A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques

MOHAMMAD SHAHNAWAZ, Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, India
MANISH KUMAR, Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, India

Modern computing devices generate vast amounts of diverse data. It means that a fast transition through various computing devices leads to big data production. Big data with high velocity, volume, and variety presents challenges like data inconsistency, scalability, real-time analysis, and tool selection. Although numerous solutions have been proposed for big data processing, they are often limited in scope and effectiveness. This survey aims to address the lack of comprehensive analysis of big data challenges in relation to machine learning (ML) and the Internet of Things (IoT) environments, particularly concerning the 7Vs of big data. It emphasizes the significance of selecting suitable tools to address each unique big data characteristic, providing a structured approach to manage these challenges effectively. The article systematically reviews big data characteristics and associated techniques, with a detailed discussion of various tools and their applications. Additionally, it analyzes existing ML methods and techniques for IoT data analytics in big data contexts. Through a systematic literature review (SLR), we examine key aspects, including core concepts, benefits, limitations, and the impact of big data on ML algorithms and IoT data analytics. We highlight groundbreaking studies addressing big data challenges to impact future research and enhance big data-driven applications.

CCS Concepts: • **Survey and Reviews → Big Data**; • **Big Data and Analytics → Characteristics and Processing Techniques**; • **Big Data Technological Impact → Machine Learning Algorithms and Internet of Things (IoT);**

Additional Key Words and Phrases: Big data, analytics, tools, techniques, survey

ACM Reference Format:
Mohammad Shahnawaz and Manish Kumar. 2025. A Comprehensive Survey on Big Data Analytics: Characteristics, Tools and Techniques. *ACM Comput. Surv.* 57, 8, Article 196 (March 2025), 33 pages. https://doi.org/10.1145/3718364

1 Introduction

The popularity of mobile and sensor devices increases big data production. The evolution of big data analytics (BDA) acknowledged the big data generation that began in the 1990s. In an International Data Corporation (IDC) prediction, by 2025, 163 zettabytes of digital data will be created

Manish Kumar IEEE Senior Member.

This work was supported under the junior research fellowship (JRF) program managed by the University Grant Commission (UGC), Government of India. NTA Ref. No.: 210510235958.

Authors' Contact Information: Mohammad Shahnawaz, Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh, India; e-mail: mshahnawazn@gmail.com; Manish Kumar, Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh, India; e-mail: manish@iiita.ac.in.

ACM 0360-0300/2025/03-ART196
https://doi.org/10.1145/3718364

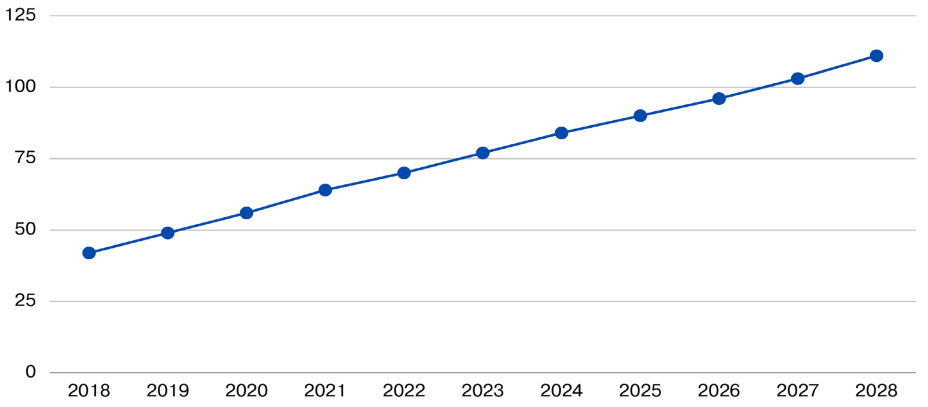

Fig. 1. Big data market growth (Billion USD).

globally [163]. Relational databases and traditional data warehouses were made to handle structured data. Structured data can be easily arranged and stored in tables with defined schemas. The recent data creation blends unstructured and partially structured data such as textual content, images, and videos. Therefore, it demands the innovation of more powerful technologies to handle. Yahoo developed Hadoop as an Apache open-source project in 2006 [124]. This distributed processing platform enables running big data applications on a clustered platform. Initially, professional institutions having on-site big data platforms carried out BDA. Google and Facebook are the ones. Organizations with limited infrastructure, however, require some form of technology to use BDA. The emergence of cloud computing services met this demand. The cloud can manage Hadoop clusters. So, cloud services have opened up BDA for organizations of any scale. The emergence of powerful technologies made extracting valuable insights from daily data easy. Hence, the big data market proliferates. BDA is important because it can scrutinize and derive insights from extensive, intricate data sets [123]. BDA can be used in healthcare for disease pattern identification, drug discovery, and personalized medicine [25]. BDA is crucial for fraud detection, risk management, and financial market analysis [42]. In marketing, BDA can analyze customer data and optimize marketing campaigns. BDA can be used in manufacturing to analyze production data and improve efficiency. BDA is essential in optimizing resource management and improving energy efficiency. BDA can be used to analyze traffic patterns and optimize transportation routes [75]. BDA can be used in the public sector for crime prevention, emergency response, and public health [147]. BDA can be used in retail for sales data analysis, inventory management, and supply chain optimization [66].

The BDA market is anticipated to expand across various industries in the coming years. Figure 1 shows the rapid adoption of BDA from 2018 to 2028 [152]. Considering the merits of big data, it yields novel prospects for upcoming researchers in knowledge discovery tasks.

However, opportunities always follow some challenges. Thus, a comprehensive data analytics ecosystem is required for today's agility in effective knowledge discovery [146]. Although BDA is extremely popular these days, a comprehensive study needs to be made available to compile and assess the current tools and approaches. Therefore, the primary objective of our research is to assess solutions for BDA. We have thoroughly gathered, organized, and reviewed the available techniques to achieve this. The following tasks make up the majority of our research:

— Proposing a thorough analysis of the current BDA problems.
— Proposing a technology categorization using current literature via established methodologies.
— Examining the benefits, drawbacks, and various evaluation methods and tools employed in current research.

— Considering the complexities of BDA present in big data and its technological advancements.
— Examining the direction of future research and the significance of BDA in multiple areas.

The scope of this study focuses on a comprehensive analysis of BDA, addressing key concepts, challenges, and emerging technologies. It covers the characteristics of big data, differentiating between datasets that qualify as big data and exploring data analysis techniques, processing tools, algorithms, databases, and big data frameworks.

This survey also highlights the "7 Vs" of big data and their associated research challenges, particularly in relation to ML and IoT-driven data analytics. Through a systematic literature review, it examines state-of-the-art technologies and methodologies, identifies research gaps, and evaluates the integration of ML and IoT for BDA. Lastly, future research directions are suggested to advance the field.

The subsequent sections of the article are organized in the following manner: Section 2 discusses the foundation of writing this article in terms of big data, its analytics, the processing technologies for big data, and model development. Relevant reviews have been discussed and compared with this article in Section 3. Section 4 encompasses the research methodology and tools employed for article selection. Section 5 investigates the persistent research challenges in BDA. Section 6 provides the State-of-the-Art discussions on BDA. Section 7 elaborates on the future scope, and the research concludes in Section 8.

2 Fundamentals of Big Data and Analytics

2.1 Characteristic of Big Data

Big data refers to vast, complex, diverse datasets that challenge conventional processing and analysis methods. Big data grows continuously in quantity and variety and originates from sources like social media, IoT devices, transactional records, machine data, geospatial information, and public datasets [88]. Initially defined by Gartner's "3 V's" [58]—volume, velocity, and variety—recently expanded and include up to 7 V's [65], with Veracity, Value, Visualization, and Variability. Although there have been claims of up to 51 V's [94], the characteristics of big data are often summarized by the original 3 V's, supplemented by the additional 4 V's, to provide a robust framework for understanding and addressing big data.

Volume: Volume refers to the vast amount of data generated and collected, which exceeds traditional systems' storage and processing capabilities [109]. It is typically measured in terabytes or petabytes and includes data from diverse sources such as social media, sensors, transactions, and electronic health records (EHRs) [16, 20, 133, 164].

Velocity: Velocity denotes the speed at which data is generated and processed [9]. Rapid data flow from digital technologies enables real-time analysis for applications like social media trends, predictive maintenance, high-frequency trading, traffic management, and fraud detection [49, 148, 149, 164].

Variety: Variety describes the range of data types and sources, including structured, unstructured, and semi-structured data [43]. This diversity creates challenges for processing but also enables richer insights, better decision-making, and personalized recommendations [16, 31, 165].

However, the original 3 V's are sufficient for defining big data. We welcome proposals for additional Vs but not as determining properties of big data. We can accept the other 4 Vs as features relevant to specific datasets and may be valuable for particular contexts of the big data study.

Veracity: Veracity in big data ensures that data is accurate and reliable. It involves checking where the data originates and whether it is precise. Since big data can be messy and contain errors, verifying its quality is essential for making accurate predictions and informed decisions [40, 63, 162].

Value: The value of big data lies in its ability to provide insights and advantages for decision-making and trend discovery. This value is variable, depending on the available data and hidden information. The challenge is to identify and extract this value for meaningful analysis [7].

Visualization: Visualization in big data involves creating readable and understandable representations of large datasets [150]. Visualizing large datasets requires developing tools that perform well in terms of functions, scalability, and response time [33]. Visualization is incredibly challenging with real-time data due to the vast amounts and high velocity.

Variability: Variability in big data refers to changes over time, including changes in data structure, format, and frequency. These changes can complicate data parsing and processing. Tracking changes and adapting advanced data processing tools are essential to manage variability [162].

We can understand what big data is with the help of these characteristics. Here are some specific examples of datasets that qualify as big data and which does not qualify.

2.1.1 Examples of Dataset That Qualifies as "Big Data".

(1) Healthcare Data: EHRs, medical imaging, and genomic data (variety) are substantial in volume. They vary significantly in format and are processed rapidly to provide real-time insights and support decision-making (value) [25, 105, 139, 177].
(2) Social Media Data: Platforms like X (formerly known as Twitter), Facebook, and Instagram generate enormous amounts (volume) of real-time (velocity) data. Social media data includes user posts, comments, likes, shares, and multimedia content, which vary widely in format and content (variety) [16, 26, 80, 167].
(3) IoT Data: IoT devices, such as smart thermostats, wearables, and industrial sensors, produce continuous data streams (volume and velocity). The variety includes numerical readings, categorical data, and time-series data, all generated at high velocity [21, 68, 82, 121, 143].
(4) Financial Transactions: Stock markets and financial institutions handle vast transactions (volume) every second (velocity). This data, including trade details, market prices, and economic news, is highly varied (variety) and generated quickly [42, 148].
(5) E-commerce Data: Online retail platforms collect data on customer behavior, transactions, product reviews, and inventory levels (variety). This data comes from various sources and formats, changing rapidly (variability) with consumer trends and interactions [30, 66].
(6) Telecommunications Data: Mobile network operators handle many call records, text messages, and data usage logs. The data is varied and generated continuously (velocity), reflecting usage patterns (visualization) and network performance [57, 142].

These examples illustrate how large datasets can also exhibit the characteristics of big data, making them challenging and valuable for analysis.

2.1.2 Examples of Dataset Which Does Not Qualify as "Big Data" [47, 65, 146, 175].

(1) Large Static Relational Database: Consider a traditional relational database teeming with structured data. This repository may house vast amounts of information. Still, if it remains relatively static, receiving infrequent updates and handling similar data types, it doesn't satisfy the nature or characteristics of Big Data.
(2) Archived Log Files: Consider the vast archives of log files collected over time. These records are significant in volume, yet they need more immediacy of real-time processing and primarily consist of structured data. They remain static, waiting to be accessed rather than actively processed.
(3) Historical Weather Data: Envision a comprehensive collection of historical weather records. While this dataset may be extensive, it remains static, confined to past information without

real-time updates or diverse data formats. Its unchanging nature distinguishes it from Big Data.

(4) Library Catalog: Library catalog may list countless books and resources, but it is structured and does not experience high-velocity updates or encompass many data types. It remains static, with its data growing incrementally.
(5) Corporate Financial Data: It's a corporation's extensive database of financial transactions accumulated over the years. Although substantial, this dataset is primarily structured, updated at regular intervals, and lacks the real-time processing and variety of Big Data.
(6) Archived Scientific Research Data: This is a database brimming with data from scientific experiments. This dataset, while voluminous, is static and structured, stored for archival purposes rather than processed in real-time.

These examples underscore that simply having a large amount of data does not qualify it as big data. True big data involves vast volumes, high velocity, variety, variability and veracity, and potential for significant value.

Organizations can gain a deeper understanding of their business by leveraging the characteristics of big data. So, BDA will lead to better decision-making, improved operational efficiency, and a more competitive position in the market.

2.2 Data Analysis

The outcome of the previous subsection promises the effectiveness of getting into the big data. So the next step will be "The Analysis". Analysis of big data can be done through various techniques tailored to specific needs, such as descriptive analytics [92], which provides an overview of what has happened; diagnostic analytics [76], which delves into why it happened; predictive analytics [51], which forecasts future trends; prescriptive analytics [57] offers recommendations for future actions and cognitive analytics mimics human thought processes to interpret complex data [49]. These fundamental types of analysis, often used together, form the backbone of BDA.

In each type of analysis, normal (small) and big data are processed differently due to their distinct characteristics. Basic tools such as structured queries are utilized to summarize datasets for descriptive analysis of normal data. In contrast, advanced techniques and distributed computing platforms, such as Hadoop and Spark, process vast amounts of information [89, 136]. Diagnostic analysis of normal data requires traditional statistical methods like regression and correlation, while ML algorithms uncover deeper insights and patterns of big data [138]. Predictive analysis transitions from using simple algorithms, such as linear regression on limited data, to employing complex ML models like neural networks for more accurate forecasts in big data [89]. Prescriptive analysis evolves from basic optimization techniques like linear programming to sophisticated algorithms for big data analysis incorporating ML and simulation models to provide actionable recommendations [129]. Cognitive analysis, which aims to mimic human thought processes, moves from standard techniques to natural language processing (NLP) and deep learning methods to interpret and derive meaning from large, complex datasets [161].

2.3 Big Data Processing

Big data processing requires specialized techniques beyond conventional infrastructures and methodologies. We have categorized these techniques into algorithms, tools, and databases, providing a clear structure for understanding the different solutions available for BDA.

Figure 2 illustrates a comprehensive model that outlines the key components and processes involved in analyzing large datasets. Initial steps include critical data preprocessing operations, which prepare the data for analysis. Following this, appropriate processing tools such as Hadoop and Spark are selected to perform analysis and achieve the desired goal.

