Recent Advances in Big Data: Features, Classification, Analytics, Research Challenges, and Future Trends

Khalid Mahmood[†], M. Rahmah[†], Muhammad Ahsan Raza[††], and Binish Raza[†††]
pcc16022.ump@gmail.com
[†] Faculty of Computing, Universiti Malaysia Pahang, Kuantan, Malaysia
[††] Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan
[†††] Department of Computer Science, Pakistan Institute of Engineering and Technology, Multan, Pakistan

Summary
Dealing with big data at present involves several procedures, including storage, processing, indexing, integration, and governance; thus, it poses a major challenge to application developers and research scientists. The challenge increases when accurate results must be obtained for queries in decision-making and cognitive applications regarding various disciplines and industries. A variety of literature on big data is available. However, big data must still be understood through various aspects, such as features, classification, analytics, research challenges, and future trends. This study provides a deep understanding of big data features, classification, and manipulation. An improved understanding of big data analytics and recent research challenges is also provided. Additionally, future trends are discussed for researchers. In conclusion, a vision to wisely and accurately improve information retrieval or search is presented to make better decisions using big data.

Key words:
Big Data, Big Data Classification, Big Data Challenges, Big Data Analytics, Semantic Search, Information Retrieval.

1. Introduction

At present, systems and humans are interacting with the web and other media by generating and extracting large-scale data, which grow continuously every day. Such data are measured in exabytes (1018) and zettabytes (1021). By 2025, Internet data storage size is predicted to exceed the brain capacity of all living creatures worldwide [4]. A National Security Agency report indicated that approximately 1.8 petabytes of data are collected on the Internet daily [24]. This fast and continuously growing data size has become possible due to developments in the Digital Age. Various devices and technologies, including communication, computation, sensor, and storage technologies, are involved in the creation and collection of huge amounts of data. The scale of data created and disseminated by public organizations, businesses, various nonprofit and industrial sectors, and scientific investigations has increased immeasurably [25]. As reported in [4], 2.5 billion gigabytes of data are being produced daily worldwide, including 90% unstructured data.

One research indicated that the estimated growth of global data volume from 2013 to 2020 is from 4.4 zettabytes to 44 zettabytes [5]. These data consist of text (unstructured, semi-structured, and structured data) and multimedia (images, audio, and videos) from multiple platforms, such as sensor networks, social media, cyber systems, machine-to-machine communication, and the Internet of Things (IoT). One of the popular terms at present is “big data,” and governments, businesses, science fields, societies, and industries are experiencing significant changes due to the impact of big data. Big data has become a major trend in various disciplines over the last few years. Big data is considered an influential raw material that can affect research in multiple disciplines. Big data exhibit several emerging potentials [27], and useful information must be extracted to facilitate decision-making and performance improvement by utilizing resources in an efficient and optimized manner [1, 24].

The remainder of this article is organized as follows. Big data and its characteristics, their relativity, the order of first occurrence, and related challenges are discussed in Section 2. After the definition, different classes of big data based on their characteristics are briefly described in Section 3. In Section 4, various big data analytics and their role in retrieving valuable information are deliberated. Section 5 provides a brief explanation of grouping big data challenges. Future trends related to the extraction of valuable information from big data are highlighted in Section 6. Section 7 concludes this study. Section 8 provides references related to the information included in this article.

2. Big Data

Big data has been defined in the literature by various authors in different contexts [23]. The following paragraph presents definitions of big data provided by various investigators. Interestingly, each definition of big

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

data focuses on features whose names start with the letter "V", such as volume, velocity, and variety. Thus, these interrelated and interdependent features have become known as 3 Vs, 7 Vs, and 10 Vs over time as new studies come into existence [42]. Fig. 1 shows the order of occurrence of each feature.

