed as streams, combining multiprocessor scalability with the efficiency of streaming algorithm techniques [86]. For stream processing tools analysis, considered the following tools: Storm, S4, SQLstream s-Server, Splunk, Apache Kafka, SAP HANA, Samza, Flink, Samoa, Millwheel, heron, cloud-based streaming, Amazon Kinesis, s2, Microsoft Azure, and IBM streaming analysis. Among this, Figure 7 gives a statistical view of popular stream processing tools.

7.1.4. Graph Processing. Graph analysis can be employed to produce recommendations and customization models for customers and to take critical decisions based on the data analysis findings [87]. This helps the enterprises potentially guide customers to buy their products, marketing approach, and customer service behavior. Several scenarios present graph databases as a more suitable match for data management than relational databases and other NoSQL data storage [83].

Graph data solutions assist in detecting fraudulent transactions in a payment processing application using related data that comprises people, transactions, products, and events. Topic modeling entails approaches for grouping documents and extracting thematic representations from the

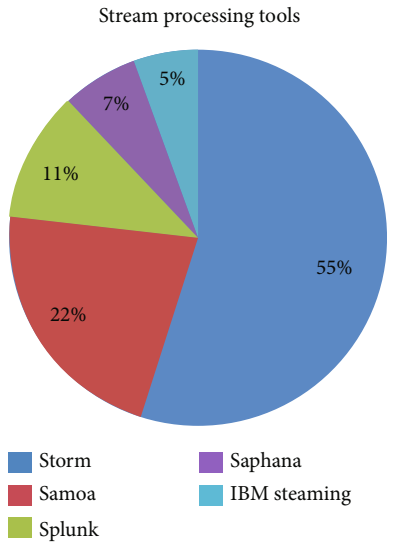

Figure 7: Stream processing tool statistics.

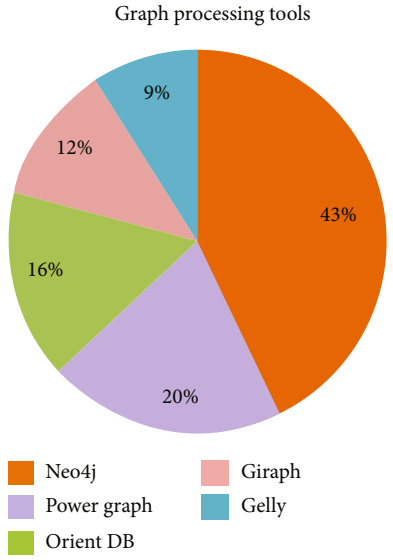

Figure 8: Graph processing tool statistics.

information included in those sources [88]. In social network applications, the shortest distances and pathways are also significant. They can be used to determine the network importance of a specific user. Closer users are more relevant than those distant away, as can be predicted. For graph processing tools analysis, considered the following tools: Neo4j, Orientdb, power graph, Gelly, Giraph, GraphLab, Hama, GraphX, and Pregel. Among this, Figure 8 gives a statistical view of popular tools in graph processing.

7.1.5. Hybrid Processing. Hybrid process systems incorporate both batch and stream processing mechanisms, which incorporate processing units that are functioning at their optimum efficiency to execute one or more process jobs. Combining the APIs and related components, it facilitates multiple data processing procedures [89]. Due to its sustainable and innovative processing choices, it has become a choice for many complex operations.

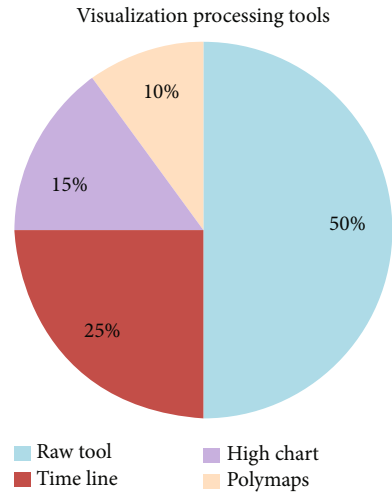

Figure 9: Visualization processing tool statistics.

7.1.6. Visualization Processing. Data visualization is one of the essential instruments for determining a qualitative understanding. It can be used to convey and highlight the key relationships in layouts and charts that are more useful to stakeholders to measure the correlation or significance [46, 54, 78, 90]. This might be useful when examining a dataset and retrieving features about it, and spotting patterns, corrupt data, outliers, and other things [38]. For analysis of visualization, processing tools considered the following: D3, raw graphs, Google charts, modest maps, open heat map, color brewer, datawrappers, digraphs, chartjs, charted, infogram, chart blocks, polymaps, and ember charts, and among this, Figure 9 presents the popular tools for visualization processing.