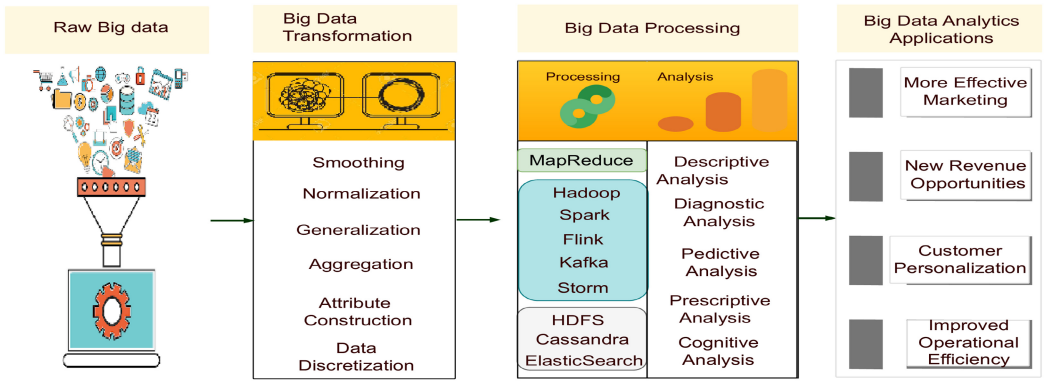

Fig. 2. Conceptual model of BDA.

2.3.1 *Algorithms.*

— **MapReduce**: MapReduce is Google's algorithm for handling large volumes of data by splitting processing into map and reduce tasks. It operates within the Hadoop framework and enables distributed, parallel data processing [39].

2.3.2 *Tools.*

— **Hadoop**: Hadoop is an open-source big data framework for distributed storage and batch processing, primarily using HDFS and MapReduce [6, 168].
— **Apache Spark**: Spark is a fast, flexible in-memory data processing engine designed for large-scale analytics [151].
— **Apache Kafka**: Kafka is a distributed platform for real-time data ingestion and streaming, commonly used with Flink and Storm for real-time data processing [91].
— **Storm**: Storm is a fault-tolerant, real-time processing system often paired with Kafka and Cassandra for fast diagnostic analytics [50].
— **Flink**: Flink is a powerful tool for efficient streaming and batch data processing, optimized for network and storage efficiency [117].
— **Hive and Pig**: Hive and Pig are data processing frameworks built for Hadoop. Hive is used for SQL-based querying of large datasets, while Pig is used for data transformation and processing [15, 156].
— **Mahout**: Mahout is an open-source tool for scalable ML, offering algorithms for clustering, classification, and regression using Hadoop [50].
— **Presto**: Presto is a high-performance distributed SQL query engine, suitable for ad-hoc queries, but lacks built-in security or transaction features [163].

2.3.3 *Databases.*

— **Hadoop Distributed File System (HDFS)**: HDFS is the distributed storage system used in Hadoop, designed for managing large datasets in a scalable manner.
— **Cassandra**: Cassandra is a column-oriented NoSQL database optimized for real-time processing, high write throughput, and low latency [14].
— **ElasticSearch**: ElasticSearch is a scalable search and analytics engine designed for unstructured data, offering fast query capabilities, though with some limitations in data consistency [163].

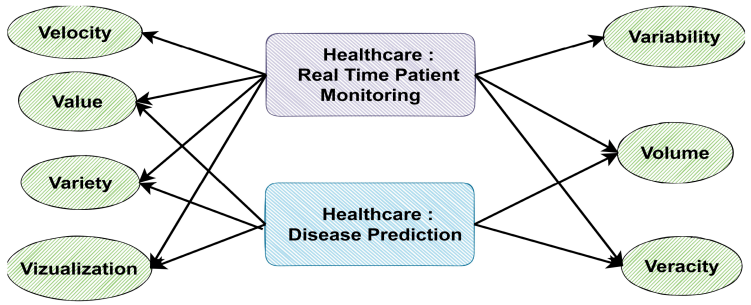

Fig. 3. Interaction of big data characteristics in real-world healthcare analytics scenario.

It's crucial to comprehend how big data characteristics interact in a real-world setting and how these tools can process them to achieve the desired outcome. We can use real-world analytics scenarios to understand this. We are taking a healthcare analytics scenario here as an example. In a healthcare analytics scenario, there are interactions of big data characteristics like volume, velocity, variety, veracity, and value, as shown in Figure 3. Hospitals handle massive amounts of data from sources like EHRs, medical imaging, and wearable devices [25, 35]. Tools like HDFS store this diverse data, while ElasticSearch helps quickly search and retrieve it [98]. Real-time patient monitoring generates high-speed data streams that need to be accurate and reliable, handled by Kafka for ingestion and Spark Streaming for real-time processing [177]. Data cleaning tools like Hive and Pig ensure accuracy, while Spark and Mahout run predictive analytics to improve patient outcomes [147].

Furthermore, addressing individual big data characteristics or the interplay of these characteristics requires specific tools [50]. Figure 4 shows one-to-many (1:M) relations between seven characteristics of big data with their respective big data processing tools.

Big data tools are essential in managing different characteristics of big data. The optimal tool selection depends on the nature of the task, desired performance metrics, scalability, and operational complexity. Table 1 presents a detailed explanation of the tools' strengths, weaknesses, best use cases, and when one tool can be used over another.

2.4 Big Data Models or Frameworks

Big data processing tools like Spark and Hadoop are software platforms used to execute data processing tasks, focusing on how data is processed (e.g., batch or real-time). In contrast, big data models or frameworks (batch processing framework, stream processing framework, etc.) define the overall approach or architecture. Big data processing frameworks provide structured methods to manage and analyze vast datasets, addressing the key characteristics of big data: volume, velocity, variety, veracity, value, variability, and visualization [3, 118, 166].

(1) Batch Processing Framework: Batch processing frameworks utilize Hadoop to handle large volumes of data by processing it in batches, making them ideal for offline analytics.
(2) Stream Processing Framework: Stream processing frameworks use *Apache Storm, Spark, and Kafka* to process real-time data streams, effectively managing high-velocity data.
(3) Hybrid Framework: Hybrid frameworks use *Apache Flink* to combine batch and stream processing, offering flexibility in handling historical and real-time data.
(4) Graph Processing Framework: Graph processing frameworks utilize *Neo4j or Amazon Neptune* to specialize in analyzing relationships and connections within data, addressing complexity and variety.

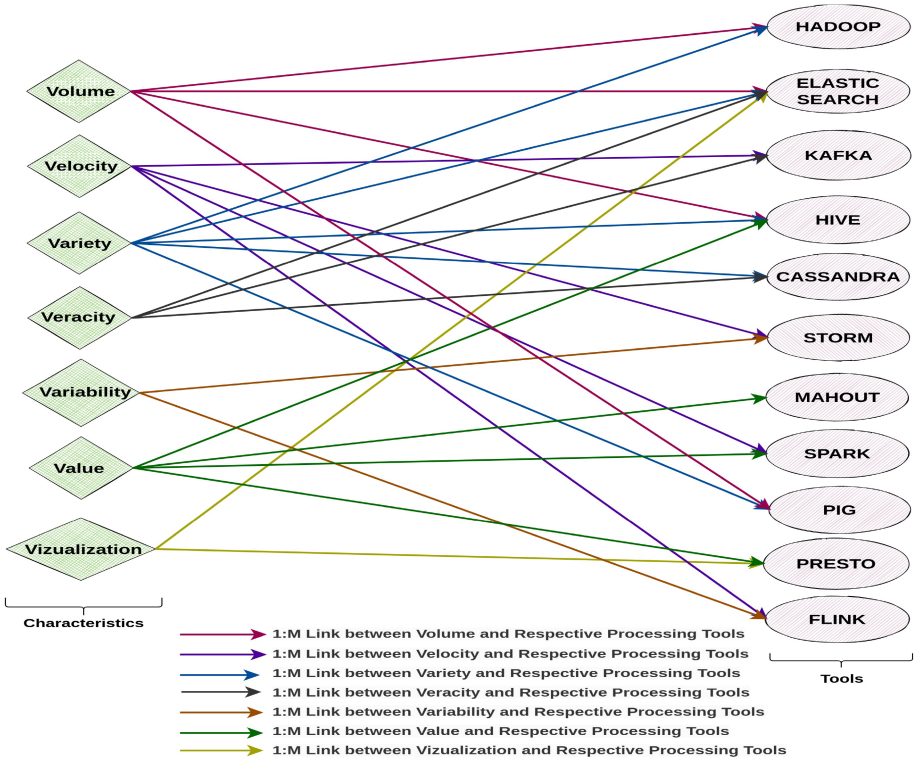

Fig. 4. Relational diagram for big data V's and corresponding tools.

(5) Interactive Analytical framework: Interactive analytical frameworks use *Apache Drill or Presto* to enable real-time, ad-hoc big data querying, facilitating quick insights, and visualization.

Table 2 presents a comparative analysis of big data development models. These frameworks often use different sets of big data processing tools to create robust platforms for managing and processing the complex interplay of big data characteristics.

3 Relevant Reviews

We discussed big data and analytics fundamentals in the previous section. In this section, the goal is to review related studies. Table 3 provides a comparative analysis of the related work with our research. In comparison to prior studies, our work stands out by offering a more comprehensive and technically detailed evaluation of BDA. Zhihan Lv et al. [110] provide a general overview of big data advances and Flasteen et al. [2] focus on framework comparisons, these works lack in-depth solutions and practical insights. Our work bridges this gap by offering practical guidance on tools and techniques, explicitly addressing challenges related to volume, velocity, variety, veracity, value, variability, and visualization.

Justin et al. [179] and Zaher et al. [8] have primarily focused on the intersection of big data and ML without evaluating real-world performance. Our study delves into how big data tools can effectively integrate with modern applications. Additionally, our study covers modern, scalable solutions that go beyond the traditional methodologies discussed by Shalini et al. [101]. The comparison with these related works positions our work as a more thorough and actionable resource

Table 1. Strengths, Weaknesses, and Use Cases of Big Data Tools

Tool	Strengths	Weaknesses	Use Case	When to Use Over Other
Hadoop (MapReduce)	- Excellent for batch processing and large-scale storage - Cost-effective for processing high volumes of data	- High-latency - Inefficient for real-time data processing	Batch-oriented data processing at scale	Use over Spark when cost is a priority and latency isn't a concern
Apache Spark	- In-memory processing - High-speed iterative tasks - Real-time and batch processing	- High memory requirements - More complex than Hadoop	Real-time analytics and iterative data processing	Use over Hadoop for speed-critical tasks
Apache Kafka	- High-throughput, durable streaming - Easy integration with other tools	- Lacks stateful stream processing - Limited in handling complex event processing	Real-time log ingestion and streaming	Use over Flink/Storm for high-throughput, simple event streaming
Apache Flink	- Advanced stateful stream processing - Event-time processing	- Higher setup complexity - Can be overkill for simple tasks	Low-latency, stateful stream processing	Use over Kafka/Storm for complex real-time data transformations
Apache Storm	- Low-latency stream processing - Simple to set up and deploy	- Not as efficient for stateful operations as Flink - Lacks windowing support	Real-time data processing with fault tolerance	Use over Kafka for simple real-time stream processing
Cassandra	- Low-latency reads/writes - High availability - Horizontal scalability	- Eventual consistency model (may sacrifice accuracy) - Complex data modeling	Distributed database for write-intensive applications	Use over HDFS for real-time reads and writes
HDFS	- Ideal for large, unstructured data storage - Scalable	- High latency - Not suitable for real-time operations	Storing vast amounts of unstructured data	Use over Cassandra for batch storage and when real-time access is not needed
Elasticsearch	- Full-text search - Real-time indexing - Scalable for logs	- Not ideal for structured analytics - Memory-intensive	Real-time search and log analytics	Use over Presto for full-text search or near real-time querying
Presto	- SQL-compliant distributed querying - High-speed queries across heterogeneous data sources	- Less efficient for real-time search or document-oriented storage	Interactive, ad-hoc querying	Use over Elasticsearch when you need SQL-based querying on distributed data
Hive	- SQL-based querying for large datasets - Runs on top of Hadoop	- High-latency - Not designed for real-time queries	Batch analytics on large datasets	Use over Spark when SQL-based querying on Hadoop is required
Pig	- Easy to use for complex data transformations - Works well with Hadoop	- Not as fast as other query engines - Limited to batch processing	Complex transformations on batch data	Use over Hive for non-SQL, batch-oriented transformations
Mahout	- ML libraries for Hadoop - Scalable for large datasets	- Limited to batch processing - Requires expertise in ML	Scalable ML on big data	Use over Spark MLlib for large-scale batch learning tasks

for handling complex big data environments, providing deeper insights and addressing a broader set of challenges than previous studies.

Our work comprehensively analyses the prevailing challenges in BDA, identifies key limitations, and explores potential solutions. To achieve this, we propose a structured categorization of big data technologies based on established methodologies and insights from current literature. Additionally, we critically examine various evaluation techniques, tools, and frameworks utilized in existing research, assessing their effectiveness, limitations, and applicability across different big data environments. Furthermore, we analyze the complexities inherent in BDA, considering both the evolving nature of big data characteristics and the impact of technological advancements in this domain. Finally, we explore future research directions, emphasizing the significance of BDA across multiple fields and its transformative potential in optimizing decision-making, automation, and intelligent data-driven applications.

4 Methodology of Research

This section employed the SLR approach to investigate strategies that address big data challenges in ML and IoT. The SLR involves a comprehensive examination of all research within a defined

Table 2. Comparative Analysis of Big Data Development Models

Comparative parameters	Big data development models				
	Batch	***Stream***	***Hybrid***	***Big Graph Processing***	***Interactive Analytics***
Data storage	HDFS	Real-time input stream	Memory/Disk	Distributed file system/ databases	In-memory databases
Processing speed	Minute	Second/ mili-second	Second	Minute	Second
Information propagation mechanism	HTTP	Message queue	Shared memory/ broadcast	Message queue	Queries
Data Characteristic	Large scale data	Real-time data	Batch and stream data	Relevant graph structure	Ad-hoc query
Processing tool used	HDFS	Apache Kafka, RabbitMQ, Amazon Kinesis	Cassandra /HBase	Neo4j/ Amazon Neptune	Apache Druid/ MemSQL
Scalability	Large scale processing	Horizontal scalability with increasing data velocity	Scalability depends on the architecture	Challenging to scale due to intricate relat-ionship analysis	Support scalable querying
Strengths	Historical analysis	Real-time insights	Flexibility and balance	Relationship analysis	Interactivity
Limitations	High latency	Data Complexity	Integration challenges	Computational complexity	Resource Intensive
Use Cases	Processing massive data offline (volume); Large scale searches for web information (value)	Full real-time analysis (velocity); Ongoing scheduling; Continuous computation	Repetitive machine learning; Ongoing incremental computation (volume, velocity)	Social network mapping (vizualization); Traffic road map analysis	Interactive sales dashboards (visualization); Marketing campaign analysis (value)
Key Publications	[67, 86, 166]	[44, 67, 96, 174, 177]	[23, 32, 67, 165]	[28, 41, 46, 77]	[4, 57]

scope, specifically focusing on challenges posed by big data characteristics. To ensure the validity of the research selection process, we also seek verification of the selection techniques used.