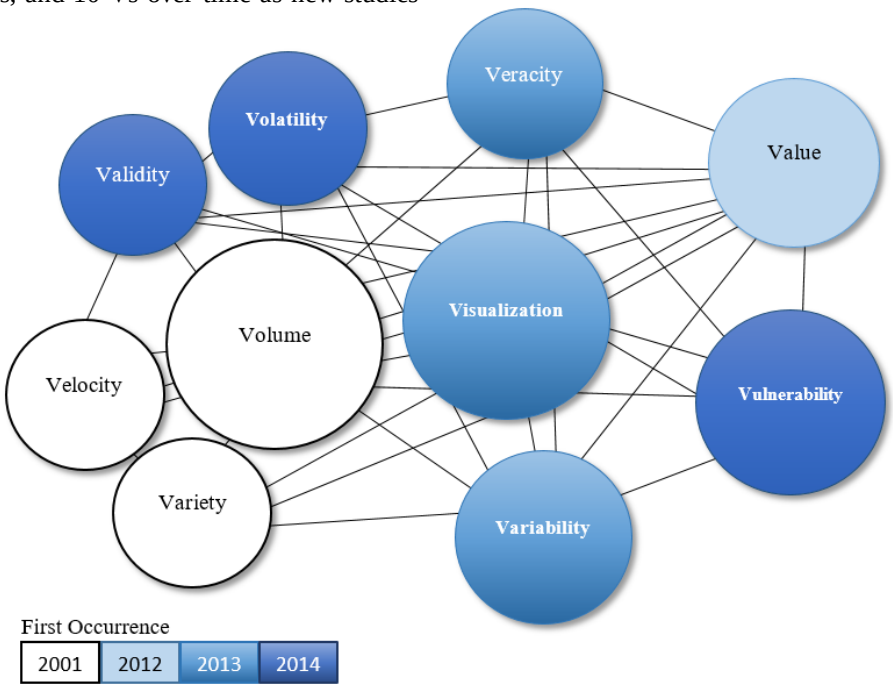

Fig. 1 Order of occurrence of 10 Vs of Big Data

Classic database software tools for collecting, storing, organizing, and evaluating data cannot handle big data because of its size [6, 8]. Traditional database tools, such as Structured Query Language (SQL), also cannot control and evaluate big data [7, 8, 9, 16]. Authors belonging to classical database and traditional database groups have focused only on the volume aspect of data when describing big data. Data that are excessively large, volatile, and cannot fit the database structure are considered big data [11]. This definition introduces another characteristics (i.e., variety and velocity) of big data. That is, big data is also defined by many researchers on the basis of huge volumes of data as comprehensively varied data that are created, collected, and processed at high velocity [16, 13, 14]. The aforementioned 3 Vs are additional features of big data that are mostly found in the literature. Various institutes, such as the Institute of Electrical and Electronics Engineers, and the authors of [6, 14] have focused on these 3 Vs. Among these three characteristics, visualization provides a considerably better tool for discovering data and investigation results. Other features are being introduced by various authors, such as the recently presented 10 Vs (as shown in Fig. 2) [30] Meanwhile, the researchers are working to sort out the problems being faced due to these features while developing big data applications and software tools [23, 4]. Nearly two decades before, the 3 Vs and 4 Vs [34, 40] of big data were introduced. Thereafter, 6 Vs, 7 Vs [20, 19], and then 10 Vs were presented [23]. Research continues to discover other big data features.

2.1 Volume

The growth of big data has been caused by an increase in storage capacity, computational power, and data size accessibility. At present, most technologies focus on different major issues, such as volume, to grasp big data problems [15, 25, 40]. Huge amounts of data are being continuously created and collected every minute by machines and humans worldwide. For example, 15 h of videos are uploaded to Facebook every minute; that is, Facebook collects more than 50 terabytes of data in 1 day [16].

Data volume has many sources of, including research studies, space images, medical networks, crime reports, server logs, broadcast audio/video streams, online banking transactions, music albums, website contents, scanned documents, and financial market data [13, 7]. Sensors

also collect and generate large data in various systems, such as weather forecasting system, natural disaster management system, Global Positioning System (GPS), and healthcare and environment analysis systems [10, 4, 7]. Corporations based on technologies, such as Yahoo, Microsoft, Google, and Amazon, preserve data in exabytes or even larger. Other popular online social media companies, such as Twitter and YouTube, collect massive amounts of data daily from billions of users. We now have an idea of the amount of data generated per day [31].

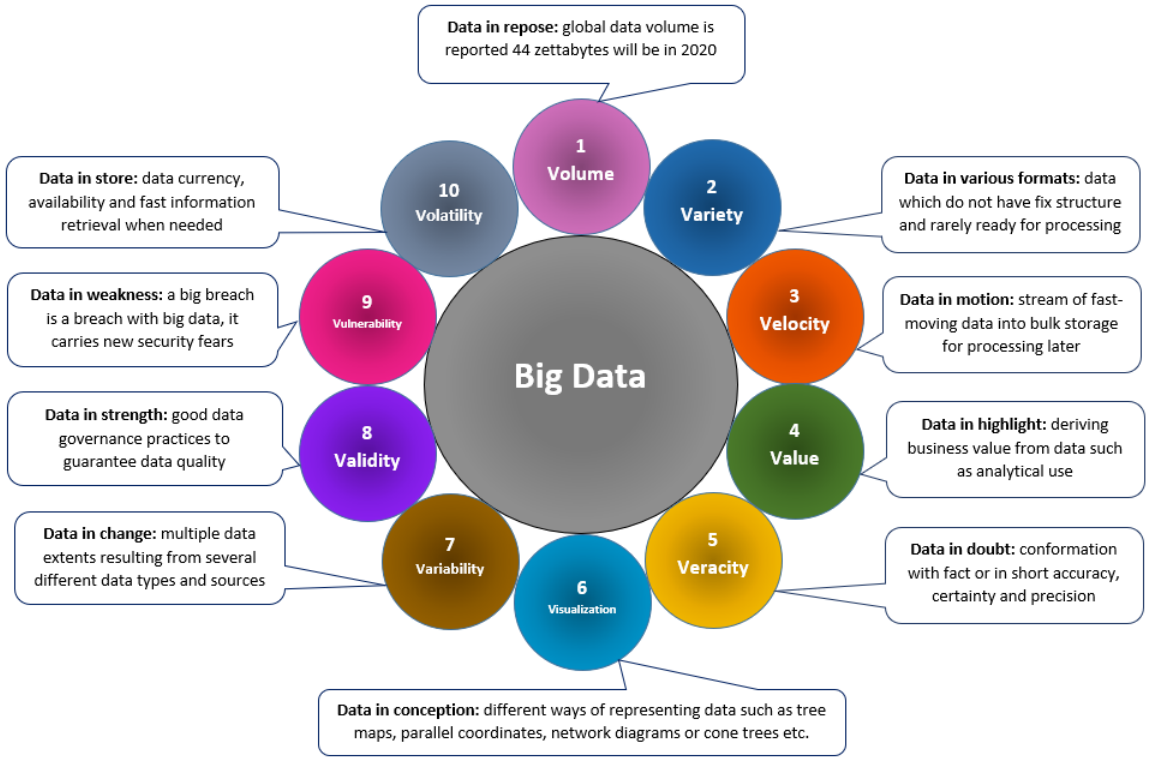

Fig. 2 Brief description of the 10 Vs of Big Data

We cannot define volume of data specifically on the basis of a predefined trend because a relative measure of data depends on the current state of an organization. Volume shows the magnitude of data, which may be considerably large or extremely big. Such large-scale data cannot be normalized in the traditional manner and are impossible to manipulate using an SQL-based methodology [15].
Among the major issues in mining big data, volume is the first one encountered by entrepreneurs and individuals; that is, various simple algorithms cannot deal with the large amount of data [3, 25]. Conventional tools cannot manage such bulk data. Therefore, specialized algorithms using novel approaches are required to address the problem of big data mining.