The above pie chart resembles popular tools in storage, batch, stream, visualization, and graph processing that are used in various application areas such as health, finance, social communications, business, and industrial areas for analysis. It also considers different literature papers and Google Trends data. The analysis in Figure 10 will be helpful in selecting the appropriate tool for a specific sector.

8. Future Directions

The results of this methodical study address various implications for the researchers who consider further exploring the consequence of big data technology and tools. In addition, this study will also contribute comprehensive information to practitioners who are involved in developing and using big data analysis and techniques to improve the quality of their services. It enables researchers to identify areas where further research efforts are needed. It is evident that even though several research works deal with the development and evaluation of advanced analytical techniques, there is still a dearth of information on the implementation of data analytics. By combining data analytics with machine learning, world can obtain more productive results. Big data tools

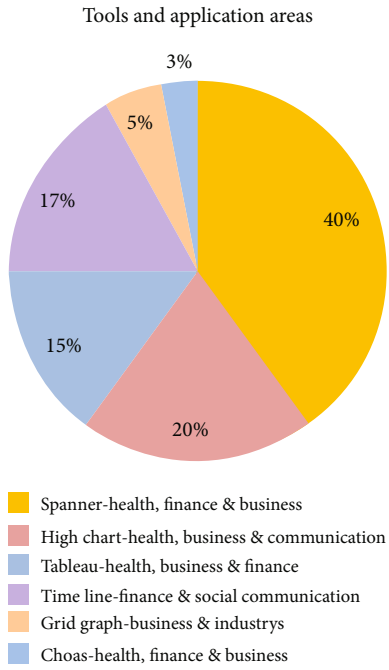

Figure 10: Popular tool and application areas.

have also started adopting new technologies to facilitate the users, but it is still required new tools to face new challenges arising from various application fields.

9. Conclusion

Technology has ushered in a new era of progress over the past two decades. This research study conceptualized the big data importance analysis of both structured and unstructured information using various tools. It is regarded as the foundation for all decision-making responsibilities and has become an essential component in the majority of company operations. The continuously growing nature of user's day-to-day activities in various sectors brings new concerns to the world. So always, data scientists can combine data with numerous new emerging methodologies to solve users' problems [91]. There is a need to build new standards to make better information systems. V's and 6S aid in becoming aware of big data features and APIs that allow clients to make potential strategies for adaptable outcomes. To properly manage massive data, working through parallel processing gives effective results. In the future, there may be scope for enhancing big data V's as there are more than 100 V's, as well as researchers need to do more explorations in this field. In the case of tool analysis, nearly 50 popular tools were considered. In the subsequent works, we would like to extend it with more tools along with PPML concepts.

Data Availability

The data supporting this systematic review are from previously reported studies, which have been cited. The raw data supporting the conclusions of this article will be made available by the author, without undue reservation.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Conceptualization was worked on by S.D. and R.K. Data curation and formal analysis were conducted by S.D. and R.K. The investigation and methodology were performed by R.K. Project administration was done by S.D. and R.K. Resources were obtained by S.D. and R.K. Supervision was conducted by R.K. Validation was accomplished by R.K. The visualization was completed by S.D. and R.K. Writing—review and editing—were done by S.D. and R.K. The authors have read and agreed to the published version of the manuscript.

Acknowledgments

I would like to convey my humble gratitude to Dr. Rajesh Kaluri my research supervisor for his effective guidance and motivation in every step of this research work.

References

[1] H. Margetts and C. Dorobantu, "Rethink government with AI," *Nature*, vol. 568, no. 7751, pp. 163–165, 2019.

[2] P. Beri and S. Ojha, "Comparative analysis of big data management for social networking sites," in *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, pp. 1196–1200, New Delhi, India, 2016.

[3] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. Mephu Nguifo, "An experimental survey on big data frameworks," *Future Generation Computer Systems*, vol. 86, pp. 546–564, 2018.

[4] I. Lee, "Big data: dimensions, evolution, impacts, and challenges," *Business Horizons*, vol. 60, no. 3, pp. 293–303, 2017.

[5] 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.

[6] M. Larch, J. Wanner, Y. V. Yotov, and T. Zylkin, "Currency unions and trade: a PPML re-assessment with high-dimensional fixed effects," *Oxford Bulletin of Economics and Statistics*, vol. 81, no. 3, pp. 487–510, 2019.

