Big Data: Big Data Analysis, Issues and Challenges and Technologies

R Rawat¹ and R Yadav²

¹ Computer Science and Engineering Department, Graphic Era Deemed to be University, Dehradun 248002, Uttarakhand, India
² Computer Science and Engineering Department, Graphic Era Hill University, Dehradun 248002, Uttarakhand, India

E-mail: ryrishikayadav@gmail.com

Abstract. The data generated at an exponential rate has resulted in Big Data. This data has many characteristics and consists of structured, unstructured, and semi-structured data formats. It contains valuable information for the different types of stakeholders based on their need however it is not possible to meet them with the help of traditional tools and techniques. Here the big data technologies play a crucial role to handle, store, and process this tremendous amount of data in real-time. Big data analytics is used to extract meaningful information or patterns from the voluminous data. It can be further divided into three types i.e. text analytics, audio analytics, video analytics, and social media analytics. Big data analytics if followed by big data analysis process plays a significant role in generating meaningful information from big data. Big data analysis process consists of data acquisition, data storage, data management, data analytics, and finally data visualization. However, it is not simple and brings many challenges that need to be resolved. This paper presents the issues and challenges related to big data, prominent characteristics of big data, big data analytics, big data analysis process, and technologies used for processing the massive data.

Keywords: — big data, big data analytics, big data processing, big data processing technologies, big data analysis

1. Introduction

The evolution of technology is taking place at a fast pace, and knowingly or unknowingly everyone is generating data in some or the other way. Different areas are producing data at an unprecedented rate such as healthcare data, social media data, sensor data, phone log data, server log data, stock market, etc. The formation of this massive amount of data through divergent sources leads to high volume, high velocity, and a variety of data that led to the coining of the term big data. It is the data to which the existing data mining techniques cannot be applied well or where the existing data mining techniques have failed to perform. Big data has some striking characteristics called 5V’s that are high volume (Data Size), velocity (Data Speed), variety (Variations in Data), veracity (the reliability and data authenticity), and value (gain from the dataset). Although many other different characteristics like variability, volatility, the validity of big data have also been highlighted by the researchers in this paper, we have
only focused on the 5 most prominent characteristics. These characteristics bring many different types of challenges in handling, processing, and analysing big data. Therefore one machine cannot store and process this data that are required for further analysis and used for decision making and thus this problem can be overcome employing using large clusters and distributing the data over these large clusters, for example, the Google's data centre which has thousands of machines [1].

The internet processes 1826 petabytes (PB) of data every day according to the National Security Agency [2]. 2.5 quintillion bytes of data were being generated every single day in the year 2018. Recently 90% of the big data was generated in the past two years. Google search engine alone accounts for 77% of the 5 billion searches performed over the internet in a single day [3]. Facebook is one of the largest social media networks where approximately 1.5 billion users are actively browsing or uploading data every day with uploading up to 300 million photos, posting 5,10,000 comments, and around 2,93,000 status updates per day [4]. Therefore, more advanced techniques or methods are required to manage, process, and generate valuable information from this massive amount of data [5].

Big Data can be categorized into either semi-structured, structure or unstructured class. Structured data is in the form of a schema, having rows and columns which can be generated through applications such as CRM (Customer Relationship Management), and Enterprises, etc. Semi-structured data generally have meta-data which describe its structure. This type of data is generated through the sensors, web feeds, network, and security systems. There is no surety that this data will be present in the format of rows and columns. Finally, unstructured data consists of text, audio, video, images, etc. it is generated by the people. Approximately 95% data is present in the form of raw data (not in any form) which brings many challenges to businesses and enterprises.

With variety and volume as the characteristics of big data, it is generated swiftly with no means of knowing about the veracity of the deposited data. This is the reason that the traditional systems are incapable of storing and processing this massive data [6]. The five characteristics of big data make it makes extremely difficult to perform data analysis [7,8].

Big data analysis is the procedure of analysing the massive datasets to uncover the unidentified relationships and associations, market trends, and valuable information. Researchers are focusing on rising real technologies for analysing big data and generating valuable information from it. This paper discusses the big data analytics techniques such as social media analytics, text analytics, video, audio analytics, the various issues and challenges related to the data, processing, and management and also throw light on batch-processing and stream processing and the technologies used in each processing and the combination of both.