2.2 Variety

The second major feature of big data is variety, i.e., dealing with data that are available in various formats and from several sources. Variety refers to different types and formats of data, in addition to various ways of data usage [19]. Big data does not have a specific structure and is presented a perfect ordered form that is ready for processing [4]. Big data can be extremely structured, semi-structured, or unstructured.

Structured Data – data stored in relational databases.

Semi-structured Data – data from different sources, such as sensors, social media, news feeds, web logs, and emails.

Unstructured Data – data in video, audio, and image forms.

Variety is the most fascinating of all data features because it represents data of various types and modalities for a specified entity. Managing and organizing data in meaningful format are difficult tasks, particularly when data are available in different formats [3, 19]. Variety is

the only characteristic that has begun to scratch the surface of big data and is a crucial challenge in big data manipulation. Data variety has a significant impact on data integrity; that is, the higher the variety, the greater the inaccuracy. Various forms and qualities of big data represent its heterogeneous nature and result in difficulty in grasping and managing such data [35].

2.3 Velocity

Big data velocity is another prominent feature that deals with the speed at which data are streamed massively and continuously from various sources, such as social network websites, business processes, mobile devices, and machines. Data streaming is extremely fast and can easily halt system operation [5, 19]. Velocity also refers to the rate of data change and how data are being created frequently. The availability of data is also referred to as velocity by researchers.

Considering velocity, big data applications must deal with data quickly and process them rapidly [5, 4]. The opportunity to manipulate data in real time is a field that requires specific attention. It helps businesses in decision-making, such as presenting personalized advertisements on websites by relying on searches, visits, and transaction history. The large volume of big data can be attributed to the high velocity of data being generated from different sources [2, 4, 41].

High-velocity data must be handled by using advanced and optimized algorithms, and organizations must be equipped with advance technology and DB engines to manipulate data as required. Two important factors must be considered to deal with high-velocity data streams [2, 5]. First, special analysis algorithms are required to analyze data and store them upon arrival. Second, applications should be able to respond in real time upon receiving data.

2.4 Value

The most important feature of big data is value; all other features are meaningless if data do not present a business value [25]. Applications of big data must not only analyze data but also provide clear, concise, and descriptive results. In the discipline of big data, data and analysis are completely inter-reliant; that is, if one is missing, then the other is useless; thus, both are crucial for the success of a business [19, 33]. Big data can provide a business with enormous value if it is handled properly and every phase is accurately dealt with.

The value of big data enables accurate, efficient, and effective decision-making regarding opportunities and threats in a business. Significant value can be obtained from big data, such as improving the understanding of customers of a business, the subsequent pursuit of these customers, process improvements, and the optimization of business performance or machine [19]. Semantic and cognitive approaches are being introduced to meet and optimize such targets.

The value of big data depends on the techniques adopted for governance. Data governance basically involves the introduction of structures and policies to bring data into balance between risks and rewards [16]. If policies and structures are not carefully scripted, then they may result in the extraction of false data value. Researchers of big data consider value an important feature and that valuable information can be found within stored data, but its extraction is a challenging task [14, 33]. Therefore, systems face the problems of storing, managing, and extracting value from data in an efficient manner.

2.5 Veracity

Uncertainty in big data is another important and challenging issue. Data with uncertain veracity or doubt result from deception, model approximations, incompleteness, latency, inconsistency, and ambiguities [19, 41]. Although big data is important, it is useless when data are inaccurate. Business applications used for automated decision-making process and/or unsupervised machine learning algorithms face the problem of producing false results when the input data are inaccurate. Good results are only possible if the input data are reliable and accurate.

Normalization is implemented in relational databases to maintain data integrity and ensure no duplication in data. Similar to normalization, big data should be cleansed deeply by using efficient algorithms in applications [40, 42]. However, the source and utilization of data may also generate trusted data.

Veracity refers to the reliability of the data source, the data context, and the analysis that determines how important and meaningful data are. For example, consider a data set of the sales of a business, i.e., what items were sold to the customers and the prices of those items over the last 5 years [19, 20]. In this regard, we need answers to queries about data source generation, data collection methodology, certain types of businesses included, summary of information by the data generator, and whether the collected information is edited or modified [25].

To satisfy these queries, the veracity of data must be determined. Increased knowledge about the veracity of data will be helpful in understanding and avoiding risks related to decisions and analysis based on this specific data set. Erroneous and confusing data must be corrected; thus, new systems for monitoring and mining unreliable data

must be developed using advanced tools and intelligent analytics [2, 4].

2.6 Visualization

The presentation of data is commonly known as visualization. Many techniques can be used to present data, such as rows and columns (an Excel worksheet), text (a Word document), and graphs or charts. Regardless of the technique used, data should be easily understandable [19, 28]. Presenting data in graphs and charts is better than other formats; hence, visualization is an important feature.

Providing big data visualization with a good graphical view is a difficult task; however, using charts and graphs to present large-scale complex data is considerably better than visualizing such data in documents and worksheets. In big data visualization, users should not only depend on typical graphs when plotting massive data points [3, 41, 42]. Different techniques must be adopted to visualize big data, including the use of tree maps, circular network diagrams, cone trees, or parallel coordinates to represent data clustering.

2.7 Variability

Variability refers to the continuous and constant change in information or data meaning. Variability is also observed through data inconsistency. Lastly, variability is an essential feature of big data [42]. Multiple dimensions of data from various data types and sources make big data variable. The inconsistent speed of big data loading in data warehouses is another cause of variability. Given the changing nature of big data, variability must be considered in performing sentiment data analysis [41]. When performing semantic search, algorithms should be able to understand the context of search queries.