[7] N. Bhandari and P. Pahwa, "Comparative analysis of privacy-preserving data mining techniques," in *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018*, vol. 56, pp. 535–541, Singapore, 2019.

[8] J. Zhou, Z. Cao, X. Dong, and X. Lin, "PPDM: a privacy-preserving protocol for cloud-assisted e-healthcare systems," *IEEE Journal on Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1332–1344, 2015.

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

[10] D. Arunachalam, N. Kumar, and J. P. Kawalek, "Understanding big data analytics capabilities in supply chain management: unravelling the issues, challenges and implications for

practice," *Transportation Research Part E: Logistics and Transportation Review*, vol. 114, pp. 416–436, 2018.

[11] R. A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: a survey," *International Journal of Information Management*, vol. 45, pp. 289–307, 2019.

[12] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018.

[13] A. Gandomi and M. Haider, "Beyond the hype: big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.

[14] A. K. Bhadani and D. Jothimani, "Big data: challenges, opportunities and realities," *Effective big data management and opportunities for implementation*, pp. 1–24, 2017, https://www.igi-global.com/chapter/big-data/157681.

[15] J. Singh and V. Singla, "Big data: tools and technologies in big data," *International Journal of Computer Applications*, vol. 112, no. 15, pp. 975–8887, 2015.

[16] F. L. F. Almeida, "Benefits, challenges and tools of big data management," *Journal of Systems Integration*, vol. 8, no. 4, pp. 12–20, 2017.

[17] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, and A. Hussain, "Big data and IoT-based applications in smart environments: a systematic review," *Computer Science Review*, vol. 39, article 100318, 2021.

[18] S. Ren, Y. Zhang, Y. Liu, T. Sakao, D. Huisingh, and C. M. V. B. Almeida, "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: a framework, challenges and future research directions," *Journal of Cleaner Production*, vol. 210, pp. 1343–1365, 2019.

[19] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[20] S. Kumar and M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools," *Big data mining and analytics*, vol. 2, no. 1, pp. 48–57, 2019.

[21] E. Park, Y. Jang, J. Kim, N. J. Jeong, K. Bae, and A. P. del Pobil, "Determinants of customer satisfaction with airline services: an analysis of customer feedback big data," *Journal of Retailing and Consumer Services*, vol. 51, pp. 186–190, 2019.

[22] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146–157, 2018.

[23] S. Ketu, P. K. Mishra, and S. Agarwal, "Performance analysis of distributed computing frameworks for big data analytics: Hadoop vs Spark," *Computación y Sistemas*, vol. 24, no. 2, pp. 669–686, 2020.

[24] M. K. Saggi and S. Jain, "A survey towards an integration of big data analytics to big insights for value-creation," *Information Processing and Management*, vol. 54, no. 5, pp. 758–790, 2018.

[25] C. A. Ardagna, P. Ceravolo, and E. Damiani, "Big data analytics as-a-service: Issues and challenges," in *In 2016 IEEE international conference on big data*, pp. 3638–3644, Washington, DC, USA, 2016.

[26] V. Marx, "The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[27] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of big data*, vol. 2, no. 1, pp. 1–21, 2015.

[28] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[29] N. Hasliza, M. Hassan, K. Ahmad, and H. Salehuddin, "Diagnosing the issues and challenges in data integration implementation in public sector," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 529–535, 2020.

[30] 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.

[31] A. Mohamed, M. K. Najafabadi, Y. B. Wah, E. A. K. Zaman, and R. Maskat, "The state of the art and taxonomy of big data analytics: view from new big data framework," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 989–1037, 2020.

[32] J. Xie, Z. Song, Y. Li et al., "A survey on machine learning-based Mobile big data analysis: challenges and applications," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 8738613, 19 pages, 2018.

[33] M. L. Song, R. Fisher, J. L. Wang, and L. B. Cui, "Environmental performance evaluation with big data: theories and methods," *Annals of Operations Research*, vol. 270, no. 1, pp. 459–472, 2018.

[34] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data*, vol. 6, no. 1, pp. 1–6, 2019.

[35] U. Islam, A. Muhammad, R. Mansoor et al., "Detection of distributed denial of service (DDoS) attacks in IOT based monitoring system of banking sector using machine learning models," *Sustainability*, vol. 14, no. 14, p. 8374, 2022.