The following subsections detail the search methodology, including formulating Research Questions (RQ) and applying the selection criteria.

4.1 Question Formalization

— **RQ 1 :** What are the characteristics of a dataset that can be classified as Big Data in a real-world scenario?
This question is answered in Section 2.

— **RQ 2 :** What are the use cases in which a big data processing tool is selected over others?
This question is answered in Section 2.

— **RQ 3 :** What are the primary challenges posed by the characteristics of big data and state-of-the-art solutions on the topic?
This question is answered in Sections 5.1 and 6.

— **RQ 4 :** What are the current trends and anticipated research directions in solving the challenges faced while using traditional ML for processing Big Data?
This question is answered in Sections 6 and 7.

— **RQ 5 :** What role do emerging technologies play in overcoming big data challenges for IoT applications?
This question is answered in Section 6.

Table 3. Review of Related Work

Authors	Main Idea	Advantage	Work Limitations	Big Data Characteristics Covered
Zhihan Lv et al. [110] 2017	The work focuses on recent advances in big data from a view of data types, storage models, analysis models and applications.	Presents the state-of-the-art in BDA, discussing challenges like scalability and real-time processing.	Comprehensive in identifying future research topics but lacks depth in addressing specific solutions to current technical challenges.	Volume, Velocity, Variety, Veracity, Value
Flasteen et al. [2] 2019	A Comparative Study on BDA Frameworks, Data Resources, and Challenges	Compares big data frameworks like Hadoop and Spark, highlighting their capabilities and shortcomings	Focuses primarily on comparing frameworks, with limited exploration of practical implementations or real-world applications of these frameworks	Volume, Velocity, Variety
Justin et al. [179] 2021	The study examined the focus of current research on BDA and ML	Provides a retrospective analysis of big data and ML through bibliometric insights	Lacks practical examination of the performance and integration of ML with big data tools	Volume, Velocity, Variety
Zaher Ali et al. [8] 2022	Explore BDA Applications and Opportunities	Reviews applications of BDA in multiple industries, identifying opportunities and challenges	Lacks a detailed technical comparison or evaluation of specific BDA tools and methodologies	Volume, Velocity, Value
Fatima et al. [157] 2023	Provides an overview of Big Data's definition, significance and challenges, with a focus on its role in AI advancements	Highlights the potential of Big Data for extracting insights using ML and improving operational efficiency	Lacks empirical data demonstrating the impact of Big Data on ML applications in real-world scenarios	Volume, Velocity, Variety, Veracity
Shalini et al. [101] 2024	Review on Essential Techniques and Methodologies in Data Analysis	Highlights the advancement in big data processing and technologies to address the associated challenges	Focuses on traditional data analysis methodologies, providing little insight into modern BDA tools and frameworks	Volume, Variety
Leonidas et al. [159] 2024	Provides a review on big data management for strategic decision-making and innovation	Comprehensive analysis on real-world applications, challenges and emerging technologies	Lacks methodological details, potential case study and future trends	Volume, Velocity, Variety, Value
Keerthana and Sherly [53] 2025	Review on big data classification challenges and techniques	A thorough analysis of challenges in big data classification with constrained computational resources	Lacks real-world validation, explanation of dataset as big data, and thorough analytical discussion	Volume, Variety
Our Work	Provide a comprehensive review on Big Data Characteristics, Tools and Techniques	In-detail evaluation of the topic	- Lack of evaluation of non-English resources - Lack of using lecturer notes	Volume, Velocity, Variety, Veracity, Value, Variability and Visualization

— **RQ 6 :** What are the most important potential solutions and unanswered questions in this field?
Section 7 puts forward this question's answer.

4.2 Article Selection Process

The article selection and search procedure for this research consists of four stages. This procedure is depicted in Figure 5. Table 4 lists the search terms and keywords used at the initial stage. The articles were retrieved through standard queries from electronic databases, including Springer Link, ACM, Scopus, Elsevier, IEEE Xplore, Emerald Insight, Taylor and Francis, MDPI, and Hindawi. Stage 1 comprises a total of 512 articles.

Stage 2 involves two processes to determine the total number of articles for review. Initially, 225 articles were selected based on the criteria outlined in Figure 6. This research focused on exploring specific solutions for BDA mechanisms. Therefore, studies offering evaluations or solutions were

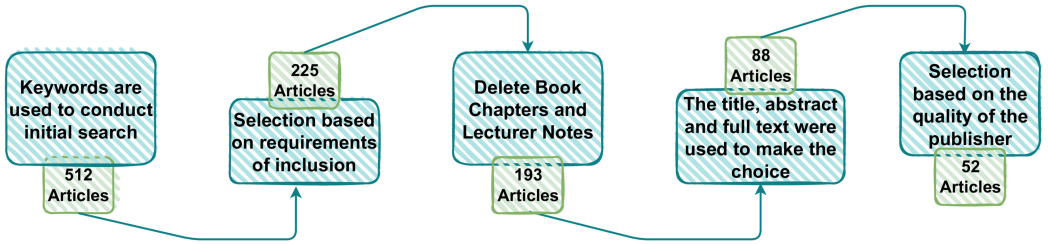

Fig. 5. The stages of the article searching and selection procedure.

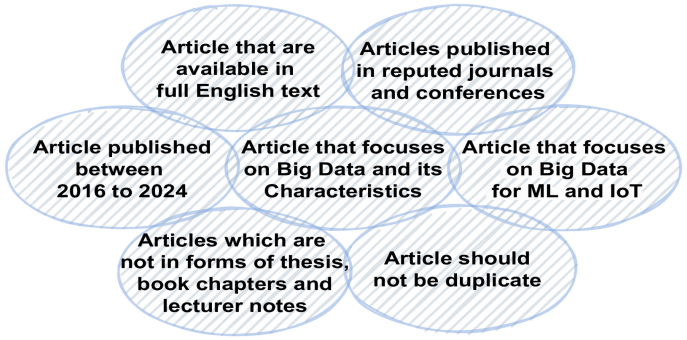

Fig. 6. The standard for selecting articles.

Table 4. Search Terms and Keywords

Search Terms and keywords	S1: *"Big Data" and "Analytics"*
	S2: *"Big Data Characteristics"*
	S3: *"ML" and "Big Data"*
	S4: *"IoT" and "Big Data"*
	S5: *"Big Data" and "AI"*
	S6: *"Big Data" and "Tools"*

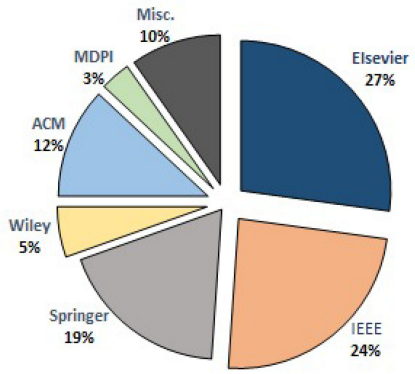

Fig. 7. Percentage of BDA articles published by various publications considered for review.

included, while others were excluded. Thesis, lecturer notes, and book chapters were omitted as their content is often covered in journals or conference papers already included. Additionally, non-peer-reviewed publications were excluded due to uncertain quality, reducing the pool by 32 papers and leaving 193 articles.

In Stage 3, the titles and abstracts of the remaining articles were examined, resulting in 88 publications that met the strict inclusion criteria. Ultimately, 52 manuscripts were selected for the final analysis, with Figure 7 displaying the distribution of journals that published these studies.

5 The 7 V's of Big Data: Research Challenges

The efficient integration of advanced technologies and analytical methodologies is essential for accurate predictions of future events. As big data continues to evolve, so too do the challenges associated with it. This section will delve into the following critical aspects:

— Key research challenges associated with the fundamental characteristics of big data.
— The specific challenges faced when ML and IoT technologies interact with big data.

By examining these challenges, we aim to contribute valuable insights into the ongoing discourse surrounding BDA and its implications for future research and practical applications.

5.1 Data

Big data persists in several challenges due to its distinct characteristics.

Volume: In big data, handling volume introduces unique challenges compared to normal data. Data quality requires more extensive cleaning due to the scale and variety of sources [158]. Data relevance and redundancy become critical in big data due to the vast amount of irrelevant or duplicate data, whereas in normal data, these issues are less frequent [36, 135]. Integration and normalization are significantly more complex in big data, as it often comes from diverse, unstructured sources, unlike normal data, which usually has more consistent formats [30, 64]. Additionally, demands for privacy, security, and storage infrastructure are much more significant in big data due to its sheer size and potential sensitivity. In contrast, normal data is less costly and easier to secure and store [62, 137].

Velocity: While velocity can provide significant benefits, such as real-time insights and faster decision-making, it can pose several challenges. High data rates can overwhelm systems and cause delays [167]. Legacy infrastructure struggles to handle fast data streams [51]. Data quality is affected as quickly generated data can be incomplete or inconsistent [172]. Some applications need real-time processing, requiring specialized systems [72].

Variety: Big data variety introduces challenges like population imbalance, where uneven subgroup sizes can skew analysis [130]. Diverse formats complicate visualization, and data integration from multiple sources becomes time-consuming and error-prone [119, 176]. Additionally, data governance is harder to manage due to unclear policies on ownership and access [83]. Unlike normal data, which is more structured, big data requires more advanced strategies for integration, visualization, and governance.

Veracity: Challenges in leveraging the veracity of big data include data quality and accuracy, data heterogeneity and complexity, data privacy and security, lack of domain expertise, and data governance and management.

Variability: Significant challenges due to the variability of big data are data quality assurance, data integration, scalability and performance, data privacy and security, domain expertise, and data governance practices.

Value: Extracting value from big data is challenging due to issues like identifying relevant data, ensuring data quality, and aligning insights with business objectives. Addressing these requires a comprehensive strategy that includes data governance, quality checks, privacy measures, advanced analytics, and effective communication of insights.

Visualization: Big data visualization has several problems: insufficient data preparation, limited interactivity, misleading visualizations, and limited processing power. It is essential to have a robust data management and preparation process in place to ensure that the data is accurate, relevant, and complete.

5.2 Technology

ML and IoT are broad categories into which the research concerns relevant to big data analysis are divided. But it's not limited.

5.2.1 ML for BDA. BDA revolutionizes processes, facilitates insights, and aids decision-making, with ML playing a critical role. However, traditional ML methods struggle to meet modern big data requirements due to challenges like high dimensionality, sparsity, and large volume. These methods rely on outdated assumptions such as centralized data storage, manageable dataset sizes, and homogeneous data formats, which limit their scalability and effectiveness in distributed, large-scale, and heterogeneous big data environments [111]. Therefore, this section aims to compile, summarize, and organize the ML challenges of dealing with Big Data. This study classifies ML challenges in Big Data based on the Big Data Vs encompassing volume, velocity, variety, veracity, variability, visualization, and value. This strategy focuses on the cause-and-effect relation, which is less frequently covered in other studies.

Volume-driven ML challenges: Volume refers to the amount and scale of data, and it is the most accessible dimension of Big Data to define [109], but it presents numerous difficulties. Here, we discuss the challenges posed by the volume of big data in ML.

(i) ***Computing performance***: The challenge of computational complexity in ML with Big Data arises as data size increases, rendering many algorithms impractical. For instance, the standard SVM algorithm exhibits training time and space complexities $O(m^3)$ and $O(m^2)$, respectively [100], where m represents the training sample count. Data size impacts data storage architecture, necessitating the development of new abstractions like Resilient Distributed Datasets (RDDs) to manage the issue [111]. Therefore, data size affects performance and requires rethinking typical algorithm development architectures.

(ii) ***Pitfalls of modularization***: This issue occurs when data size surpasses memory or single-file capacity, hindering some ML algorithms like iterative graphs and gradient descent. Adapting to distributed computing paradigms like MapReduce poses challenges, requiring more robust solutions for specific algorithms, such as k-means [111].

(iii) ***Imbalanced classes***: Class imbalance in datasets challenges the assumption of uniform class distribution, impacting ML performance. Japkowicz and Stephen have shown that algorithms like decision trees, neural networks, and SVMs are sensitive to this issue [84], particularly when some classes have few samples. In small datasets, class imbalance leads to overfitting and bias in the model [61]. Solutions include creating more data, balancing the classes, using cross-validation, and making the model sensitive to minority classes, whereas, in big data, the main challenges are the need for lots of computing power and difficulty finding minority class examples. Solutions include using powerful computing systems, scalable data balancing techniques, advanced algorithms, reducing data complexity, and processing data in batches [112]. So, addressing the class imbalance issue is more challenging than achieving adequate ML results in big data contexts.

(iv) ***Challenges of high-dimensional data***: Challenges of high-dimensional data involve working in a high-dimensional space, relating to the number of attributes within the dataset [37]. The study [105, 111] says that the impact of dimensionality reduction on algorithms is unpredictable. Spectral clustering is robust in reducing the feature space.

(v) ***Feature Engineering***: High-dimensional datasets and feature engineering are two different challenges in ML. Feature engineering involves creating new features using domain expertise to enhance learning outcomes. However, choosing the best features is a time-consuming process [55]. As datasets expand vertically and horizontally, developing highly relevant features becomes more complex. Similar to dimensionality, the challenges of feature engineering grow

with dataset size. Feature engineering adds new features to improve learning outcomes. Feature selection aims to pick the most pertinent parts [17]. Feature selection can be challenging in high-dimensional settings due to erroneous correlations and unintentional endogeneity [18, 111].

(vi) ***Bonferroni's principle***: According to Leskovec et al. [104], The Bonferroni principle implies higher event probability when searching for specific events in data, potentially compromising ML model accuracy. This statistical problem is frequently called spurious correlation [54]. Calude et al. [31] have studied the frequency and importance of incorrect correlations in big data. They demonstrated many correlations will likely be spurious when the dataset is large enough. When conducting ML with Big Data, it is essential to consider a technique for avoiding these false positives.