2. Big Data Analytics

The evolution of big data has resulted in the analysis of not only a high volume of data but myriad types of data which is usually streaming in nature. The unique features of big data such as the high dimension, heterogeneous, unstructured, incomplete, erroneous, noisy shows that the data need to be polished from the different perspectives such as volume, velocity, etc. Else we can lose the effectiveness of big data using traditional tools. Therefore, to make the process big data analysis more accurate, fast, and precise several innovative tools and techniques have been incorporated and practiced such as natural language processing, deep learning, artificial intelligence, machine learning, etc. These tools and techniques discover the hidden pattern, find the unknown-correlation, and extract meaningful information from the data deluge. For example, the analysis of the data may show where the housing price is high in a city, the analysis of the patient report may show an early detection of a disease in a patient and a timely decision can be taken at an early stage [9], sales trends helps the management making better policies to retain the customers in business [10]. The uncertainty brings new challenges to the data, it is the situation when the data has unknown and imperfect data that are present in all the phases of the big data analytics process. Big data analytics may adversely impact the result of the data in terms of effectiveness and accuracy in output data. The techniques used for analysing the data are as follows.
2.1. Text Analytics
The text mining process is used to extract eloquent facts or figures from unstructured data like the blogs, corporate documents, and online forums, etc.

2.2. Audio Analytics
It is the process of dig out the audio data from the unstructured data. The widely used applications of audio data are smart speakers, applications in the healthcare sector for patient monitoring, customer-care centres, etc.

2.3. Social Media Analytics
It consists of the data from various online platforms such as Facebook, LinkedIn, blogs, micro-blogs, Instagram owned by Facebook, YouTube, etc. It is the analysis of data present in various forms. Social graphs and activity graphs are the two types of graphs that form the structure of a social network [11].

2.4 Predictive Analytics
It makes use of historical data and current data for predicting future outcomes. Predictive analytics can be employed to all sectors such as the market basket customer behaviour, churn of employees from the company, etc. The predicting techniques are SVM, Neural Networks, decision tree, and linear regression.

3. Issues and Challenges
Challenges in big data can be broadly alienated into three types the first type is data challenges, the second type is data process challenges, and the third type are data management. Data challenges are the challenges that are associated with the characteristics of big data. Process challenges are the challenges that faced during the processing of data whereas management challenges pertaining to tackling the data such as providing security. The characteristics of big data bring many challenges to it such as its high volume, variety, etc. Process challenges are related to data acquisition, pre-processing, data analysis, and data visualization whereas management challenges are related to privacy and security. Figure 1 shows the different types of challenges associated with the different phases of the big data analysis process.

3.1. Data Challenges
Researchers have given many definitions of big data and based on their understanding towards they come up with several new characteristics of big data. [12] researchers discussed the 3V’s characteristics of data (Volume, Variety, and Volume), 4th V was introduced by IBM as veracity [13], researchers have discussed 5th and 6th V’s as variability, and value. The 10 V’s of big data are taken under consideration [14], there are many worth mentioning prominent challenges associated with the characteristics of data. Some of the prominent challenges are discussed as follow.

3.1.1. Volume Challenges. The unprecedented increase in data through internal and external sources has resulted in a massive amount of data. This high volume of data brings the challenges to the data itself such as the storage of the data for processing is not possible through traditional tools and thus more innovative methods should be developed to handle this deluge[15].

3.1.2. Variety Challenges. The challenge associated with variety is related to its different forms. The massive data can be present in the form of structured, semi-structured, and unstructured. Research studies show that 95% of the data is present in unstructured form. Therefore, converting it into a form so that the analysis can be performed is a big challenge.

3.1.3. Velocity Challenges. Velocity indicates the speed of the data generated through the devices. Data can be processed in two ways batch processing and real-time processing. In batch processing, the data is stored and then processed whereas real-time processing is continuous. In online shopping, real-time processing is required to generate value for customers.
3.1.4. Veracity Challenges. Data veracity indicates the quality and accuracy of data. It deals with the fabrications, imprecision, messiness, and misplaced evidence in data. It defines the trustworthiness of data when a significant decision needs to be taken. In social networking sites, user opinion can be classified as positive, negative, or neutral.

3.1.5. Value Challenges. Value is one of the most significant features of big data characteristics. Big data contains valuable information that needs to be extracted from the large datasets. This brings a big challenge to data as extracting the high information from data in a cost-effective manner and making use of it for business intelligence, health sectors, etc.