[36] A. A. Najar and S. Manohar Naik, "DDoS attack detection using MLP and random forest algorithms," *International Journal of Information Technology*, vol. 14, no. 5, pp. 2317–2327, 2022.

[37] S. Ananthu, N. Sethumadhavan, and H. Narayanan Ag, "Credit card fraud detection using Apache Spark analysis," in *2021 5th international conference on trends in electronics and informatics (ICOEI)*, pp. 998–1002, Tirunelveli, India, 2021.

[38] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: a survey," *Journal of Internet Technology*, vol. 21, no. 2, pp. 393–405, 2020.

[39] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 85–101, 2020.

[40] N. J. Shoumy, L. M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, vol. 149, article 102447, 2020.

[41] D. Kshirsagar and S. Kumar, "An efficient feature reduction method for the detection of DoS attack," *ICT Express*, vol. 7, no. 3, pp. 371–375, 2021.

[42] T. R. Rao, P. Mitra, R. Bhatt, and A. Goswami, "The big data system, components, tools, and technologies: a survey," *Knowledge and Information Systems*, vol. 60, no. 3, pp. 1165–1245, 2019.

[43] N. Bindra and M. Sood, "Detecting DDoS attacks using machine learning techniques and contemporary intrusion detection dataset," *Automatic Control and Computer Sciences*, vol. 53, no. 5, pp. 419–428, 2019.

[44] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–30, 2019.

[45] L. Chen, Y. Zhang, Q. Zhao, G. Geng, and Z. Yan, "Detection of dns ddos attacks with random forest algorithm on spark," *Procedia computer science*, vol. 134, pp. 310–315, 2018.

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

[47] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data*, vol. 2, no. 1, pp. 1–36, 2015.

[48] C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," *Computer science review*, vol. 17, pp. 70–81, 2015.

[49] N. R. Vajjhala and E. Ramollari, "Big data using cloud computing-opportunities for small and medium-sized enterprises," *European Journal of Economics and Business Studies*, vol. 4, no. 1, p. 129, 2016.

[50] D. Zhang, "Inconsistencies in big data," in *IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, pp. 61–67, New York, NY, USA, 2013.

[51] P. Struijs, B. Braaksma, and P. J. H. Daas, "Official statistics and big data," *Big Data & Society*, vol. 1, no. 1, article 205395171453841, 2014.

[52] R. Sahal, J. G. Breslin, and M. I. Ali, "Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case," *Journal of Manufacturing Systems*, vol. 54, pp. 138–151, 2020.

[53] O. M. Araz, T. M. Choi, D. L. Olson, and F. S. Salman, "Role of analytics for operational risk management in the era of big data," *Decision Sciences*, vol. 51, no. 6, pp. 1320–1346, 2020.

[54] R. H. Hamilton and W. A. Sodeman, "The questions we ask: opportunities and challenges for using big data analytics to strategically manage human capital resources," *Business Horizons*, vol. 63, no. 1, pp. 85–95, 2020.

[55] L. Hong, M. Luo, R. Wang, P. Lu, W. Lu, and L. Lu, "Big data in health care: applications and challenges," *Data and information management*, vol. 2, no. 3, pp. 175–197, 2018.

[56] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: a systematic review," *International Journal of Medical Informatics*, vol. 114, pp. 57–65, 2018.

[57] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, 2019.

[58] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data - evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, pp. 63–71, 2019.

[59] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.

[60] M. Anshari, M. N. Almunawar, S. A. Lim, and A. Al-Mudimigh, "Customer relationship management and big data enabled: personalization & customization of services," *Applied Computing and Informatics*, vol. 15, no. 2, pp. 94–101, 2019.

[61] S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, and S. J. Childe, "How to improve firm performance using big data analytics capability and business strategy alignment?," *International Journal of Production Economics*, vol. 182, pp. 113–131, 2016.

[62] S. F. Wamba, A. Gunasekaran, S. Akter, S. J. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: effects of dynamic capabilities," *Journal of Business Research*, vol. 70, pp. 356–365, 2017.

[63] M. Ge, H. Bangui, and B. Buhnova, "Big data for internet of things: a survey," *Future Generation Computer Systems*, vol. 87, pp. 601–614, 2018.

[64] O. Müller, M. Fay, and J. Vom Brocke, "The effect of big data and analytics on firm performance: an econometric analysis considering industry characteristics," *Journal of Management Information Systems*, vol. 35, no. 2, pp. 488–509, 2018.

[65] D. S. Terzi, R. Terzi, and S. Sagiroglu, "A survey on security and privacy issues in big data," in *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pp. 202–207, London, UK, 2016.