(vii) ***Variance and Bias***: ML models face a bias-variance tradeoff, impacting accuracy, especially with big data. Bias reflects how much predictions deviate from actual values on average, while variance measures how much predictions fluctuate with different datasets. Increasing data volume can reduce bias in big data but may increase variance if the additional data is noisy or unrepresentative. Mathematically, this tradeoff is represented as Equation (1) [116]:

$$E\left[(y-\hat{f}(x))^2\right] = \text{Bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x)) + \sigma^2, \tag{1}$$

where y is the actual outcome, $\hat{f}(x)$ is the predicted outcome, and σ^2 is the irreducible error. Unlike normal datasets, where bias and variance may be less pronounced, big data amplifies these challenges due to scale and potential noisiness.

Velocity-driven ML challenges: The velocity aspect of Big Data pertains to the data generation rate and the analysis pace required. Smartphones, real-time sensors, and emerging technologies like smart homes necessitate rapid interaction with the environment. Therefore, the velocity of Big Data has become a crucial factor to consider. Here, we discuss the challenges posed by the big data velocity in ML.

(i) ***Online processing***: Traditional ML methods face challenges processing constant data streams, particularly in real-time. Real-time processing involves immediate data analysis, distinct from model updates with new data arrival. Real-time processing systems provide instantaneous reactions essential for algorithmic trading, stock market forecasting, and traffic management systems [148, 149]. Integrating complex ML algorithms into real-time big data streaming systems faces challenges related to algorithm complexity and the availability of effective online learning solutions.

(ii) ***Concept Drift***: Concept drift alters target variable statistics over time, hampering ML model performance. Types include incremental, gradual, sudden, and recurring, each posing unique challenges. According to Meng Han et al., [73], concept drift can pose several challenges in ML, such as reduced accuracy, increased error rates, and decreased model interpretability. When learning from high-volume streaming data, it is crucial to address concept drift effectively [107], with several proposed techniques broadly classified into active and passive methods, which involve modifying the model or the data to adapt to concept drift.

(iii) ***IID Random Variables***: In ML, i.i.d. assumptions aid convergence, yet real-world data may only sometimes adhere to this. Certain algorithms, like Markov sequences, depend on specific probability distributions. Big data can mitigate i.i.d. constraints due to various factors [60, 111, 127]:

— The i.i.d. assumption requires data to be randomly ordered, which is only sometimes the case in many datasets, making it impractical to randomize the data when dealing with big data.

— Big Data is characterized by its fast and continuous nature, making it impractical to randomize an incomplete dataset or wait for all data to arrive.

The assumption of i.i.d. is essential for many ML algorithms, including neural networks and support vector machines, as demonstrated by Dundar et al. [45]. Due to its high likelihood, it is crucial to address the challenge of broken i.i.d. assumptions in big data.

Variety-driven ML challenges: The variety of Big Data encompasses its structure, data types, semantic interpretation, and sources, presenting challenges that significantly affect data analysis [111]. Here are some issues related to the variety of big data dealing with ML.

(i) ***Heterogeneous data learning***: The challenges of heterogeneous data learning in ML arise due to the structural variation of the dataset, data types, semantic interpretation, and sources [181]. This diversity can lead to difficulties in identifying patterns and developing effective models. Additionally, it can be challenging to integrate multiple data sources with varying levels of reliability and accuracy. These challenges require specialized techniques such as multi-view learning and domain adaptation to handle the complexity of diverse multimedia big data.

(ii) ***Fusing semantics into multimedia neural networks***: The challenge of fusing semantics into multimedia neural networks in the context of extensive data variety refers to the difficulty of incorporating the meaning and interpretation of diverse data types, such as images, videos, and text, into ML models [181]. This challenge arises due to big data's structural, semantic, and source heterogeneity, making it challenging to create unified and practical models to handle such diverse data.

Veracity-driven ML challenges: Big data veracity focuses on ensuring data reliability and quality, which is more complex compared to normal data [111]. This includes data provenance, tracing the origin and processing of massive datasets, and data uncertainty, which arises from noise and incomplete information [74, 173]. Unlike normal data, big data presents challenges with scalability, privacy, and handling diverse, unstructured data, making it harder to manage data tracking and accuracy in ML models, potentially leading to more errors.

Value-driven ML challenges: ML faces challenges in extracting value from big data due to the complexity of processing and analyzing vast datasets. Interpretable ML is crucial in high-stakes fields like healthcare, where decisions need clear explanations, but big data's diversity complicates this. Additionally, big data-driven collaborative decision-making involves multiple users, where explainable ML is needed to provide transparency in how decisions are made [181].

Variability-driven ML challenges: Variability pertains to fluctuations in the data flow that can arise when it becomes challenging to sustain. Such data flow disruptions may be due to the erratic, growing demand for social media data [56]. Variability can pose challenges for ML algorithms, which need consistent and accurate data to provide meaningful insights and predictions. As a result, managing variability is a critical challenge in utilizing big data for ML.

Visualization-driven ML challenges: Data visualization is about making data attractive. Visualization can help users or decision-makers gain better insights into Big Data [94]. Visualization is essential to understand and analyze such big data in remote sensing [94], complex medical image data [138], smart city urban data [121], big graph data [41], and so on to bring out data details relevant to the current aims or objectives. Using ML algorithms in big data visualization presents challenges such as interpretability, scalability, and the need for domain expertise to interpret the results effectively.

5.2.2 Big IoT-Data Analytics. Many organizations have accepted IoT and BDA due to their impact on routine life and business. These technologies are still in their early stages. Several research

challenges still need to be addressed. The following subsections will present big IoT data analytics challenges according to the V's of big data.

Volume-driven IoT challenges: IoT generates vast amounts of data, measured in zettabytes. For handling such massive data, distributed storage, and processing solutions are essential [24]. However, existing distributed solutions face scalability problems for large-scale IoT applications that require real-time and historical data storage. The massive volume of IoT data also causes a strain on network bandwidth, making it difficult for data-intensive applications from multiple stakeholders to use the network efficiently.

Velocity-driven IoT challenges: The speed at which data is ingested is called velocity. The ability of big data solutions to receive and process quickly arriving data is the topic of this subject. Real-time streaming data from the IoT reaches computing infrastructures at a high-speed rate [20]. Additionally, because the rate at which data is ingested from different sources can vary, flow management strategies are needed to govern this rate.

Variety-driven IoT challenges: The characteristic of having various data forms is called variety. Variety poses a challenge for effective data modelling required for storing, querying, processing, and analyzing diverse data formats. IoT applications often use multi-source data, such as WSN, sensor-based IoT devices, multimedia, and so on. Such multi-source data is usually produced in various formats, making it difficult to manage and analyze effectively [69].

Veracity-driven IoT challenges: Veracity refers to uncertainty about the accuracy of acquired data due to a lack of trust in the data source. Inaccuracy of IoT data is introduced at the provenance level due to sensor failure, unreliable data sources, and purposeful misinformation in some MCS applications. Such inaccuracies decrease data quality and reduce the value of analytics decisions based on inconsistent data, even at the data origin stage [24].

Value-driven IoT challenges: Value refers to data-driven information that guides intelligent IoT device actions and decisions. IoT knowledge extraction faces challenges due to the real-time, fine-grained, weakly semantic, and space-time-linked nature of IoT data, data quality issues, and the need to process real-time and historical data. However, analysis quality can be improved by maximizing variables such as the chosen analysis method, data sample, and clean and reliable data [20].

Variability-driven IoT challenges: Variability refers to the dynamic nature of data or data sources. In the IoT, dynamicity arises from IoT devices' mobility or sensor data fluctuations due to real-time events of interest. Rapidly moving IoT devices may experience sporadic data loss during frequent inter-network migrations [24].

Visualization-driven IoT challenges: Visualizing IoT data is challenging as it requires presenting distilled knowledge to users using engaging graphics or visuals. The enormous dimensionality of data, visual imprecision of some data formats like multimedia, ontologies, and social data, and the dynamic nature of real-time scenarios make IoT visualization difficult [69].

6 Discussions: State-of-the-art on BDA

A systematic cause-effect mapping in the previous section identified the challenges related to V's of big data and its intersection with ML and IoT. Professionals in the field of BDA will have significant research potential due to the identification of these challenges. The list of big data challenges highlights the issues and opportunities in current big data solutions. ML and IoT have had a substantial influence on big data technology. This section encompasses an in-depth exploration of prevailing methodologies to address the persistent challenges arising from the V's of big data and their intricate intersections of big data with ML and IoT. Table 6 concisely discusses scholarly solutions for big V's challenges. Further in this section, we will empirically delve into discussing scholarly solutions addressing big data, ML, and IoT intersection challenges.

Table 5. Review of Literature on Proposed Solutions for Big Data Challenges

V's	Challenges	Ref.	Work Highlights	'+'	'-'	Anticipated Opportunities
Volume	Data Quality	[25, 109, 158]	A holistic approach for managing big data quality.	Continuous quality management throughout the big data processing lifecycle.	Lack of qualitative evaluation	Expand the framework to include unstructured big data quality assessment Integration with emerging technologies.
	Data Relevance	[36, 133]	The article explains how to manage data inventories, putting significance and relevance at the core of data explanation.	Framework for assessing data relevance, identifying key factors, addressing challenges, providing recommendations, and case studies.	Limited generalizability.	Integration of emerging technologies
	Data Redundancy	[16, 65, 135]	Explored the significance of data cleansing in the context of big data	Provides a comprehensive review and analysis of various redundant data removal methods specifically tailored for big data.	Limited discussion on data cleansing in streaming or real-time environments	Further research is needed to develop scalable data cleansing techniques that effectively handle data volume, variety and velocity, focusing on optimizing parallel processing.
	Data Normalization	[20, 64]	It focuses on data normalization in the context of massively parallel processing databases that handle big data.	Provide an insight into the strengths and limitations of different normalization approaches and their impact on the performance when dealing with big data.	Lack of real-world validation. Limited comparison and evaluation	Future research could focus on conducting in-depth case studies and real-world applications of the proposed normalization techniques in different domains or industries.
	Data Integration	[30, 164]	Data-driven design and its significance in the context of early product design.	Data integration in the design process is explored from a design and data science perspective.	Context-specific challenges may have limited applicability to other domains.	Future studies could focus on validating and evaluating the effectiveness of data-driven design methods in real-world applications.
	Data Storage and Processing	[66, 137]	Introduces a novel data warehouse architecture to store and handle a large amount of data efficiently.	The proposed "Lake Data Warehouse Architecture" integrates traditional DW, Hadoop, and Apache Spark, enhancing scalability and cost-efficiency.	The article lacks a clear resolution for data quality and accessibility challenges.	Integration with new technologies (such as fog and edge-based big data framework), fostering widespread Big Data adoption.
	Data Privacy and Security	[42, 62, 148]	The article conducts an in-depth analysis and identification of privacy and security challenges in a big data environment.	The study presents a comprehensive analysis of state-of-the-art privacy and security solutions in the context of a big data environment.	Insufficient depth and limited literature review.	The future prospect involves the implementation of cutting-edge security and privacy mechanisms that efficiently safeguard complex big data networks.
Velocity	High Data Rate	[9, 164, 167]	Introduced a distributed and incremental computation method for big data in Cyber-Physical-Social Systems	Proposed distributed, incremental approach for high-quality big data extraction, validated through tests and simulations.	Constrained generalization potential	Future research may emphasize multi-order distributed/incremental model for knowledge discovery from big data
	Infrastructure Limitations	[16, 51, 149]	This article presents a real-time high-load Infrastructure transaction status output prediction model.	Advanced infrastructure utilizing mathematical modelling for error prevention	The model's limitation is reliance on a pre-collected, cleaned dataset, excluding missing data.	The findings facilitate the development of high-quality data projects through the integration of BDA and quantum computing.
	Data Quality	[148, 172]	A method for incorrect data detection in big data cleaning	The proposed method effectively detects missing and abnormal segments, aiding big data cleaning	Inadequate resolution of data inconsistency due to data diversity	Exploration of optimal algorithm such as genetic algorithm to reduce the time complexity of the model.
	Real-Time Processing	[49, 72, 80]	Introduced real-time processing using enhanced distributed recurrent neural networks for social BDA	Extensive experiments validate the solution's accuracy improvement for deep learning models	The context-specific nature of the work restricts its applicability to other domains	Exploring alternative strategies to enhance model performance and testing across diverse datasets would be intriguing.

(Continued)

Table 6. Continued

V's	Challenges	Ref.	Work Highlights	'+'	'-'	Anticipated Opportunities
Variety	Population Imbalance	[43, 130]	Deals with the class imbalance issue of big data.	The Hybrid SMOTE has good scalability within the framework proposed.	Overall classification time increases due to increments in several mapper functions.	Future work may focus on optimizing oversampling algorithms and reducing the classification time.
	Visualization	[31, 119]	Proposed a technical comparison of big data visualization tools.	The analysis uses the latest data, methodologies, techniques and tools.	The implementation details are outside the scope of this article.	In the future, a detailed problem analysis can be pursued through implementation and further analysis.
	Data Integration	[16, 176]	The article proposed a case study leveraging BDA to create sustainable and efficient smart cities.	The study offers tangible insights into how BDA, resource orchestration, and digital sustainability principles can be effectively applied in smart city development.	Lack of comparative analysis	Comparative studies and longitudinal analysis can be the potential future work.
	Data Governance	[83, 165]	The article discusses challenges and approaches to data governance for big data	Proposed a comprehensive framework for data governance specifically tailored to address the challenges posed by big data algorithmic systems	Assumption of data quality and availability	Some possible avenues for future work include empirical validation of the proposed framework and developing interoperability standards for data governance.
Veracity	Truthfulness and Reliability	[19, 40, 63, 162]	The study on methods for improving data veracity	A comprehensive analysis of existing technical solutions to improve data veracity has been presented	Non-experimental approach	Investigation of real cases from related industries to ensure data veracity
Value	Valuable Insights	[7, 22, 105]	The study appraises prevailing literature concerning different aspects of big data value.	Provides a holistic view of the discourse around realizing value from big data, highlighting diverse perspectives.	The article should have addressed real-time developments in big data technologies and methodologies.	The article might point out areas where consensus is lacking, indicating potential avenues for future research.
Variability	Scalability and Integration	[79, 162]	The article examines student reasoning on variability through data preparation in public datasets.	The article contributes a deeper understanding of how variability affects data preparation processes.	The research does not cover real-time developments in data preparation techniques.	There might be potential to integrate advanced statistical methods further to analyze variability during data preparation.
Visualizaton	Misleading Visualization	[33, 114, 150]	This study examines different data visualization phases, their types and numerous applications.	The study presents an intellectual history of data visualization, depicting essential concepts.	The research's concentration on exploratory study limits detailed investigation and validation.	Comparative studies and empirical analysis of big data visualization techniques can be the potential future work.

