Pakistan Journal of Engineering and Technology, PakJET
ISSN (p): 2664-2042, ISSN (e): 2664-2050
Volume: 5, Number: 2, Pages: 216- 220, Year: 2022

Big Data Issues, Challenges and Techniques: A Survey

Tariq Mehboob, Irshad Ahmed Sumra, and Ayesha Afzal

Department of Information Technology, Lahore Garrison University Lahore, 54000, Pakistan
Corresponding author: Tariq Mehboob (ranatariqm@gmail.com)

Received: 25/05/2022, Revised: 20/08/2022, Accepted: 20/09/2022

***Abstract*— The data is growing daily due to the revolution in Information Technology, and the data size is much bigger than traditional data. Big data is the combination of data sets which is huge and diverse. Big data has multifaceted characteristics like Volume, Variety, Veracity, Velocity and Value (5Vs). Due to these characteristics and versatility, many issues and challenges occur in data retrieving and manipulation. The data management and processing issues are also facing data analysts and researchers. The existing traditional tools and algorithms are not capable of resolving these issues. In this survey paper, we will provide the current issues and challenges in the field of big data and discuss in detail the big data tools (Hadoop, Apache Spark) to use in different applications to serve the end users.**

Keywords—Big Data, Velocity, Variety, Heterogeneity,Hadoop, Apache Spark, Yottabyte.

I. INTRODUCTION

A collection of large and complex data is called big data in which processing activities involves by using management systems and software techniques but cannot be processed by traditional database methods and tools. The big data is identifying datasets that are large in size and complex [1], due to this current technology regarding analysis is obsolete to analyze big data. The big data is a varied combination of structured and unstructured data. The big data mining has capability to remove valuable data from large datasets or data chains that were not possible first because of its quantity, variety and speed. The big data is achieving more attention due to revolutionizing technology called the internet of things (IOT) producing large amounts of data. Thesecurity functions of big data are necessary for work flow of various components of hardware, operating system and network domains. There are many kinds of data like textual, graphical, streaming which are having five attributes

i.e. volume, velocity, variety, Veracity and value. Growth rate in the quantity of data collection is astonishing. Data growth rate is a big challenge for researchers and IT practitioners. In general, big data means large and complex amount of data collected from different sources like web, enterprise applications, mobile devices, and sensors generated data. Two method have been proposed for the solutions of big data issues (a) Batch based stored data processing and (b) Real time data stream processing. The both above methods are implemented through two techniques i.e Hadoop MapReduce and Apache Spark. The given paper is divided into the following section. The Section II is about big data types which elaborates in Figure I and section III is about characteristics of big data which are 5'vs. The section IV is on issues in big data. Section V is challenges for big data. Section VI shows techniques for big data. Section VII shows conclusion to understand and specify to point out issues and challenges in big data and its techniques.

II. BASIC CONCEPT OF BIG DATA AND ITS TYPES

The big data is depending on structured and unstructured data types. Taking advantage of big data to integrate information for analysis and data management is important. In the structured data, sorted / labeled and store in warehouse same as the concept of data warehousing. The unstructured data type it is random and tough to analyze in big data. The following Fig. 1 elaborates in detail big data types with examples.

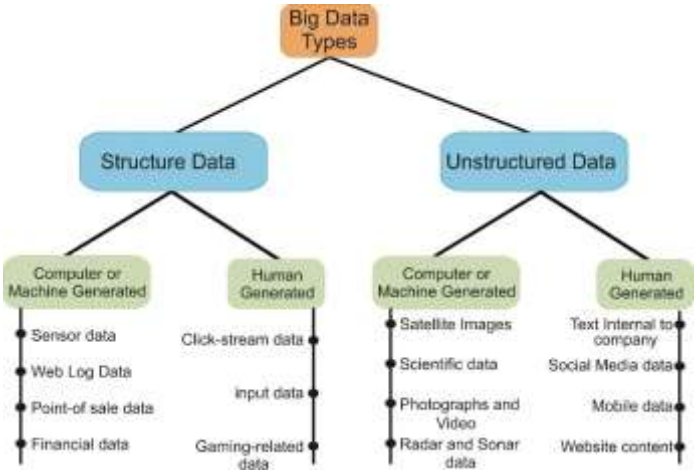

FIGURE 1. Big Data Types [20]