3.2 Process Challenges
Process challenges are related to processing and analyzing large datasets. It brings a significant challenge to the process as the data is present in different forms and conversion of it into one form for analysis purpose is a challenging task. It can be divided into four parts: Data Acquisition and Storage, Data Preprocessing, Data Analysis and Modeling, Data Visualization.

3.2.1. Data Acquisition and Storage. Data acquisition is the process of acquiring and storing the data for the future utilizing some valuable information. The data is acquired from various sources such as from sensors, social networking sites, blogs, etc. and hence the data is present in different forms (structured, semi-structured, and unstructured) bring a significant challenge to data. The second challenge is associated with the storage because the data generated through various devices does not mean that whole data carry meaning to it therefore the smart filter must be applied for generating the relevant datasets. Storing this massive dataset can result in high-cost scalable systems to handle the data.

3.2.2. Data Preprocessing. Data preprocessing is the process to collect the quality data from large datasets as low-quality data leads to low-quality knowledge. Therefore, data preprocessing plays a significant role in knowledge discovery. In this stage noise, missing values, inconsistent and superfluous data, etc. are removed before applying the big data mining techniques to its data [16]. In big data preprocessing most of the efforts are done in the Feature Selection method whereas some of the families of it are ignored such for instance reduction, missing value imputations, noise treatments.

3.2.3. Data Analysis and Modelling. Data analysis is the process that discovers the hidden information from the data and helps the organizations to make a better decision. To efficiently extract the knowledge from the large datasets extraordinary techniques are required. To generate the hidden pattern from the large datasets Wal-Marts employs statistical and machine learning techniques.

3.2.4. Data Visualization. A big data visualization technique presents the analytical data visual form. It makes usage of various types of graph for representing the valuable information for decision making[17]. As per the research studies, the visual report has a better impact on information seeker rather than the text reports. Visualization tools like Tableau and Qlik View are the tools used for visualization however according to the researcher these tools cannot be fruitful shortly where data is growing every second by each one of us.

3.3 Management Challenges
Management challenges are related to those challenges encountered by an organization which is related to the privacy, security, governance of data. Management challenges are also faced because we have a lack of data for skilled professionals who know the latest tool and techniques to employ the correct method for dealing with each phase of data. Security and privacy will always be the major concerns as data are highly sensitive such as financial data, military data, insurance codes and contains different kinds of information that can ruin if the unauthorized user has access to it.
4. Big Data Processing Technologies

Before the invention of the big data processing technologies, the companies were not able to capture and store the massive datasets. Even the invented processing tools are also not capable of generating the complete results within a period however it has shown a great performance in different sectors such as building the business models, decision making, etc. Minimizing the hardware cost, reducing processing costs and value generations are the aims of these technologies.

The data processing techniques can be divided into two types based on batch processing and stream processing. Figure 2 shows Batch processing and stream processing are the two types of processing. Both deal with two different aspects. Batch processing works on stored data whereas Real-time processing deals on time-sensitive that or data that must be processed then and their only. According to the myriad types of data processing in big data, processing framework can be categorized into three types first is the Batch processing framework (Hadoop), second is Stream only framework (Apache Storm) and third is Hybrid frameworks (Apache Spark).

4.1. Apache Hadoop

Apache Hadoop is used for batch data. It was the first open-source framework that has significant power. It makes the process of massive amounts of data. Hadoop makes use of the Map-Reduce function to process such a large volume of data. Map-Reduce work on the concept of the divide and conquer method by giving out the problem in the subparts.

Hadoop infrastructure is made up of Master and worker nodes. The master node assigns the work to the worker nodes. When worker nodes are done with the task assigned to them then they return the output work to the master node. Master node amalgamates all the output sent by the worker nodes and generates the solution to the problem. Apache Hadoop technology is used in many different fields. Organizations such as Amazon, Microsoft, Google are using Hadoop technology. The advantages of Hadoop are it is best suited for large datasets, perform distributed data processing, and simply handle partial failure, simple programming model; however it has many disadvantages to name a few are like the processing of Hadoop becomes slow due to the amalgamate functioning of the master node, a single master node, the unapparent configuration of the nodes, restrictive programming model.

4.2. Apache Storm

The stream processing of big data with the basic 3 V’s (volume, variety, and velocity) characteristics is a laboring task. Hadoop can process these 3V’s but while processing the data in real-time where velocity is a major concern then Apache Storm comes into the picture. The challenges associated with Hadoop are ingestion of real-time data into its environment; parallel processing on time is not possible and at last, getting the output from the processing again on time is a challenging task as the working of Hadoop doesn’t support it.