Machine Learning and Big Data:

Optimizing ML algorithms for big data involves various techniques. Many algorithm-specific methods to reduce time complexity and improve efficiency have been introduced in recent research. The study [59] addressed SVM algorithm improvements for big data, addressing issues related to noise, outliers, and complex dimensions. The study reduced time complexity by incorporating fuzzy membership and optimizing solutions, enabling the accurate fitting of 98 percent for large-scale data. In [125], efficiently handling significant data partitioning without accuracy loss was addressed. An advanced classification model [10] is developed through the optimized integration of multi-kernel SVM (MKSVM) with hyper-heuristic salp swarm optimization (HHSSO). The most informative features are initially obtained using the Population and Global Search Improved Squirrel Search Algorithm (PGS-ISSA). Subsequently, the proposed HHSSO-MKSVM model is employed to classify large datasets based on the selected features. This approach sets optimal kernel functions and parameters, improving accuracy and reducing computational time for big data classification.

Another study [70] examines the relationship between pattern length and algorithm accuracy/-time complexity. Results demonstrate varying execution times as pattern length increases. SVM Linear shows the lowest execution time at length 5 (0.0035 s) and length 25 (0.0012 s) while achieving the highest accuracy (0.963) and F1 score (0.97) among tested algorithms, indicating superior performance in DNA sequence classification. Herrera et al. [78] comprehensively describe the development of a novel Random Forest algorithm implementation on the High-Performance Computing Cluster (HPCC) Systems Platform from LexisNexis. The algorithm was previously unavailable on that platform.

The learning process of Random Forest, which relies on recursive partitioning, posed a challenge due to the prohibition of recursion in HPCC's programming language. However, the recursive partition algorithm was successfully adapted as an iterative split/partition process. Additionally, the identified flaws in the initial implementation are analyzed. All necessary modifications to overcome the iterative split/partition process bottleneck are thoroughly outlined. The study includes optimizing the data gathering of selected independent variables for the node's best-split analysis. Essentially, the article details optimizing the initial Random Forest implementation, transforming it into an efficient distributed ML implementation for Big Data. By fully leveraging the HPCC Systems Platform's Big Data processing and analytics capabilities, the data gathering method was improved from an inefficient "Pass them All and Filter" approach to an entirely parallelized "Fetching on Demand" system. Finally, based on the results of a runtime comparison of these two approaches in the learning process, the speedup of the optimized Random Forest implementation is confirmed.

The Algorithm-specific Techniques that have been discussed are one solution to deal with the computing performance issue of ML algorithms when applied to big data. We can use other methods, such as (a) Sampling and Subset Selection [131]: A random subset can be used for training instead of the entire dataset, significantly reducing the computation time. (b) Feature Selection and Dimensionality Reduction [99]: Choosing the most relevant features can reduce the problem's dimensionality and speed up training. (c) AutoML and Hyperparameter Tuning [52]: Automated machine learning (AutoML) techniques can optimize hyperparameters and algorithms for faster convergence on specific problems. (d) Sparse and Low-Rank Representations [108]: Incorporating sparsity and low-rank structures in algorithms can lead to faster convergence and more efficient training.

Addressing the "pitfalls of modularization" problem in ML when confronted with large data volumes necessitates a multi-faceted approach. It begins with utilizing distributed computing frameworks such as Apache Spark or Hadoop, which enable parallel processing and distributed storage, thus alleviating the computational burden associated with data volume [27]. Optimizing

data pipelines to minimize unnecessary data transfers can streamline the data flow between modules [29].

Furthermore, In-memory processing techniques, which store and manipulate data in RAM, can expedite data access and computation, making modularized algorithms more adept at handling extensive datasets [111]. Lastly, employing data compression strategies can further enhance the efficiency of modular machine-learning algorithms, ensuring they remain effective as data volumes grow [141]. Addressing class imbalance in ML involves effective strategies. Resampling techniques, such as SMOTE, generate synthetic data or reduce the majority class to balance datasets [169]. Algorithmic approaches, like cost-sensitive learning, modify algorithms to focus on minority classes [34]. Ensemble methods like Random Forests or EasyEnsemble combine models for robustness. Advanced loss functions, like Focal Loss [160], prioritize complex examples. Data-level solutions, such as augmentation [71], enhance diversity in the minority class. Combining these approaches ensures effective handling of class imbalance in ML.

Addressing the challenges associated with high-dimensional data in ML for big data necessitates a range of sophisticated techniques. These encompass feature selection and dimensionality reduction methods such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) [22] for projecting data into lower-dimensional spaces while preserving critical information. Sparse representations via L1 regularization and feature extraction through deep learning approaches are instrumental in mitigating dimensionality [79]. Manifold learning algorithms like Isomap and Locally Linear Embedding (LLE) reveal lower-dimensional structures in complex data. Furthermore, employing domain-specific knowledge and advanced methods like Elastic Net or Recursive Feature Elimination (RFE) contributes to effective dimensionality reduction [114]. Integrating random projections [12], tree-based techniques [48], kernel functions, and incremental learning [85] offer additional strategies for handling high-dimensional data intricacies. The selection of these techniques is contingent upon the dataset's attributes and the specific machine-learning problem, often benefiting from a hybrid approach or domain expertise. False positives from many tests are minimized in ML for massive data to solve Bonferroni's Principle. Techniques such as feature selection, dimensionality reduction, regularization, and error control methods (e.g., Bonferroni correction and FDR control) are employed [5, 31, 54]. The approach is tailored to the specific data and problem, often combining techniques for improved outcomes.

In ML for big data, regularization techniques like early stopping, Lasso, and Ridge regression can address the bias-variance tradeoff. These techniques enhance model generalization and reduce overfitting, which is crucial when dealing with vast datasets. However, it's important to note that they introduce additional parameters requiring tuning, typically performed through cross-validation and grid search. While effective, these processes can be computationally intensive, particularly with large datasets. Therefore, further research is needed to optimize and evaluate the efficiency of regularization techniques in mitigating the bias-variance problem in the context of Big Data [38]. Managing high data velocity requires specialized strategies. Stream processing frameworks such as Apache Kafka enable real-time data handling, ensuring that incoming data is processed promptly. Additionally, parallel and distributed processing frameworks like Apache Spark and Hadoop MapReduce [126] facilitate the simultaneous processing of data across multiple nodes or clusters, efficiently handling large volumes of high-velocity data.

Online processing is essential for handling data as it arrives in real-time. Online learning algorithms are designed to adapt continuously to incoming data, updating model parameters incrementally. Moreover, micro-batching techniques, often integrated into stream processing frameworks [151], allow data to be processed in small, manageable chunks. This combination

of online processing and micro-batching balances real-time analysis and maintains some of the benefits of batch processing.

Concept drift spans a range of problems in the context of large data velocity, each requiring a unique approach to identification and prevention. Sudden concept drift, marked by abrupt shifts, demands continuous monitoring and immediate retraining. Incremental drift, characterized by gradual changes, requires sliding windows, periodic updates, and online learning. Recurring drift involves techniques like sliding windows and online learning [115]. Virtual drift, where changes have minimal impact, relies on statistical tests for differentiation. Concept drift from external factors requires external data integration and robust model development. Label drift is managed through regular label validation and drift detection, while feature drift involves continuous feature distribution monitoring and model updates [106]. These strategies, grounded in proactive monitoring, data preprocessing, model adaptation, and retraining, can be tailored to specific concept drift types and problem domains, ensuring models remain robust in dynamic, high-velocity data environments. Feature engineering techniques are crucial to maintaining the assumption of Independent and Identically Distributed (IID) random variables. These techniques involve carefully manipulating features to ensure variables follow the IID belief, which is fundamental to many ML algorithms. Data preprocessing steps, including data scaling, encoding, and transformation, help standardize data and remove dependencies and systematic patterns that may violate the IID assumption [144]. These technical strategies address the challenges of big data velocity in ML. They use specialized frameworks, algorithms, and techniques to ensure that ML models can effectively handle large volumes of rapidly changing data while maintaining their accuracy and statistical properties.

A set of specialized approaches is essential to confront the complexities of big data variety challenges, such as heterogeneous data learning and integrating semantics into multimedia neural networks. Heterogeneous data learning necessitates applying methodologies like transfer learning [134], multi-modal learning [90], ensemble techniques [122, 170], and feature engineering [90]. These techniques empower handling diverse data types, harnessing their unique attributes for more resilient analytical processes. In infusing semantics into multimedia neural networks, the strategies encompass the utilization of semantic embeddings, multi-modal architectures, attention mechanisms, and cross-modal retrieval methods [93, 97]. These tactics are devised to bridge the semantic divide between disparate data modalities, ultimately enhancing the models' comprehensive comprehension and predictive capabilities.

Addressing veracity challenges in big data, such as data tracking and tracing, as well as data uncertainty, involves deploying specialized solutions. For data tracking and tracing, blockchain technology ensures the immutability and transparency of data [76, 140], while audit trails and robust metadata systems provide extensive data lineage documentation [11, 171]. When dealing with data uncertainty, probabilistic models like Bayesian networks [87] and Monte Carlo simulations [178] incorporate uncertainty into predictions and decisions. Additionally, uncertainty quantification [1] and data fusion [132] are utilized to assess and mitigate uncertainty, and expert systems are harnessed to handle uncertain or ambiguous data. The solution choice depends on the precise veracity challenges encountered, intending to enhance data quality and reliability for more accurate insights.

Addressing value-related challenges in big data, including interpretable ML for decision-making and collaborative decision-making among multiple users, entails the implementation of specialized solutions. In interpretable ML, techniques such as feature importance assessment [13], explainable model architectures [113], and methods like LIME [113] and SHAP [13] are employed to elucidate model predictions and enable stakeholders to comprehend the rationale behind decisions. On the other hand, in the context of collaborative decision-making driven by big data,

strategies encompass collaborative filtering algorithms for personalized recommendations [139], distributed decision support systems for combined data analysis [103], and the application of multi-criteria decision analysis (MCDA) techniques to facilitate the weighing and prioritization of criteria [103]. These solutions are tailored to empower users with transparent insights and promote effective collaborative decision-making while harnessing the full potential of big data. The choice of approach is contingent on the specific requirements and complexities inherent in the decision-making processes and the diverse characteristics of the involved data and users.

The challenge of big data variability is a prevalent issue, particularly pronounced within social media datasets. Social media is primarily attributed to the extensive diversity among users and the vast spectrum of topics discussed across these platforms. Social media content exhibits significant variability, necessitating the application of advanced analytical techniques. One such method is information diffusion modelling, which aids in comprehending how information spreads and evolves within the intricate networks of social media [80]. By closely monitoring the propagation and dissemination of information, researchers and analysts can gain valuable insights into the dynamic nature of social media data, enabling more effective analysis and decision-making processes. Big data visualization presents challenges, including data volume, complexity, and real-time rendering.

General solutions include data aggregation, summarization, and utilizing distributed computing for handling extensive data [120]. For complex data, employing dimensionality reduction and GPU acceleration aids in rendering. Maintaining real-time interactivity is possible through progressive loading and responsive design. Collaborative filtering methods enhance recommendation systems, and dashboard design facilitates storytelling [155]. Additionally, adherence to web accessibility standards ensures inclusivity and data anonymization safeguards privacy [145]. These solutions enable effective visualization and interpretation of vast, intricate big data, supporting data-driven insights and decision-making processes. Further, emerging frameworks like Apache Spark MLlib and TensorFlow Extended (TFX) are specifically designed to handle large datasets in a distributed cluster environment. These frameworks parallelize the computations, allowing ML algorithms to scale across multiple nodes, thereby overcoming the limitations of traditional, single-node methods [81].

IoT and Big Data:

Managing large volumes of IoT data requires scalable solutions. Distributed processing frameworks like Apache Hadoop and Spark offer horizontal scalability, enabling efficient processing of massive IoT datasets. Additionally, by strategically placing computing nodes closer to IoT devices, implementing edge computing reduces the need for extensive data transmission over network infrastructures [82]. Applying these methods alleviates network bandwidth strain and enhances real-time data processing capabilities. Real-time data management is crucial in IoT applications. Stream processing technologies such as Apache Kafka and Apache Flink are instrumental in handling high-velocity IoT data streams in real-time [68]. Complex Event Processing (CEP) systems add another intelligence layer by detecting and responding to intricate patterns and events within rapidly flowing IoT data streams [95]. IoT data often comes in various formats, posing data integration challenges. To address this, organizations can deploy data integration tools and platforms to harmonize and transform data from diverse sources into a standardized schema. Alternatively, they can adopt a schema-on-read approach, which interprets data during analysis, allowing flexibility in handling different data formats [69]. Ensuring the reliability of IoT data is essential. Robust data validation techniques can be applied to identify and filter out unreliable data from potentially trustworthy sources. Furthermore, integrating federated learning [153] and blockchain or distributed ledger technology can create immutable, verifiable data records, enhancing data veracity [143].

Extracting valuable insights from IoT data in real-time requires advanced techniques. ML and artificial intelligence (AI) play a crucial role by enabling predictive maintenance, anomaly

detection, and immediate decision-making based on data-driven insights [21]. CEP engines are also valuable for identifying critical events and trends in real-time data streams, facilitating rapid responses [154]. Organizations can implement data buffering and checkpoint mechanisms to mitigate the risk of data loss during inter-network migrations or disruptions, ensuring data resilience [180]. Data visualization tools and platforms like Tableau and Power BI offer advanced capabilities for creating custom and interactive visualizations. These tools enable in-depth analysis of IoT data, making it more accessible and actionable. Moreover, organizations can develop custom dashboards tailored to specific IoT applications, providing situational awareness and insights for technical teams [102].

Instead of having these solutions, it has been observed that the limitation within various solutions for Big Data, IoT, and ML lies in the absence of a universally applicable approach. This drawback arises from the intricate nature of diverse data scenarios, particularly involving heterogeneous data types like streaming, video, image, and numerical data. Each data type necessitates specific preprocessing, normalization, or analytical methods tailored to its distinct characteristics and requirements [128]. Consequently, it is common to encounter situations where a technique well-suited for one data type may experience reduced effectiveness when applied to another, highlighting the need for adaptable and versatile solutions.