2.8 Validity

Data validity is similar to data veracity; however, the two concepts are not the same. Validity may mean that data should be clean, accurate, reliable, valid, and useful for later processing. If validity is considered, then data should be correct and accurate for their intended application [41]. Data may be invalid if not appropriately understood, but the same data may not exhibit veracity problems. The validity of the same data may also vary from application to application. Valid data are clearly a basic requirement for decision-making [19].
Users can benefit from big data analytics if and only if the primary data are good. Therefore, good data control practices must be implemented to guarantee reliable data value, metadata, and common definitions. In this regard, relationships must be verified, and relationships among data components that are being used must be defined to validate these data for their intended application as much as possible [15, 25].

2.9 Vulnerability

Data breach is a serious concern in the present era of advance technology. Hackers are continuously and constantly trying to breach systems and databases to steal information. Big data breach is a big breach, and thus, vulnerability is also a challenging and important feature of big data because securing information from unauthorized and unauthenticated access is a basic requirement [41, 42]. The three most important identified vulnerabilities that must be considered and addressed are big data security, privacy, and lack of standards.
Ensuring the security of big data from malware and other criminal activities while millions of devices remain connected through the Internet is a newly identified feature. Many applications are not developed with data security as a basic function, leading to several major big data security issues [26]. The following are major big data security problems.

Distributed Frameworks – Various big data applications perform distributed processing jobs across systems for efficient and fast data analysis. When more systems are involved in processing, security must be enhanced to secure data from illegal use.
Non-relational Data Warehouses – Databases, such as NoSQL, lack security, which may be provided by middleware.

Endpoints – Data are accessed from endpoints, requiring the authentication of endpoints for improved analysis.

Real-time Data Applications – Such applications generate huge amounts of data, requiring the identification and resolution of false positive issues.

Data Mining – The basic task in the big data environment; data mining solutions must provide security not only from external threats but also from internal threats given that the right to retrieve sensitive information must also be provided.

Access Control – Systems must have an encrypted user authentication or verification functionality to determine user permission or access levels.

Legal data processing and storage must be completely secure against unauthorized access and should be

considered critically. Big data manipulators must be concerned about the possible vulnerabilities of data and take preemptive strategies to implement security mechanisms [20, 26]. A strong authentication mechanism must be adopted, and data access channels should be secure across distributed environments. Smart big data analytic schemes can drive new business strategies; thus, they must also demonstrate strong security.

2.10 Volatility

A retention policy is mostly implemented to deal with structured data or relational databases in various data applications. A retention policy implies that data should be destroyed from the database when the retention period (i.e., how long the data should be stored) expires [41]. The duration of the relevance of data to current specific study is difficult to determine.

Thus, implementing a retention policy in big data is a challenging task. The retention period of big data may be too long. Therefore, additional storage and strong security are necessary, and maintenance expenses may increase considerably to meet challenges. Given the volume, velocity, and veracity of big data, volatility must be considered wisely and carefully [23, 42].

In general, these characteristics of big data are categorized into five groups [8]: 1 – collection of data, 2 – processing of data, 3 – data integrity, 4 – data presentation, and 5 – data worth. We add another important category that must be considered in the present era of advanced technology, i.e., 6 – data security. The characteristics, i.e., the 10 Vs of big data, are grouped in accordance with the relevant category as shown in Table 1.

Table 1: Categorization of Big Data features

Category	***Characteristics***
Data Collection	Variety, Veracity
Data Processing	Volume, Velocity
Data Integrity	Variability, Volatility, Validity
Data Presentation	Visualization
Data Security	Vulnerability
Data Worth	Value

Other parameters of big data are also considered by researchers and developers, namely, **availability** (data should be available whenever and wherever needed either in case of failure), **heterogeneity** (a concern in data variety), scalability (whether a system supports the efficient processing of large-scale data), **integrity** (refers to the veracity and validity of data), and **resource optimization** (efficient usability of available resources) [1].

3. Classification of Big Data

Big data is classified into different categories to understand its characteristics. Fig. 3 provides a brief overview of these categories. Big data classification is helpful in determining challenges and their solutions because such large-scale data manipulation is impossible in a single solution. Big data can be classified in accordance with these aspects: data source, content format, data stores, data staging, and data processing [12]. Each category exhibits complexities and characteristics, and some of them are defined as follows. Important classes and their branches are also discussed, and classification may provide elaboration.

3.1 Data Source

A site from where data originates in an application is known as a data source. The following are some of the most common data sources from which big data are generated continuously [12].

Social Media – It is the source of data creation via a Universal Resource Locator (URL) used for exchanging and sharing data through groups, communities, and trends over a network, such as Facebook, Twitter, and blogs.

Data Generation Machines – These devices generate data automatically without human intervention. Examples are computers, machine controllers, and medical devices.

Sensors – These devices sense physical quantities (e.g., blood pressure, speed, temperature, and humidity) and convert them into analogue or digital signals to process electronically.