Apache Storm can take care of all the above-mentioned challenges Kafka can be used to handle the ingestion of data, the second challenge can be handled by Strom which takes the data from Kafka.
and does the parallel processing of data, and finally generating the correlation of large datasets is performed by Storm. The main aim is providing low latency and thus the best option for stream processing. It can handle large datasets. It also works on the scheduling of Directed Acyclic Graphs (DAGs). Storm technology is composed of three things: first streams, second is spouts and third is bolts.

The stream is the immeasurable data unremittingly arriving at the system generating through various mediums and sites. Apache spout pulls the data from the stream through some technologies such as Kafka and makes it available to Bolt. Bolt has the processing logic that makes use of streams and applies operate it and then finally generates the stream as a result. At last, the Bolt output can be used as an input to the systems. Strom does not guarantee the chronological order of messages. It offers at-least-once processing which means each message will be processed at least once. However, the disadvantage is it might give you duplicate messages in case of failure.

**Figure 2. Batch processing vs Real-time processing**

4.3. *Apache Spark*

The Apache Spark technology is used for both types of processing. It can process both types of data. Apart from processing the high volume of data on time it also focuses on accelerating the batch processing workloads[18]. Spark provides full in-memory computation and thus optimizes the processes [19]. Its built process is also using many of Hadoop MapReduce functions but the main reason for using spark is its speed which is due to its in-memory calculation power and advancement in the directed acyclic graph (DAG) scheduling it can process datasets faster than Hadoop. Apache stream can handle an unrestricted volume of data however they process single stream or limited items at a time. This one
stream process stands for factual stream processing and very few stands for micro-batch processing and it must maintain a state in between records.

The advantages of Apache spark are its usefulness. It can be deployed as a separate cluster or united with a present Hadoop cluster. A single cluster can handle both the processing styles and thus perform stream processing and batch processing. Spark is easier to write than MapReduce. It has libraries for interactive queries, machine learning, etc.

5. Big Data Applications

The big data analytics help companies, entrepreneurs to make more informed business decision by providing analytics and predictive techniques. The application area of big data analytics are as follows:

5.1. Health care
Electronic health records have generated massive data set [20]. The data created in hospital or clinic are categorized into three types, clinic data, patient data and machine generated/ sensor data. Most of the health records contain quantitative data, qualitative data and data related to transactions. The big data analytics techniques impower the traditional methods by handling structured and unstructured data.

Big data creates the observational base for clinical questions. Big data can help to make personalized medicinal initiatives into clinical practice by proving the contingency to use analytical capabilities. This can combine the system biology and health record. Big data is used various applications e.g. healthcare data solution, anti-cancer therapy, tracking patient vitals, hospital administration improvement, promotion of business development, fraud detection and prevention for health insurance company etc.

5.2. Educational data mining and learning analytics
The online [21] education has become much more popular spatially during the pandemic period. Online activities of the students produce large amount of unutilized data. The big data techniques become very necessary in learning environment. Big data learning analytics techniques can be used for performance prediction, attrition risk detection, data visualization, intelligent feedback, course recommendations, student skill estimation, and behavior detection etc.

5.3. Process safety and risk management
Organization [5] have started to use of the big data in process safety and risk management. Big data can improve the analysis of quality and improve risk management by providing the statistics. Thus, the management can take quick and necessary decision.

5.4. Smart agriculture
Big data can be used in management of smart farming process [22]. New technology enables external big data source for e.g. market data, weather data with farms that contributing to development of smart farms. The big data transforming agriculture by boosting productivity, predicting yields, risk management, and food safety.

The big data is so powerful tool that it can be in many fields. The applications are used in various other like government, social media analytics, fraud detection, call center analytics, banking, marketing, telecom sector etc.

6. Conclusion

The data is generating at an exponential rate and the universe will grow by 50-fold by the current year 2020 and thus the study of managing the whole data is very important. Big data are generated through internal and external sources of data thus existing systems fail to handle the unprecedented data therefore
high-performance, highly scalable systems with advanced techniques are required to process the
valuable information.
The paper discussed the most imperative characteristics of big data. The paper also presented the various
types of big data analytics, big data analysis process, and some prominent challenges about big data.
Study shows that the current tool and technology need to be updated with the time as the data is growing
continuously.
Future research can be done on pre-processing part as major researchers have not focused on this part
and researcher studies also show that pre-processing is one of the major phases that play a significant
role in generating quality output.