7 Future Research Directions

The present study has significantly advanced our understanding of optimizing ML algorithms for big data. However, it is imperative to acknowledge that this field remains ripe with untapped potential, offering a plethora of unexplored avenues for future research and innovation. Within the realm of ongoing research priorities and the pressing needs of the hour, several critical areas demand focused attention and investigation.

In ML, future research holds promising avenues for advancing the big data field capabilities. These avenues encompass the development of specialized variant models tailored to specific data types and applications aimed at enhancing ML adaptability and effectiveness. Innovative **kernel functions optimization** for handling high-dimensional data in big data scenarios presents an area of exploration. Adapting solution spaces and kernel functions based on data characteristics via automated methods can improve model performance. Researchers must optimize the computational efficiency of traditional ML algorithms to process vast data sets swiftly.

Moreover, scalability, resource efficiency, and exploration of parallel and distributed computing are imperative for addressing the challenges of increasing data volumes. Integrating feature selection methods compatible with alternative encoding approaches, **advanced hyperparameter optimization techniques**, and developing **deep clustering models** to capture global low-rank information constitute vital research directions. Furthermore, enhancing *SMOTE variants* for handling imbalanced datasets is essential. These collective endeavors pave the way for more adaptable, efficient, and robust ML models in the era of big data. Future exploration in dimensionality reduction technique (DRT) involves advancing deep learning-based methods, improving efficiency, ensuring adaptability, addressing fairness and privacy, handling multi-modal data, developing interactive DRTs, standardizing benchmarks, and tailoring DRTs to specific domains, among other areas. These efforts aim to enhance the versatility and effectiveness of DRTs.

Despite online learning algorithms and micro-batching techniques, optimizing the efficiency of real-time data processing remains an evolving concern. Enhancements in algorithm efficiency and the development of **specialized hardware accelerators** are required to handle ever-increasing data velocity while maintaining low latency. Adapting concept drift detection to rapidly changing data is challenging; **sophisticated automated methods** for diverse shifts are crucial. Ensuring

ML models remain robust in concept drift is a continuous challenge. Techniques for timely model retraining and adaptation must be refined to cope with dynamic data.

Additionally, future research can address the tradeoff between model stability and adaptability. Rising data velocity demands scalable, parallel processing frameworks. Optimizing ML systems for distributed processing across clusters is vital.

A vital issue is resource balance in managing high-velocity data. Innovations in **resource-efficient ML algorithms** are essential to avoid bottlenecks. Ensuring ethical AI, including bias mitigation and fairness, in high-velocity data contexts is critical. Ongoing research and algorithm development are necessary for ethical AI promotion. Ensuring data reliability and mitigating uncertainty is critical and needs further research. Advancing interpretability in complex AI models will be pivotal for fostering user confidence. Innovative analytical techniques will be required to cope with the variability in social media data. The potential impact of **quantum computing** on BDA is significant, as it can address current limitations and unlock new research and application opportunities. Quantum computing, capable of processing complex computations at exponentially faster rates, has the potential to revolutionize BDA by boosting the speed and efficiency of ML tasks, currently hindered by classical computation. This could lead to breakthroughs in handling high-dimensional data and complex models. Future research should prioritize the development of indispensable scalable visualization solutions to cope with the ever-expanding volumes of data.

In big IoT data, several critical technical challenges persist, along with emerging priorities that demand attention. First and foremost, exploring advanced methods for scaling up data processing beyond traditional frameworks is essential. **Federated learning** framework with its ability to train ML models across decentralized devices without the need to share raw data. This not only helps protect user privacy but also reduces the challenges associated with data centralization, especially with growing concerns about data governance and compliance. It allows for scalable analytics across distributed environments, fostering collaboration across organizations without the need for data pooling.

Real-time applications need low-latency edge computing, demanding improvements in infrastructure and algorithms. **Edge computing** addresses the latency and bandwidth issues in centralized big data systems by processing data closer to its source. This decentralization reduces the need for data transfer to central servers, making real-time analytics more feasible, especially in IoT and sensor-driven environments. It also addresses data privacy and security concerns by minimizing data movement. Enhancing stream processing is vital for managing high-speed data and detecting complex events. Interoperability standards are critical for seamless data sharing among IoT devices. Incorporating **Edge AI** and ML into IoT is a crucial research area, enabling real-time decisions and predictive maintenance. Addressing resource limitations means developing lightweight models and ensuring they work well with IoT applications.

Data governance, regulatory compliance, and energy-efficient designs remain essential. Achieving effective data integration, semantic understanding, and **edge-cloud synergy** is vital for better data analytics. Lastly, resilience and adapting to regulations require ongoing collaboration in the IoT and big data communities.

8 Conclusion

The article provides an extensive and in-depth exploration of BDA, offering significant value to researchers and practitioners. It serves as a comprehensive resource for understanding the current state of analytics, encompassing foundational concepts, challenges, advanced solutions, and future prospects in the light of big data V's. It extensively presents the state of BDA tools, models and technologies. This makes it an indispensable reference for those involved in BDA and allied areas.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76, 12 (2021), 243–297.

[2] Flasteen Abuqabita, Razan Al-Omoush, and Jaber Alwidian. 2019. A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science* 13, 7 (2019), 1–14.

[3] Flasteen Abuqabita, Razan Al-Omoush, and Jaber Alwidian. 2019. A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science* 13, 7 (2019), 1–14.

[4] Debi Prasanna Acharjya and Kauser Ahmed. 2016. A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications* 7, 2 (2016), 511–518.

[5] Ji Seon Ahn, Kyungun Jhung, Jooyoung Oh, Jaeseok Heo, Jae-Jin Kim, and Jin Young Park. 2022. Association of resting-state theta–gamma coupling with selective visual attention in children with tic disorders. *Frontiers in Human Neuroscience* 16, 9 (2022), 1017703.

[6] Bilal Akil, Ying Zhou, and Uwe Röhm. 2017. On the usability of Hadoop MapReduce, Apache Spark and Apache flink for data science. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, USA, 303–310.

[7] Khaled Saleh Al-Omoush, Fernando Garcia-Monleon, and José Manuel Mas Iglesias. 2024. Exploring the interaction between big data analytics, frugal innovation, and competitive agility: The mediating role of organizational learning. *Technological Forecasting and Social Change* 200, 3 (2024), 123188.

[8] Zaher Ali Al-Sai, Mohd Heikal Husin, Sharifah Mashita Syed-Mohamad, Rasha Moh'd Sadeq Abdin, Nour Damer, Laith Abualigah, and Amir H. Gandomi. 2022. Explore big data analytics applications and opportunities: A review. *Big Data and Cognitive Computing* 6, 4 (2022), 157.

[9] Ali M. Al-Salim, Taisir E. H. El-Gorashi, Ahmed Q. Lawey, and Jaafar M. H. Elmirghani. 2018. Greening big data networks: Velocity impact. *IET Optoelectronics* 12, 3 (2018), 126–135.

[10] Issa Mohammed Saeed Ali and D. Hariprasad. 2023. Hyper-heuristic salp swarm optimization of multi-kernel support vector machines for big data classification. *International Journal of Information Technology* 15, 2 (2023), 651–663.

[11] Haneen Alosert, James Savery, Jennifer Rheaume, Matthew Cheeks, Richard Turner, Christopher Spencer, Suzanne S. Farid, and Stephen Goldrick. 2022. Data integrity within the biopharmaceutical sector in the era of Industry 4.0. *Biotechnology Journal* 17, 6 (2022), 2100609.

[12] Majid Altuwairiqi. 2023. Combining extreme learning machine through random projections for dimensional information taxonomy and assembling. In *Proceedings of the 2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)*. IEEE, 488–491.

[13] Mohammad Alwadi, Girija Chetty, and Mohammad Yamin. 2023. A framework for vehicle quality evaluation based on interpretable machine learning. *International Journal of Information Technology* 15, 1 (2023), 129–136.

[14] Fairuz Amalina, Ibrahim Abaker Targio Hashem, Zati Hakim Azizul, Ang Tan Fong, Ahmad Firdaus, Muhammad Imran, and Nor Badrul Anuar. 2019. Blending big data analytics: Review on challenges and a recent study. *IEEE Access* 8 (2019), 3629–3645.

[15] apache. 2009. hive 2009. [Online]. Available:. https://hive.apache.org/. Access date: 11th September 2023.

[16] Elisa Arrigo, Caterina Liberati, and Paolo Mariani. 2021. Social media data and users' preferences: A statistical analysis to support marketing communication. *Big Data Research* 24, 2 (2021), 100189. DOI: https://doi.org/10.1016/j.bdr.2021.100189

[17] Naziya Aslam, Shashank Srivastava, and M. M. Gore. 2023. A comprehensive analysis of machine learning-and deep learning-based solutions for DDoS attack detection in SDN. *Arabian Journal for Science and Engineering* 49, 3 (2023), 1–41.

[18] Naziya Aslam, Shashank Srivastava, and M. M. Gore. 2024. DDoS SourceTracer: An intelligent application for DDoS attack mitigation in SDN. *Computers and Electrical Engineering* 117, 5 (2024), 109282.

[19] Fatmah Assiri. 2020. Methods for Assessing, predicting, and improving data veracity: A survey. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 9, 4 (2020), 5.

[20] Safa Ben Atitallah, Maha Driss, Wadii Boulila, and Henda Ben Ghézala. 2020. Leveraging deep learning and IoT big data analytics to support the smart cities development: Review and future directions. *Computer Science Review* 38, 4 (2020), 100303.

[21] Serkan Ayvaz and Koray Alpay. 2021. Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. *Expert Systems with Applications* 173, 11 (2021), 114598.

[22] Ashley Babjac, Taylor Royalty, Andrew D. Steen, and Scott J. Emrich. 2022. A comparison of dimensionality reduction methods for large biological data. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 1–7.

[23] Omar Backhoff and Eirini Ntoutsi. 2016. Scalable online-offline stream clustering in apache spark. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW'16)*. IEEE, 37–44.
[24] Maggi Bansal, Inderveer Chana, and Siobhán Clarke. 2020. A survey on iot big data: Current status, 13 v's challenges, and future directions. *ACM Computing Surveys* 53, 6 (2020), 1–59.
[25] Kornelia Batko and Andrzej Ślęzak. 2022. The use of big data analytics in healthcare. *Journal of Big Data* 9, 1 (2022), 3.
[26] Gema Bello-Orgaz, Jason J. Jung, and David Camacho. 2016. Social big data: Recent achievements and new challenges. *Information Fusion* 28, 2 (2016), 45–59.
[27] Sana Ben Hamida, Ghita Benjelloun, and Hmida Hmida. 2021. Trends of evolutionary machine learning to address big data mining. In *Proceedings of the International Conference on Information and Knowledge Systems*. Springer, 85–99.
[28] Siddharth Bhatia and Rajiv Kumar. 2018. Review of graph processing frameworks. In *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 998–1005.
[29] Matthew Brendel, Chang Su, Zilong Bai, Hao Zhang, Olivier Elemento, and Fei Wang. 2022. Application of deep learning on single-cell RNA sequencing data analysis: a review. *Genomics, Proteomics and Bioinformatics* 20, 5 (2022), 814–835.
[30] Tristan Briard, Camille Jean, Améziane Aoussat, and Philippe Véron. 2023. Challenges for data-driven design in early physical product design: A scientific and industrial perspective. *Computers in Industry* 145, 2 (2023), 103814.
[31] Cristian S. Calude and Giuseppe Longo. 2017. The deluge of spurious correlations in big data. *Foundations of Science* 22, 3 (2017), 595–612.
[32] Paris Carbone, Asterios Katsifodimos, Stephan Ewen, Volker Markl, Seif Haridi, and Kostas Tzoumas. 2015. Apache flink: Stream and batch processing in a single engine. *The Bulletin of the Technical Committee on Data Engineering* 38, 4 (2015), 28–38.
[33] C. L. Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, 22 (2014), 314–347.
[34] Yingying Chen, Zijie Hong, and Xiaowei Yang. 2023. Cost-sensitive online adaptive kernel learning for large-scale imbalanced classification. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 10554–10568.
[35] Muskaan Chopra, Sunil K. Singh, Anshul Gupta, Kriti Aggarwal, Brij B. Gupta, and Francesco Colace. 2022. Analysis & prognosis of sustainable development goals using big data-based approach during COVID-19 pandemic. *Sustainable Technology and Entrepreneurship* 1, 2 (2022), 100012.
[36] Naomi Clarke. 2019. How to ensure provision of accurate data to enhance decision-making. *Journal of Securities Operations and Custody* 11, 2 (2019), 112–127.
[37] Adolfo Crespo Márquez. 2022. *The Curse of Dimensionality*. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-97660-6_7
[38] Yehuda Dar, Vidya Muthukumar, and Richard G. Baraniuk. 2021. A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning. *arXiv preprint arXiv:2109.02355*.
[39] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Communications of the ACM* 51, 1 (2008), 107–113.
[40] Natarajan Deepa, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, and Pubudu N. Pathirana. 2022. A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems* 131, 6 (2022), 209–226.
[41] Wajdi Dhifli, Sabeur Aridhi, and Engelbert Mephu Nguifo. 2017. MR-SimLab: Scalable subgraph selection with label similarity for big data. *Information Systems* 69 (2017), 155–163. DOI: https://doi.org/10.1016/j.is.2017.05.006
[42] Grazia Dicuonzo, Graziana Galeone, Erika Zappimbulso, Vittorio Dell'Atti, et al. 2019. Risk management 4.0: The role of big data analytics in the bank sector. *International Journal of Economics and Financial Issues* 9, 6 (2019), 40–47.
[43] Papa Senghane Diouf, Aliou Boly, and Samba Ndiaye. 2018. Variety of data in the ETL processes in the cloud: State-of-the-art. In *Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD)*. IEEE, 1–5.
[44] Jaschar Domann, Jens Meiners, Lea Helmers, and Andreas Lommatzsch. 2016. Real-time news recommendations using apache spark. In *Proceedings of the CLEF (Working Notes)*. 628–641.
[45] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. 2007. Learning classifiers when the training data is not IID. In *Proceedings of the IJCAI*. 756–61.
[46] Benedikt Elser and Alberto Montresor. 2013. An evaluation study of bigdata frameworks for graph processing. In *Proceedings of the 2013 IEEE International Conference on Big Data*. IEEE, 60–67.
[47] Isitor Emmanuel and Clare Stanier. 2016. Defining big data. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*. 1–6.
[48] Zi Fang, Zhuang Fu, Linhui Zhou, Zeyu Fu, and Yisheng Guan. 2023. Prediction for loosening life of bolted joints using IMUs with dimensionality reduction. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–17.