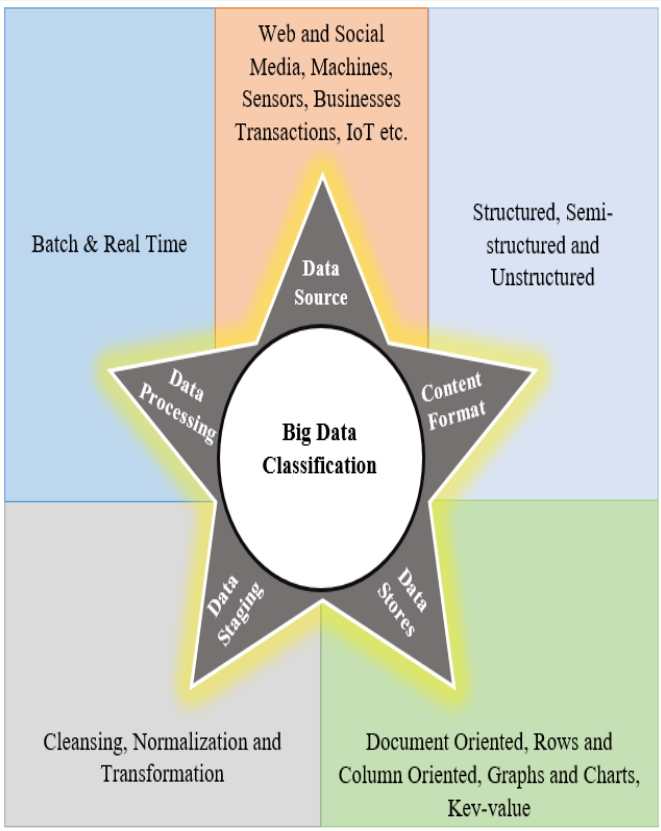

Fig. 3 Classification of Big Data

Business Transactions – These data are generated through financial activities and compromises from financial institutions.
IoT – It refers to a set of uniquely identifiable devices over the Internet, such as digital cameras, smartphones, and tablets. These devices enable smart services to support basic environmental, health, and economic needs [39].

3.2 Content Format

Content format is an encoded format used to convert a particular type of data into a displayable form of information. Numerous content formats (e.g., document file format, natural language format, audio data encoding, visual data encoding, expert language format, motion graphic encoding, and instruction encoding) can be stored and manipulated in various data formats, such as structured, semi-structured, and unstructured [19, 20].

3.3 Data Stores

Data stores are warehouses for storing, handling, and distributing data collections. They include simple databases, files, and emails. A data store is a broad term for any storage used for all types of data generated and used by organizations. Data stores are categorized on the basis of organizational needs [12].
Document-oriented – These data stores are designed to store and manage document collections consisting of complex data formats (e.g., XML and JSON) and binary data formats (e.g., MS Word and PDF).

Column-oriented – Such data stores are designed to store data in columns separately from rows.

Graph Database – Such databases are designed to store and represent data by using graph models along with nodes, edges, and properties related with one another.

Key Value – This substitute relational database management system stores and manages large-scale data.

3.4 Data Staging

Data staging is an intermediate data storage area for manipulating data before loading them to a data warehouse. However, data staging architectures are designed to hold data for only a specified time for troubleshooting or archival purpose. The following functionalities are performed in data staging [12].

Cleansing – The process of identifying and removing inaccurate or invalid data from source systems.

Consolidation – The merging of data from numerous source systems is one of the primary functions of data staging.

Transformation – The process of converting data into a suitable form such that data can be analyzed accordingly.

Normalization – The process of database structuring to reduce redundancy.

The heterogeneous nature of big data is a major issue that causes research and implementation problems relevant to data staging. That is, a variety of data collected from different sources cannot be presented in a structured format [14, 15]. Cleaning and converting such unstructured data are challenging issues to enable further processing of data.

3.5 Data Processing

The manipulation of data to yield significant information is commonly considered the processing of data. The conversion or manipulation of data is performed by

computers by using sequential operations. The data results are processed in different desired formats, such as graphs, images, text, audio, video, and tables. Data collection, validation, aggregation, storage, sorting, analysis, and presentation are some stages involved in data processing. The following are some types of data processing [12].

Batch Processing – a well-organized approach for manipulating large-scale data wherein a pool of data is collected over valuable periods.

Real-time Processing – real-time processing or manipulation of data that comprises persistent input, manipulation, and data output over a short period.

The data processing cycle must consist of an ordered series of interdependent steps to retrieve valuable information from raw collections of data. The storage and result production phases may repeat in multiple data processing cycles; therefore, this management is also challenged to deal with bulk data sets [1, 15]. The generated results during processing cycles must be stored temporarily or permanently because they play an important role in effective decision-making.

4. Big Data Analytics

The success of any organization or system is always attributed to effective decision-making, which is only possible if the input data (facts) are meaningful. Valuable or meaningful information retrieval from large volumes of data is important for decision-making [2, 25]. A large variety of big data potentials exists, but the lack of technologies, skills, tools, and efficient algorithms restricts the application of big data. Big data analytics are techniques or methods for observing and extracting valuable information from large-scale datasets [6, 11]. Big data analytics are facing two basic challenges: 1) the overwhelming requirements for storage and processing regarding the volume, variety, and velocity of big data; and 2) the complex analytic techniques and algorithms that make big data analysis an intensive task [28, 6]. An analytic can be a sub-analytic belonging to the entire procedure, as depicted in Fig. 4, of information extraction from big data.

Organizations collect data regularly and introduce analytics to support effective decision-making. The best analytical method selection is evidently beneficial for highlighting the support of big data in prolific decision-making [22]. In [31], big data analytics were classified into various divisions, namely, descriptive, inquisitive, predictive, prescriptive, and pre-emptive analytics.

4.1 Descriptive Analytics

The simplest form of big data analytics is descriptive. Descriptive analytics are used to summarize and describe knowledge patterns by using simple statistical techniques, such as the mean, median, mode, standard deviation, variance, and frequency measurements of particular occurrences in big data. Descriptive analytics also consider backward observation and find occurrences [2].