7. References

[1] Sivarajah U, Kamal M, Irani Z and Weerakkody V 2016 Critical analysis of Big Data challenges
and analytical methods J. Bus. Res. 70 263–86
[2] K.U J and M.David J 2014 Issues, Challenges and Solutions : Big Data Mining CS IT-CSGP 4 131–
40
[3] Yaqoob I, Hashem I A T, Gani A, Mokhtar S, Ahmed E, Anuar N B and Vasilakos A V 2016 Big
data: from beginning to future Int. J. Inf. Manage. 36 1231–47
[4] Gandomi A and Haider M 2015 Beyond the hype: Big data concepts, methods, and analytics
Int. J. Inf. Manage. 35 137–44
[5] Hariri R H, Fredericks E M and Bowers K M 2019 Uncertainty in big data analytics: survey,
opportunities, and challenges J. Big Data 6 44
[6] Martin A 2011 a Framework for Business Intelligence Application Using Ontological
Classification Int. J. Eng. Sci. Technol. 3 1213–21
[7] Ma C, Zhang H H and Wang X 2014 Machine learning for Big Data analytics in plants Trends
Plant Sci. 19 798–808
[8] Tsai C-W, Lai C-F, Chao H-C and Vasilakos A V 2015 Big data analytics: a survey J. Big Data 2 21
[9] Galetsi P, Katsaliaki K and Kumar S 2020 Big data analytics in health sector: Theoretical
framework, techniques and prospects Int. J. Inf. Manage. 50 206–16
[10] Khan N, Yaqoob I, Hashem I, Inayat Z, Kamaleldin W, Alam M, Shiraz M and Gani A 2014 Big
data: survey, technologies, opportunities, and challenges Sci. World J. 2014
[11] Saha P, Mittal M, Gupta S and Sharawi M 2017 Big Data trends and analytics: A survey Int. J.
Comput. Appl. 180 9–20
[12] Shah T, Rabbi F and Ray P 2015 Investigating an ontology-based approach for Big Data
analysis of inter-dependent medical and oral health condi Cluster Comput. 18 351–67
[13] Schroock M, Shockley R, Smart J, Romero Morales D and Tufano P Analytics: the real-world
use of big data: How innovative enterprises extract value from uncertain data, Executive
Report IBM Inst. Bus. Value Said Bus. Sch. Univ. Oxford
[14] Khan N, Alsaker M, Shah H, Badsha G, Abbasi A and Salehian S 2018 The 10 Vs, issues and
challenges of big data Proceedings of the 2018 International Conference on Big Data and
Education (New York, NY, USA: Association for Computing Machinery) pp 52–6
[15] Nair L R, Shetty S D and Shetty S D 2018 Applying spark based machine learning model on
streaming big data for health status prediction Comput. Electr. Eng. 65 393–9
[16] Garcia S, Ramirez-Gallego S, Luengo J, Benitez J M and Herrera F 2016 Big data preprocessing:
metholds and prospects Big Data Anal. 1
[17] Hashem I A T, Yaqoob I, Anuar N B, Mokhtar S, Gani A and Ullah Khan S 2015 The rise of “big
data” on cloud computing: Review and open research issues Inf. Syst. 47 98–115
[18] Dritsas E, Livieris I E, Giotopoulos K and Theodorakopoulos L 2018 An Apache Spark
implementation for graph-based hashtag sentiment classification on Twitter *Proceedings of the 22nd Pan-Hellenic Conference on Informatics* pp 255–60

[19] Oo, Myat Cho Mon and Thein T 2019 An efficient predictive analytics system for high dimensional big data *J. King Saud Univ. - Comput. Inf. Sci.*

[20] Murdoch T B and Detsky A S 2013 The inevitable application of big data to health care *J. Am. Med. Assoc.* **309** 1351–2

[21] Sin K and Muthu L 2015 "Application of big data in education data mining and learning analytics – a literature review" *ICTACT J. Soft Comput.* **05** 1035–49

[22] Wolfert S, Ge L, Verdouw C and Bogaardt M-J 2017 Big Data in smart farming – A review *Agric. Syst.* **153** 69–80