[49] Farzam Farbiz, Yuan Miaolong, and Zhou Yu. 2020. A cognitive analytics based approach for machine health monitoring, anomaly detection, and predictive maintenance. In *Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA'20)*. IEEE, 1104–1109.

[50] Marzieh Fathi, Mostafa Haghi Kashani, Seyed Mahdi Jameii, and Ebrahim Mahdipour. 2022. Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering* 29, 2 (2022), 1247–1275.

[51] Solomia Fedushko, Taras Ustyianovych, and Michal Gregus. 2020. Real-time high-load infrastructure transaction status output prediction using operational intelligence and big data technologies. *Electronics* 9, 4 (2020), 668.

[52] Konstantinos Filippou, George Aifantis, George A. Papakostas, and George E. Tsekouras. 2023. Structure learning and hyperparameter optimization using an automated machine learning (AutoML) pipeline. *Information* 14, 4 (2023), 232.

[53] Keerthana G. and Sherly Puspha Annabel L. 2025. A survey on big data classification. *Data and Knowledge Engineering* 156, 2 (2025), 102408. DOI: https://doi.org/10.1016/j.datak.2025.102408

[54] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 2 (2015), 137–144.

[55] Amir H. Gandomi, Fang Chen, and Laith Abualigah. 2022. Machine Learning Technologies for Big Data Analytics. *Electronics* 11, 3 (2022), 421 pages.

[56] Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband, and Fariza Hanum. 2016. A survey on indexing techniques for big data: Taxonomy and performance evaluation. *Knowledge and Information Systems* 46, 2 (2016), 241–284.

[57] Preeti Garg and Vineet Sharma. 2014. An efficient and secure data storage in mobile cloud computing through RSA and hash function. In *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT'14)*. IEEE, 334–339.

[58] Gartner.com. 2016. What is Big Data? By Gartner Research, Sep. 2016. [Online]. Available: http://www.gartner.com/it-glossary/big-data/. Access date: 11th September 2023.

[59] Babacar Gaye, Dezheng Zhang, and Aziguli Wulamu. 2021. Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering* 2021, 1 (2021), 1–9.

[60] Zoubin Ghahramani. 2015. Probabilistic machine learning and artificial intelligence. *Nature* 521, 7553 (2015), 452–459.

[61] Mojgan Ghanavati, Raymond K. Wong, Fang Chen, Yang Wang, and Chang-Shing Perng. 2014. An effective integrated method for learning big imbalanced data. In *Proceedings of the 2014 IEEE International Congress on Big Data*. IEEE, 691–698.

[62] Parth Goel, Radhika Patel, Dweepna Garg, and Amit Ganatra. 2021. A review on big data: Privacy and security challenges. In *Proceedings of the 2021 3rd International Conference on Signal Processing and Communication (ICPSC'21)*. IEEE, 705–709.

[63] David Goldston. 2008. Big data: Data wrangling. *Nature* 455, 7209 (2008), 15.

[64] Nikolay Golov and Lars Rönnbäck. 2017. Big data normalization for massively parallel processing databases. *Computer Standards and Interfaces* 54, 7 (2017), 86–93.

[65] Cristian González García and Eva Álvarez-Fernández. 2022. What is (not) big data based on its 7Vs challenges: A survey. *Big Data and Cognitive Computing* 6, 4 (2022), 158.

[66] P. R. C. Gopal, Nripendra P. Rana, Thota Vamsi Krishna, and M. Ramkumar. 2024. Impact of big data analytics on supply chain performance: An analysis of influencing factors. *Annals of Operations Research* 333, 2 (2024), 769–797.

[67] Vairaprakash Gurusamy, Subbu Kannan, and K. Nandhini. 2017. The real time big data processing framework: Advantages and limitations. *International Journal of Computer Sciences and Engineering* 5, 12 (2017), 305–312.

[68] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, and Muhammad Imran. 2019. Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management* 45, 2 (2019), 289–307.

[69] Yosra Hajjaji, Wadii Boulila, Imed Riadh Farah, Imed Romdhani, and Amir Hussain. 2021. Big data and IoT-based applications in smart environments: A systematic review. *Computer Science Review* 39, 1 (2021), 100318.

[70] Belal A. Hamed, Osman Ali Sadek Ibrahim, and Tarek Abd El-Hafeez. 2023. Optimizing classification efficiency with machine learning techniques for pattern matching. *Journal of Big Data* 10, 1 (2023), 124.

[71] Soufiane Hamida, Oussama El Gannour, Ayyad Maafiri, Yasser Lamalem, Ikram Haddou-Oumouloud, and Bouchaib Cherradi. 2023. Data balancing through data augmentation to improve transfer learning performance for skin disease prediction. In *Proceedings of the 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET'23)*. IEEE, 1–7.

[72] Badr Ait Hammou, Ayoub Ait Lahcen, and Salma Mouline. 2020. Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. *Information Processing and Management* 57, 1 (2020), 102122.

[73] Meng Han, Zhiqiang Chen, Muhang Li, Hongxin Wu, and Xilong Zhang. 2022. A survey of active and passive concept drift handling methods. *Computational Intelligence* 38, 4 (2022), 1492–1535.
[74] Reihaneh H. Hariri, Erik M. Fredericks, and Kate M. Bowers. 2019. Uncertainty in big data analytics: Survey, opportunities, and challenges. *Journal of Big Data* 6, 1 (2019), 1–16.
[75] Islam Hassanin, Aya ElGarhy, and Sobhy Mostafa. 2021. The role of big data analytics for enhancing the internet of things applications in megacities: The case of traffic optimization and control. *Global Business and Management Research* 13, 3 (2021), 297–306.
[76] Pawan Hegde and Praveen Kumar Reddy Maddikunta. 2023. Amalgamation of blockchain with resource-constrained IoT devices for healthcare applications–state of art, challenges and future directions. *International Journal of Cognitive Computing in Engineering* 4, 1 (2023), 220–239.
[77] Safiollah Heidari, Yogesh Simmhan, Rodrigo N. Calheiros, and Rajkumar Buyya. 2018. Scalable graph processing frameworks: A taxonomy and open challenges. *ACM Computing Surveys* 51, 3 (2018), 1–53.
[78] Victor M. Herrera, Taghi M. Khoshgoftaar, Flavio Villanustre, and Borko Furht. 2019. Random forest implementation and optimization for Big Data analytics on LexisNexis's high performance computing cluster platform. *Journal of Big Data* 6, 1 (2019), 1–36.
[79] Liangchen Hu, Jingke Xu, Lei Tian, and Wensheng Zhang. 2020. Self-centralized jointly sparse maximum margin criterion for robust dimensionality reduction. *Knowledge-based Systems* 206, 19 (2020), 106343.
[80] Adriana Iamnitchi, Lawrence O. Hall, Sameera Horawalavithana, Frederick Mubang, Kin Wai Ng, and John Skvoretz. 2023. Modeling information diffusion in social media: Data-driven observations. *Frontiers in Big Data* 6 (2023), 1135191.
[81] Md Johirul Islam. 2020. *Towards Understanding the Challenges Faced by Machine Learning Software Developers and Enabling Automated Solutions*. Ph.D. Dissertation. Iowa State University.
[82] Mustafa Musa Jaber, Mohammed Hasan Ali, Sura Khalil Abd, Ahmed Alkhayyat, et al. 2023. Application of edge computing-based information-centric networking in smart cities. *Computer Communications* 211, 15 (2023), 46–58.
[83] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly* 37, 3 (2020), 101493.
[84] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6, 5 (2002), 429–449.
[85] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. 2022. Feature dimensionality reduction: A review. *Complex and Intelligent Systems* 8, 3 (2022), 2663–2693.
[86] Hui Jiang. 2019. Research and practice of big data analysis process based on hadoop framework. In *Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC'19)*. IEEE, 2044–2047.
[87] Xue-Bo Jin, Zhong-Yao Wang, Wen-Tao Gong, Jian-Lei Kong, Yu-Ting Bai, Ting-Li Su, Hui-Jun Ma, and Prasun Chakrabarti. 2023. Variational Bayesian network with information interpretability filtering for air quality forecasting. *Mathematics* 11, 4 (2023), 837.
[88] Stephen Kaisler, J. Alberto Espinosa, William Money, and Frank Armour. 2023. Big data and analytics: Issues and challenges for the past and next ten years. In *Proceedings of the 56th Hawaii International Conference on System Sciences*.
[89] Sema A. Kalaian, Rafa M. Kasim, and Nabeel R. Kasim. 2019. Descriptive and predictive analytical methods for big data. In *Proceedings of the Web Services: Concepts, Methodologies, Tools, and Applications*. IGI Global, 314–331.
[90] Simon Kamm, Sushma Sri Veekati, Timo Müller, Nasser Jazdi, and Michael Weyrich. 2023. A survey on machine learning based analysis of heterogeneous data in industrial automation. *Computers in Industry* 149, 6 (2023), 103930.
[91] Yaghoob Karimi, Mostafa Haghi Kashani, Mohammad Akbari, and Ebrahim Mahdipour. 2021. Leveraging big data in smart cities: A systematic review. *Concurrency and Computation: Practice and Experience* 33, 21 (2021), e6379.
[92] Harkiran Kaur and Aanchal Phutela. 2018. Commentary upon descriptive data analytics. In *Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC'18)*. IEEE, 678–683.
[93] Muhammad Jaleed Khan, John G. Breslin, and Edward Curry. 2022. Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications. *IEEE Internet Computing* 26, 4 (2022), 21–27.
[94] Nawsher Khan, Arshi Naim, Mohammad Rashid Hussain, Quadri Noorulhasan Naveed, Naim Ahmad, and Shamimul Qamar. 2019. 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*. 19–24.
[95] Behnam Khazael, Mojtaba Vahidi Asl, and Hadi Tabatabaee Malazi. 2023. Geospatial complex event processing in smart city applications. *Simulation Modelling Practice and Theory* 122, 1 (2023), 102675.
[96] Taiwo Kolajo, Olawande Daramola, and Ayodele Adebiyi. 2019. Big data stream analysis: A systematic literature review. *Journal of Big Data* 6, 1 (2019), 47.

[97] Lalit Kumar and Dushyant Kumar Singh. 2023. A comprehensive survey on generative adversarial networks used for synthesizing multimedia content. *Multimedia Tools and Applications* 82, 26 (2023), 1–40.
[98] Praveen Kumar, Parveen Kumar, Nabeel Zaidi, and Vijay Singh Rathore. 2018. Analysis and comparative exploration of elastic search, Mongodb and Hadoop big data processing. In *Proceedings of the Soft Computing: Theories and Applications SoCTA 2016, Volume 2*. Springer, 605–615.
[99] Makhan Kumbhkar, Pranjal Shukla, Yashwardhan Singh, Rohib Adrianto Sangia, and Dharmesh Dhabliya. 2023. Dimensional reduction method based on big data techniques for large scale data. In *Proceedings of the 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS'23)*. IEEE, 1–7.
[100] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya. 2016. The anatomy of big data computing. *Software: Practice and Experience* 46, 1 (2016), 79–105.
[101] Shalini Lambaa, Harsha Sahni, and Aanya Sharma. 2024. The art of data analysis: Review on essential, techniques and methodologies. *TEJAS Journal of Technologies and Humanitarian Science* 3, 2 (2024), 1–7.
[102] Addepalli Lavanya, Sakinam Sindhuja, Lokhande Gaurav, and Waqas Ali. 2023. A comprehensive review of data visualization tools: Features, strengths, and weaknesses. *International Journal of Computer Engineering in Research Trends* 10, 1 (2023), 10–20.
[103] George Lăzăroiu, Mihai Andronie, Mariana Iatagan, Marinela Geamănu, Roxana Ștefănescu, and Irina Dijmărescu. 2022. Deep learning-assisted smart process planning, robotic wireless sensor networks, and geospatial big data management algorithms in the internet of manufacturing things. *ISPRS International Journal of Geo-Information* 11, 5 (2022), 277.
[104] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of Massive Data Sets*. Cambridge University Press, USA.
[105] Wei Li, Yuanbo Chai, Fazlullah Khan, Syed Rooh Ullah Jan, Sahil Verma, Varun G. Menon, and Xingwang Li. 2021. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile Networks and Applications* 26, 1 (2021), 234–252.
[106] Zeng Li, Wenchao Huang, Yan Xiong, Siqi Ren, and Tuanfei Zhu. 2020. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm. *Knowledge-based Systems* 195, 9 (2020), 105694.
[107] Jie Lu, Anjin Liu, Yiliao Song, and Guangquan Zhang. 2020. Data-driven decision support under concept drift in streamed big data. *Complex and Intelligent Systems* 6, 1 (2020), 157–163.
[108] Yingcong Lu, Yipeng Liu, Zhen Long, Zhangxin Chen, and Ce Zhu. 2023. O-Minus decomposition for multi-view tensor subspace clustering. *IEEE Transactions on Artificial Intelligence* 5, 3 (2023), 1–14.
[109] Abdalwali Lutfi, Mahmaod Alrawad, Adi Alsyouf, Mohammed Amin Almaiah, Ahmad Al-Khasawneh, Akif Lutfi Al-Khasawneh, Ahmad Farhan Alshira'h, Malek Hamed Alshirah, Mohamed Saad, and Nahla Ibrahim. 2023. Drivers and impact of big data analytic adoption in the retail industry: A quantitative investigation applying structural equation modeling. *Journal of Retailing and Consumer Services* 70, 1 (2023), 103129.
[110] Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, and Minho Jo. 2017. Next-generation big data analytics: State-of-the-art, challenges, and future research topics. *IEEE Transactions on Industrial Informatics* 13, 4 (2017), 1891–1899.
[111] Alexandra L'heureux, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. 2017. Machine learning with big data: Challenges and approaches. *IEEE Access* 5 (2017), 7776–7797.
[112] Alexandra L'heureux, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. 2017. Machine learning with big data: Challenges and approaches. *IEEE Access* 5 (2017), 7776–7797.
[113] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. 2020. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Computers in Biology and Medicine* 126, 11 (2020), 104041.
[114] Samuel Mcmurray and Ali Hassan Sodhro. 2023. A study on ML-based software defect detection for security traceability in smart healthcare applications. *Sensors* 23, 7 (2023), 3470.
[115] Hassan Mehmood, Panos Kostakos, Marta Cortes, Theodoros Anagnostopoulos, Susanna Pirttikangas, and Ekaterina Gilman. 2021. Concept drift adaptation techniques in distributed environment for real-world data streams. *Smart Cities* 4, 1 (2021), 349–371.
[116] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. 2019. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports* 810, 25 (2019), 1–124.
[117] Azlinah Mohamed, Maryam Khanian Najafabadi, Yap Bee Wah, Ezzatul Akmal Kamaru Zaman, and Ruhaila Maskat. 2020. The state-of-the-art and taxonomy of big data analytics: View from new big data framework. *Artificial Intelligence Review* 53, 2 (2020), 989–1037.
[118] Azlinah Mohamed, Maryam Khanian Najafabadi, Yap Bee Wah, Ezzatul Akmal Kamaru Zaman, and Ruhaila Maskat. 2020. The state-of-the-art and taxonomy of big data analytics: View from new big data framework. *Artificial Intelligence Review* 53, 2 (2020), 989–1037.