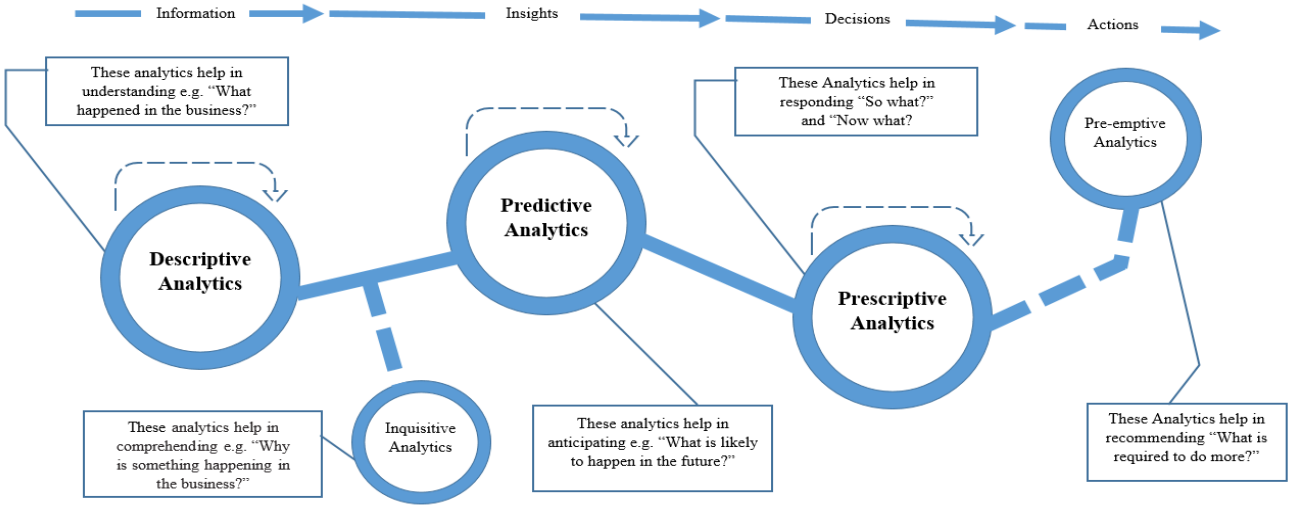

Fig. 4 Classification of Big Data Analytic

The results of descriptive analytics are used to find outcomes of predictive analytics, such as forecasting future trends. Descriptive analytics indicate current positions on the basis of data in a pattern and/or exception in the form of reports and alerts [6].

4.2 Predictive Analytics

Predictive analytics are concerned with predicting and statistical modeling to define upcoming possibilities based on learning models (supervised, unsupervised, and semi-supervised). These analytics are mostly based on statistical techniques and explore patterns and find relations in data [2]. Predictive analytics are divided into two categories: 1) regression and 2) machine learning methods. Lastly, predictive analytics forecast future possibilities on the basis of current and historical data analysis [6].

4.3 Prescriptive Analytics

Prescriptive analytics are used for process optimization on the basis of the results provided by predictive analytics. Prescriptive analytics typically define the causal relationships between analytical results and policies [6]. Prescriptive analytics contribute to handling information instability and evaluating business process models continuously [2]. In general, prescriptive analytics help analysts in decision-making after the determination of actions and their effects on objectives, constraints, and requirements.

5. Research Challenges

Many schemes and solutions have already been introduced to address the challenges of big data in different domains and directions. Despite the various available solutions, existing and new challenges must still be addressed efficiently [3]. The size and complexity of large-scale datasets and the capability to process and manage such datasets remain critical challenges. Challenges regarding big data are grouped as shown in Fig. 5. In [31], the challenges of big data characteristics, lack of management skills, and limitations of data processing systems are grouped together.

5.1 Data Challenges

A group of challenges relevant to the characteristics of big data [2, 36], including volume, velocity, variety, veracity, visualization, validity, volatility, vulnerability, variability, and value, as a set of two or more challenges and as individual challenges. Some challenges are highlighted along with the discussion of each characteristic in the big data section.

5.2 Process Challenges

Challenges regarding data processing and analysis starting from data collection to the presentation and interpretation of results belong to this group. Big data is nearly non-relational and unstructured, and dealing with such semi-structured large datasets is a significant issue [3]. Process challenges refer to a series of how procedures: how to collect data, how to incorporate data, how to convert data, how to select accurate model, and how to present results. The review of various articles indicates that process challenges are classified into five steps [31].

1) Data Acquisition and Warehousing – refer to the acquisition of data from various sources and data storage for valuable information generation and further processing. Related and valuable information collection and keen, intelligent, and robust filtration algorithms are required and must be able to eradicate inconsistencies from data [15]. Efficient analytical schemes are necessary to comprehend the background of data, and compression techniques are needed to efficiently utilize storage space.

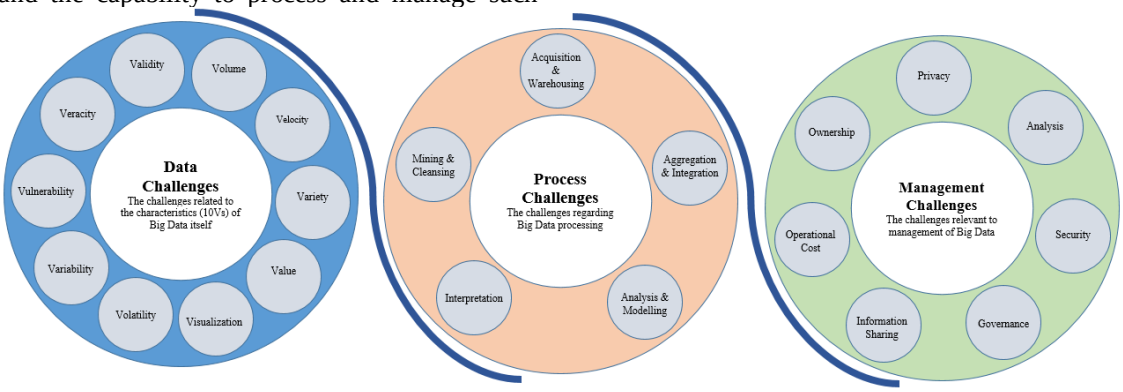

Fig. 5 Grouping of Big Data challenges

2) Data Mining and Cleansing – are relevant for data retrieval from stored large-scale data and cleansing or refinement for subsequent procedures. To convert large-scale data in usable form, schemes that extract desired information and transform it in a standard format that is easy to comprehend are crucial [3, 40]. Developing such extractions schemes is a continuous challenge.