[66] P. Pääkkönen and D. Pakkala, "Reference architecture and classification of technologies, products and Services for big data systems," *Big data research*, vol. 2, no. 4, pp. 166–186, 2015.

[67] P. Tabesh, E. Mousavidin, and S. Hasani, "Implementing big data strategies: a managerial perspective," *Business Horizons*, vol. 62, no. 3, pp. 347–358, 2019.

[68] Y. Sun, Y. Shi, and Z. Zhang, "Finance big data: management, analysis, and applications," *International Journal of Electronic Commerce*, vol. 23, no. 1, pp. 9–11, 2019.

[69] V. Grover, R. H. L. Chiang, T. P. Liang, and D. Zhang, "Creating strategic business value from big data analytics: a research framework," *Journal of management information systems*, vol. 35, no. 2, pp. 388–423, 2018.

[70] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big data research*, vol. 2, no. 2, pp. 59–64, 2015.

[71] T. S. Ing, T. C. Lee, S. W. Chan, J. Alipal, and N. A. Hamid, "An overview of the rising challenges in implementing industry 4.0," *International Journal of Supply Chain Management*, vol. 8, no. 6, pp. 1181–1188, 2019.

[72] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbour, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, article 100071, 2022.

[73] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big data technologies: asurvey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[74] S. Boubiche, D. E. Boubiche, A. Bilami, and H. Toral-Cruz, "Big data challenges and data aggregation strategies in wireless sensor networks," *IEEE Access*, vol. 6, pp. 20558–20571, 2018.

[75] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in *2013 Sixth international conference on contemporary computing (IC3)*, pp. 404–409, Noida, India, 2013.

[76] J. L. Torrecilla and J. Romo, "Data learning from big data," *Statistics & Probability Letters*, vol. 136, pp. 15–19, 2018.

[77] M. Bansal, I. Chana, and S. Clarke, "A survey on IoT big data," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1–59, 2021.

[78] N. Khan, A. Naim, M. R. Hussain, Q. N. Naveed, N. Ahmad, and S. Qamar, "The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges," in *Proceedings of the international conference on omni-layer intelligent systems*, pp. 19–24, Heraklion, Crete, Greece, 2019.

[79] G. George, M. R. Haas, and A. Pentland, "Big data and management," *Academy of management Journal*, vol. 57, no. 2, pp. 321–326, 2014.

[80] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[81] X. Zhao, J. Zhang, and X. Qin, "k NN-DP: handling data skewness in kNN joins using MapReduce," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 3, pp. 600–613, 2018.

[82] M. Zaharia, R. S. Xin, P. Wendell et al., "Apache Spark," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[83] H. Yan, D. Sun, S. Gao, and Z. Zhou, "Performance analysis of storm in a real-world big data stream computing environment," in *Collaborative Computing: Networking, Applications and Worksharing: 13th International Conference, CollaborateCom 2017*, pp. 624–634, Springer International Publishing, Edinburgh, UK, 2018.

[84] L. Abualigah and B. Al Masri, "Advances in MapReduce big data processing: platform, tools, and algorithms," *Artificial Intelligence and IoT*, vol. 85, pp. 105–128, 2021.

[85] N. Tantalaki, S. Souravlas, and M. Roumeliotis, "A review on big data real-time stream processing and its scheduling techniques," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 35, no. 5, pp. 571–601, 2020.

[86] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, "Streaming feature selection algorithms for big data: a survey," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 113–135, 2022.

[87] S. Heidari, Y. Simmhan, R. N. Calheiros, and R. Buyya, "Scalable graph processing frameworks," *ACM Computing Surveys*, vol. 51, no. 3, p. 60, 2019.

[88] L. Belcastro, F. Marozzo, and D. Talia, "Programming models and systems for big data analysis," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 34, no. 6, pp. 632–652, 2019.

[89] D. Cheng, X. Zhou, Y. Wang, and C. Jiang, "Adaptive scheduling parallel jobs with dynamic batching in spark streaming," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 12, pp. 2672–2685, 2018.

[90] P. Kulkarni, M. Awwad, R. Bapna, and A. Marathe, "Big data analytics in supply chain: a literature review," in *Proceedings of the international conference on industrial engineering and operations management*, vol. 2018, pp. 418–425, Washington DC, USA, September 2018.

[91] S. Madden, "From databases to big data," *IEEE Internet Computing*, vol. 16, no. 3, pp. 4–6, 2012.