[119] Luay Thamer Mohammed, AbdAllah A. AlHabshy, and Kamal A. ElDahshan. 2022. Big data visualization: A survey. In *Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA'22)*. 1–12. DOI: https://doi.org/10.1109/HORA55278.2022.9799819

[120] Luay Thamer Mohammed, AbdAllah A. AlHabshy, and Kamal A. ElDahshan. 2022. Big data visualization: A survey. In *Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA'22)*. IEEE, 1–12.

[121] Reza Mortaheb and Piotr Jankowski. 2023. Smart city re-imagined: City planning and GeoAI in the age of big data. *Journal of Urban Management* 12, 1 (2023), 4–15.

[122] Mohammad Moshawrab, Mehdi Adda, Abdenour Bouzouane, Hussein Ibrahim, and Ali Raad. 2023. Reviewing multimodal machine learning and its use in cardiovascular diseases detection. *Electronics* 12, 7 (2023), 1558.

[123] Oliver Müller, Iris Junglas, Jan vom Brocke, and Stefan Debortoli. 2016. Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems* 25, 4 (2016), 289–302.

[124] Mike Olson. 2010. Hadoop: Scalable, flexible data storage and analysis. *IQT Quart* 1, 3 (2010), 14–18.

[125] Amrit Pal, Abishi Chowdhury, Satakshi, Husnu S. Narman, Arkabandhu Chowdhury, and Manish Kumar. 2022. Random partition based adaptive distributed kernelized SVM for big data. *IEEE Access* 10 (2022), 95623–95637.

[126] Amrit Pal and Manish Kumar. 2018. Pattern mining for large distributed dataset: A parallel approach (PMLDD). *KSII Transactions on Internet and Information Systems* 12, 11 (2018), 5287–5303.

[127] Charles Parker. 2012. Unexpected challenges in large scale machine learning. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. 1–6.

[128] Andreas P. Plageras, Kostas E. Psannis, Christos Stergiou, Haoxiang Wang, and Brij B. Gupta. 2018. Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings. *Future Generation Computer Systems* 82, 9 (2018), 349–357.

[129] Shakthi Poornima and Mullur Pushpalatha. 2016. A journey from big data towards prescriptive analytics. *ARPN Journal of Engineering and Applied Sciences* 11, 19 (2016), 11465–11474.

[130] Mini Prince and P. M. Joe Prathap. 2023. An imbalanced dataset and class overlapping classification model for big data. *Computer Systems Science and Engineering* 44, 2 (2023), 1009–1024.

[131] M. Priyadharsini and K. Karuppasamy. 2023. Heterogeneous ensemble feature selection model (HEFSM) for big data analytics. *Computer Systems Science and Engineering* 45, 2 (2023), 2187–2205.

[132] Jun Qi, Po Yang, Lee Newcombe, Xiyang Peng, Yun Yang, and Zhong Zhao. 2020. An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure. *Information Fusion* 55, 3 (2020), 269–280.

[133] Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* 2, 1 (2014), 1–10.

[134] Hatice Catal Reis and Veysel Turk. 2023. Transfer learning approach and nucleus segmentation with medclnet colon cancer database. *Journal of Digital Imaging* 36, 1 (2023), 306–325.

[135] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. 2019. A review on data cleansing methods for big data. *Procedia Computer Science* 161, 16 (2019), 731–738.

[136] Uwe Röhm, Lexi Brent, Tim Dawborn, and Bryn Jeffries. 2020. SQL for data scientists: Designing SQL tutorials for scalable online teaching. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2989–2992.

[137] Emad Saddad, Ali El-Bastawissy, Hoda M. O. Mokhtar, and Maryam Hazman. 2020. Lake data warehouse architecture for big data solutions. *International Journal of Advanced Computer Science and Applications* 11, 8 (2020), 417–424.

[138] Houneida Sakly, Aline Sgnolf Ayres, Suely Fazio Ferraciolli, Claudia da Costa Leite, Felipe Kitamura, and Mourad Said. 2022. *Radiology, AI and Big Data: Challenges and Opportunities for Medical Imaging*. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-11199-0_3

[139] Arun Kumar Sangaiah, Samira Rezaei, Amir Javadpour, and Weizhe Zhang. 2023. Explainable AI in big data intelligence of community detection for digitalization e-healthcare services. *Applied Soft Computing* 136, 5 (2023), 110119.

[140] A Sasikumar, Logesh Ravi, Ketan Kotecha, Ajith Abraham, Malathi Devarajan, and Subramaniyaswamy Vairavasundaram. 2023. A secure big data storage framework based on blockchain consensus mechanism with flexible finality. *IEEE Access* 11 (2023), 56712–56725.

[141] Malte Schilling, Barbara Hammer, Frank W. Ohl, Helge J. Ritter, and Laurenz Wiskott. 2023. Modularity in nervous systems–a key to efficient adaptivity for deep reinforcement learning. *Cognitive Computation* 16, 5 (2023), 1–16.

[142] Rishika Shree, Tanupriya Choudhury, Subhash Chand Gupta, and Praveen Kumar. 2017. KAFKA: The modern platform for data management and analysis in big data domain. In *Proceedings of the 2017 2nd International Conference on Telecommunication and Networks (TEL-NET'17)*. IEEE, 1–5.

[143] Saurabh Shukla, Subhasis Thakur, Shahid Hussain, John G. Breslin, and Syed Muslim Jameel. 2021. Identification and authentication in healthcare internet-of-things using integrated fog computing based blockchain model. *Internet of Things* 15, 3 (2021), 100422.

[144] Shafaq Siddiqi, Faiza Qureshi, Stefanie Lindstaedt, and Roman Kern. 2023. Detecting outliers in Non-IID data: A systematic literature review. *IEEE Access* 11 (2023), 70333–70352.
[145] Gurpreet Singh, Jaspreet Singh, and Chander Prabha. 2022. Data visualization and its key fundamentals: A comprehensive survey. In *Proceedings of the 2022 7th International Conference on Communication and Electronics Systems (ICCES'22)*. IEEE, 1710–1714.
[146] Neelam Singh, Devesh Pratap Singh, and Bhasker Pant. 2022. Big data knowledge discovery as a service: Recent trends and challenges. *Wireless Personal Communications* 123, 2 (2022), 1789–1807.
[147] Rajesh Kumar Singh, Saurabh Agrawal, Abhishek Sahu, and Yigit Kazancoglu. 2023. Strategic issues of big data analytics applications for managing health-care sector: A systematic literature review and future research agenda. *The TQM Journal* 35, 1 (2023), 262–291.
[148] Tinku Singh, Riya Kalra, Suryanshi Mishra, and Manish Kumar. 2022. An efficient real-time stock prediction exploiting incremental learning and deep learning. *Evolving Systems* 14, 6 (2022), 1–19.
[149] Tinku Singh, Vinarm Rajput, Umesh Prasad, and Manish Kumar. 2023. Real-time traffic light violations using distributed streaming. *The Journal of Supercomputing* 79, 7 (2023), 7533–7559.
[150] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. 2017. Critical analysis of big data challenges and analytical methods. *Journal of Business Research* 70, 1 (2017), 263–286.
[151] Apache Spark. *Unified Engine for Large-Scale Data Analytics*. Retrieved August 25, 2023 from https://spark.apache.org/
[152] Statista.com. 2023. Retrieved 11th September 2023 from https://www.statista.com/statistics/254266/global-big-data-market-forecast/
[153] Christos L. Stergiou, Konstantinos E. Psannis, and Brij B. Gupta. 2021. InFeMo: Flexible big data management through a federated cloud system. *ACM Transactions on Internet Technology* 22, 2 (2021), 1–22.
[154] Alexander Y. Sun, Zhi Zhong, Hoonyoung Jeong, and Qian Yang. 2019. Building complex event processing capability for intelligent environmental monitoring. *Environmental Modelling and Software* 116, 6 (2019), 1–6.
[155] Guodao Sun, Zihao Zhu, Gefei Zhang, Chaoqing Xu, Yunchao Wang, Sujia Zhu, Baofeng Chang, and Ronghua Liang. 2023. Application of mathematical optimization in data visualization and visual analytics: A survey. *IEEE Transactions on Big Data* 9, 4 (2023), 1018–1037.
[156] C. Swarna and Zahid Ansari. 2017. Apache Pig-a data flow framework based on Hadoop Map Reduce. *International Journal of Engineering Trends and Technology* 50, 5 (2017), 271–275.
[157] Fatima Tahir and Muskan Khan. 2023. *Big Data: The Fuel for Machine Learning and AI Advancement*. Technical Report. EasyChair.
[158] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui, and Rachida Dssouli. 2021. Big data quality framework: A holistic approach to continuous quality management. *Journal of Big Data* 8, 1 (2021), 1–41.
[159] Leonidas Theodorakopoulos, Alexandra Theodoropoulou, and Yannis Stamatiou. 2024. A state-of-the-art review in big data management engineering: Real-life case studies, challenges, and future research directions. *Eng* 5, 3 (2024), 1266–1297.
[160] Jiao Tian, Pei-Wei Tsai, Kai Zhang, Xinyi Cai, Hongwang Xiao, Ke Yu, Wenyu Zhao, and Jinjun Chen. 2023. Synergetic focal loss for imbalanced classification in federated XGBoost. *IEEE Transactions on Artificial Intelligence* 5, 2 (2023), 1–13.
[161] J. Gregory Trafton, Laura M. Hiatt, Benjamin Brumback, and J. Malcolm McCurry. 2020. Using cognitive models to train big data models with small data. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1413–1421.
[162] Muhammad Fahim Uddin and Navarun Gupta. 2014. Seven V's of big data understanding big data to extract value. In *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. IEEE, 1–5.
[163] Ms V. Vaidehi and Viji Vinod. 2019. A survey on big data and deep learning: Methods, tools, applications, challenges and future trends. *Journal of Advanced Research in Dynamical and Control Systems* 11, 9 (2019), 988–992.
[164] Chunzhi Wang, Yichao Wang, Zhiwei Ye, Lingyu Yan, Wencheng Cai, and Shang Pan. 2018. Credit card fraud detection based on whale algorithm optimized BP neural network. In *Proceedings of the 2018 13th International Conference on Computer Science and Education (ICCSE'18)*. IEEE, 1–4.
[165] Hai Wang, Zeshui Xu, Hamido Fujita, and Shousheng Liu. 2016. Towards felicitous decision making: An overview on challenges and trends of big data. *Information Sciences* 367, 31 (2016), 747–765.
[166] Jin Wang, Yaqiong Yang, Tian Wang, R. Simon Sherratt, and Jingyu Zhang. 2020. Big data service architecture: A survey. *Journal of Internet Technology* 21, 2 (2020), 393–405.
[167] Xiaokang Wang, Wei Wang, Laurence T. Yang, Siwei Liao, Dexiang Yin, and M. Jamal Deen. 2018. A distributed HOSVD method with its incremental computation for big data in cyber-physical-social systems. *IEEE Transactions on Computational Social Systems* 5, 2 (2018), 481–492.
[168] Tom White. 2012. *Hadoop: The Definitive Guide*. "O'Reilly Media, Inc.", California, USA.

[169] Tarid Wongvorachan, Surina He, and Okan Bulut. 2023. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information* 14, 1 (2023), 54.
[170] Hongxin Wu, Meng Han, Zhiqiang Chen, Muhang Li, and Xilong Zhang. 2023. A weighted ensemble classification algorithm based on nearest neighbors for multi-label data stream. *ACM Transactions on Knowledge Discovery from Data* 17, 5 (2023), 1–21.
[171] Jianbin Wu, Sami Ahmed Haider, Manish Bhardwaj, Aditi Sharma, and Piyush Singhal. 2022. Blockchain-based data audit mechanism for integrity over big data environments. *Security and Communication Networks* 2022, 1 (2022), 1–9.
[172] Xuefang Xu, Yaguo Lei, and Zeda Li. 2019. An incorrect data detection method for big data cleaning of machinery condition monitoring. *IEEE Transactions on Industrial Electronics* 67, 3 (2019), 2326–2336.
[173] Ilkay Melek Yazici and Mehmet S. Aktas. 2023. A systematic literature review on data provenance visualization. In *Proceedings of the Computational Intelligence, Data Analytics and Applications: Selected Papers from the International Conference on Computing, Intelligence and Data Analytics (ICCIDA'23)*. Springer, 479–493.
[174] ChuanTao Yin, Zhang Xiong, Hui Chen, JingYuan Wang, Daven Cooper, and Bertrand David. 2015. A literature survey on smart cities. *Science China Information Sciences* 58, 10 (2015), 1–18.
[175] Youngjin Yoo. 2015. It is not about size: A further thought on big data. *Journal of Information Technology* 30, 1 (2015), 63–65.
[176] Dan Zhang, L. G. Pee, Shan L. Pan, and Lili Cui. 2022. Big data analytics, resource orchestration, and digital sustainability: A case study of smart city development. *Government Information Quarterly* 39, 1 (2022), 101626. DOI: https://doi.org/10.1016/j.giq.2021.101626
[177] Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li, and Kai Hwang. 2015. A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications. *Future Generation Computer Systems* 43, 2 (2015), 149–160.
[178] Jiaxin Zhang. 2021. Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics* 13, 5 (2021), e1539.
[179] Justin Zuopeng Zhang, Praveen Ranjan Srivastava, Dheeraj Sharma, and Prajwal Eachempati. 2021. Big data analytics and machine learning: A retrospective overview and bibliometric analysis. *Expert Systems with Applications* 184, 22 (2021), 115561.
[180] Zhan Zhang, Tianming Liu, Yanjun Shu, Siyuan Chen, and Xian Liu. 2023. Dynamic adaptive checkpoint mechanism for streaming applications based on reinforcement learning. In *Proceedings of the 2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS'23)*. IEEE, 538–545.
[181] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. 2017. Machine learning on big data: Opportunities and challenges. *Neurocomputing* 237, 19 (2017), 350–361.

Received 7 November 2023; revised 10 November 2024; accepted 3 February 2025