3) Data Aggregation and Integration – In this step, mined and cleaned data with diverse meanings and representations are aggregated and integrated for collaboration and good decision-making. This procedure is a major challenge. Data provenance and uncertainty are also key challenges in data aggregation and integration [3, 37]. Clean operational data should be stored for real-time decision-making systems, and the development of such systems is currently a serious issue.

4) Data Analysis and Modeling – After capturing, storing, mining, cleansing, aggregating, and integrating data, the analysis and modeling of big data follow [21]. Schemes for data analysis and modeling must be introduced to obtain valuable results. Other efficient and intelligent schemes are also required to be able to forecast future trends after analysis and modeling [40].

5) Data Interpretation – Presenting results from the data analysis and modeling phase to decision makers to find knowledge for interpretation is a huge challenge. Another issue is introducing scientific schemes to allow access, aggregate, analyze, and interpret big data [15, 21]. The lack of skilled professionals with analytical skills to interpret data is another issue.

5.3 Management Challenges

This group of challenges related to big data comprises data access, management, and governance. Schemes for ensuring privacy, strong security infrastructure, improved governance, information sharing, operational cost computation, and ownership of big data are urgently required [3].

The major concern in big data is privacy and how to maintain privacy in this era of advanced technologies. Several schemes have been introduced, but privacy processes must still be streamlined by hiring skilled data analysts [8, 15, 26]. The analysis of logs, network streams, and past events for intrusion detection and forensics is a huge challenge in the security of big data. Less refined infrastructure ensures the security of data in terms of confidentiality, availability, integrity, and accountability; and providing such infrastructure for the secure manipulation of big data is challenging for researchers and developers [8, 37].

Describing and supervising which data should be stored, retrieved, and analyzed are key challenges in the infrastructure of data centers and the provision of data scalability [15, 21]. Big data categorization, mapping, and modeling upon collection and storage are also governance challenges [32]. The issue of mining and analyzing the quality of data from large repositories of unstructured and complex data must be addressed [20, 40].

Several departments with distributed data warehousing based on different technologies, vendors, and platforms are reluctant to share sensitive information under privacy conditions. Integrating and sharing key data and information in distant systems are also challenges [37, 38]. A variety of data intensive operations that require large storage spaces and extraordinary computational resources exist, incurring high costs for data storage and processing [15, 18]. Therefore, cost optimization is a huge challenge. Another critical issue in manipulating large-scale and real-time data is ownership. Big data applications must be able to identify who owns the data [31, 32], but several application do not focus on this concern.

6. Future Trends

At present, the efficient utilization and management of big data are only possible through the retrieval of useful and valuable abstract information or knowledge. Semantic indexing is an open and future challenge in dealing with big data, with the central goal of retrieving better and valuable information. Semantic indexing techniques, instead of storing raw data, can perform better in the efficient utilization of storage and the semantic retrieval of valuable and abstract information, which will evidently be helpful in making better decisions. This abstract data representation may also be used to improve indexing techniques [22, 29]. The development of semantic search techniques is a huge challenge in introducing next-generation search engines [17].

Various open research trends exist in the discipline of big data, including better decision-making schemes, minimization of operational resistance, real-time data access and manipulation, and data overlapping. Another future trend is the development of knowledge based on inter-data similarity techniques in the context of semantic data or information mining [9]. The data of information fusion are integral and essential fragments of the IoT. Big data exhibits a heterogeneous and dynamic nature, leading to single-source analysis techniques. Data fusion is the only solution that can provide integration of numerous data and knowledge into an accurate, consistent,

and useful depiction to extract valuable information for reliable decision-making support [39]. Thus, exploring new techniques for the resolution of fusion problems in IoT is important.

7. Conclusion

A common goal of all big data schemes is to extract useful and valuable information to satisfy business requirements. The development and production of such schemes or analytics have a crucial need for understanding big data features to add to the effective ability to store, discover, access, and retrieve usable information from large-scale data. The schemes that comprise big data analytics must provide all technical aspects of collection, processing, integrity, presentation, value, and security. Big data is extremely helpful in the practice of making strategic, operational, and pre-emptive decisions. Considerable potentials in the utilization of big data can be identified in several areas, including healthcare, agriculture, finance, transport, retail, media, entertainment, manufacturing, energy, and public organizations. However, despite the various potentials of big data, vulnerabilities regarding security, confidentiality, and normalization remain. In the future, we plan to enhance the study of big data spanning technologies and potentials in detail to highlight huge issues. We are also going to conduct a detailed survey of big data security challenges and introduce competent schemes to address the issues of semantic indexing and valuable information retrieval.

References

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

[2] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35(2), pp. 137-144, 2015.

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

[4] C. K. Emani, N. Cullot, C. Nicolle, "Understandable Big Data: A survey", Computer Science Review, vol. 17, pp. 70-81, Jun. 2015.

[5] D. Brent, "Big Data: Forget Volume and Variety, Focus On Velocity" Forbes, 28 June 2017.

[6] D. Maltby, "Big data analytics," in 74th Annual Meeting of the Association for Information Science and Technology (ASIST), pp. 1–6, 2011.

[7] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," Journal of Internet Services and Applications, vol. 6(1), pp. 1-15, 2015.

[8] F. Almeida, "Big Data: Concept, Potentialities and Vulnerabilities", Emerging Science Journal Vol. 2, February, 2018.

[9] F. Benedetti, D. Beneventano, S. Bergamaschi, G. Simonini, "Computing inter-document similarity with Context Semantic Analysis," Information Systems, vol. 80, pp. 136-147, Feb. 2019.

[10] H. He, Z. Du, W. Zhang and A. Chen, "Optimization strategy of Hadoop small file storage for big data in healthcare," J. Supercomput., pp 1-12, 2015.

[11] H. Hu, Y. Wen, T. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial", IEEE Access, vol. 2, pp. 652–687, 2014.

[12] I. Abaker, T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues", Information Systems, vol. 47, pp. 98-115, 2015.

[13] IDC, Digital Universe in 2020.

[14] J. Abawajy, "Comprehensive analysis of big data variety landscape," International Journal of Parallel, Emergent and Distributed Systems, vol. 30(1), pp. 5-14, 2015.

[15] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou. "Big data challenge: a data management perspective," Frontiers of Computer Science, vol. 7(2), pp. 157–164, 2013.

[16] J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big Data for Dummies, ISBN: 978-1-118-50422-2

[17] K. Hong and H. Kim, "A Semantic Search Technique with Wikipedia-based Text Representation Model," in International Conference on Big Data and Smart Computing (BigComp), IEEE, Hong Kong, China, 18-20 Jan. 2016.

[18] L. Gu, D. Zeng, P. Li, and S. Guo, "Cost minimization for big data processing in geo distributed data centers," Cloud Networking for Big Data, pp. 59-78, 2015.

[19] M. A. Khan, M. FahimUddin and N. Gupta, "Seven V's of Big Data Understanding Big Data to extract Value," in Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3-5 April, IEEE 2014.

[20] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, "A Survey on Deep Learning in Big Data," in IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017.

[21] M. M. Najafabadi, F. Villanustre, T. M Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," Journal of Big Data, vol. 2(1), 2015.

[22] N. Elgendy and A. Elragal, "Big Data analytics in support of the decision making process," Procedia Computer Science, vol. 100, pp. 1071-1084, 2016.

[23] O'Reilly Media, "Big Data Now: 2016 Edition," O'Reilly Media, 2017.

[24] Q. Zhanga, L. T. Yang, Z. Chenc and P. Li, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146-157, 2018.

[25] R. Agarwal, and V. Dhar, "Editorial – big data, data science, and analytics: the opportunity and challenge for is research", Information Systems Research, vol. 25(3), pp. 443–448, 2014.
[26] R. Krishnamurthy, and K. C. Desouza, "Big data analytics: the case of the social security administration," Information Polity, vol. 19(3/4), pp. 165-178, 2014.
[27] R. Rialti, G. Marzi, C. Ciappei and D. Busso, "Big data and dynamic capabilities: a bibliometric analysis and systematic literature review", Management Decision, 2019, https:// doi.org/10.1108/MD-07-2018-0821
[28] S. Khan, X. Liu, K. A. Shakil and M. Alam, "A survey on scholarly data: From big data perspective," Information Processing and Management, vol. 53(4), pp. 923-944, 2017.
[29] T. A. Letsche and M. W. Berry, "Large-scale information retrieval with latent semantic indexing," Information Sciences, vol. 100(1-4), pp. 105–137, 1997.
[30] U. Sivarajah, Z. Irani, and V. Weerakkody, "Evaluating the use and impact of Web 2.0 technologies in local government," Government Information Quarterly, vol. 32(4), pp. 473-487, 2015.
[31] U. Sivarajah, M. M. Kamal, Z. Irani and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods" Journal of Business Research, vol. 70, pp. 263-286, Jan. 2017.
[32] V. Khatri and C. V. Brownk, "Designing data governance," Communications of the ACM, vol. 53(1), pp. 148–152, 2010.
[33] W. A. Günther, M. H. R. Mehrizi, M. Huysman and F. Feldberg, "Debating big data: A literature review on realizing value from big data," The Journal of Strategic Information Systems, vol. 26(3), pp. 191-209, Sep. 2017.
[34] W. Wang, J. Liu, F. Xia and I. King, "Shifu: Deep Learning Based Advisor-advisee Relationship Mining in Scholarly Big Data," in Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 03 - 07, 2017. pp. 303-310.
[35] W. Yu, J. Li, M. Z. A. Bhuiyan, R. Zhang, and J. Huai, "Ring: Realtime emerging anomaly monitoring system over text streams," IEEE Transactions on Big Data, pp. 99, 2017.
[36] X. Jin, B. W. Wah, X. Cheng and Y. Wang, "Significance and challenges of big data research," Big Data Research, vol. 2(2), pp. 59–64, 2015.
[37] X. L. Dong and D. Srivastava, "Big data integration," In Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pp. 1245–1248. IEEE, 2013.
[38] Z. Irani, A. Sharif, M. M. Kamal, and P. E. Love, "Visualising a knowledge mapping of information systems investment evaluation," Expert Systems with Applications, vol. 41(1), pp. 105-125, 2014.
[39] Z. Yan, J. Liu, L. T. Yang and N. Chawla, "Big Data Fusion in Internet of Things," Information Fusion, 2017, doi: 10.1016/j.inffus.2017.04.005
[40] Z. Zhi-Hua, N. V. Chawla, J. Yaochu, G. J. Williams, "Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives," IEEE Computational Intelligent Magazine, 2014.
[41] https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
[42] https://www.elderresearch.com/blog/42-v-of-big-data, April 1, 2017.