III. BIG DATA CHARACTERISTICS

The characteristics of big data are sets of parameters that describe big data with the analytical approach in which the main features of big data mostly refer as 5vs i.e. volume, variety, velocity, Veracity and value. These are represented in Fig. 2.

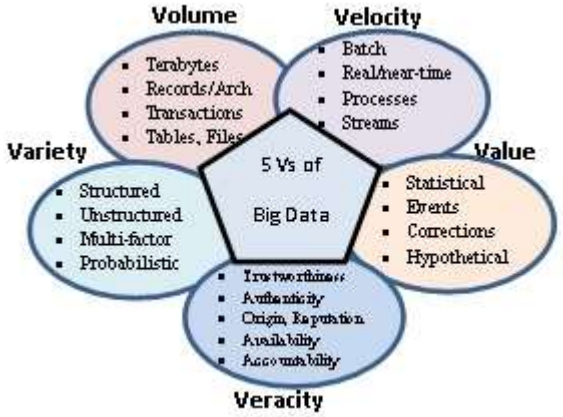

FIGURE 2: 5Vs Big Data Characteristics [21]

- VOLUME

In today's world, the I.T is involved in every era/field of work for ease of work but the storage of data at every second, minute or hour is a big issue which needs proper planning. In the volume clause, the issue is handling the big data in our digitized world. The data is in different shapes: satellite imaginary data, social media photos galleries, sensor's unstructured data, documents and much more different kinds of data on internet which is uploaded daily 24/7. The issue is maintaining users' data and the volume of data parallel, which can be in the size of gigabytes, terabytes, petabytes, and zettabytes.

- VARIETY

Variety means different forms of data can come in text, images, sounds, and geographic data; computer-generated images are increasing daily. Using statistics methods, the scattered data / data in bulk form can be divided by characteristics/nature.

- VELOCITY

The velocity here means the speed of large data generating or transmitting from one end to another and data retrieval for analyzing and storing. In the real-time, the data processing, production rate is important for the availability of data as well as the purpose of big data analysis. This speedily allows the privacy, search, viewing and purchase history on the web sites. In the big data, the searching or processing of activities are important to analyze and take advantages; otherwise, it will be a waste of opportunities.

- VERACITY

Veracity defines the quality of big data and data reliability. Veracity relates to prejudice, noise and exception of data. There is a lot of sources of big data veracity like statistical biases, untrustworthy data sources and software bugs, so if data is invalid, duplicated and incomplete or outdated, collectively known as dirty data [3]. If accuracy does not occur, then data is useless.

- VALUE

Value is the important characteristic of big data in which unstructured data make meaningful and valuable and is use for process, predictive analysis or hypotheses. The data will depend on the processes they represent for example, stochastic, probabilistic, regular/random. The collected data and its storage depend that how much important for a particular process or event also the value of the data in this regard is the volume of data.

IV. ISSUES OF BIG DATA

In big data processing and manipulation, many issues and challenges are occurring, in this regard trying to summarize most relevant big data issues and challenges which are given below.

A. BIG DATA ISSUES/CHALLENGES RELATED TO CHARACTERISTICS OF BIG DATA

- DATA VOLUME

The first issue is data storage, and necessary data retrieval requires fast speed to extract the best result. As the volume of data maximizes, the value of data records decreases in proportion to period type, majesty, and quality [4].

- DATA VELOCITY

Today's the computer system is creating data by users and the subsequent working etc. People want maximum data retrieval within a short time / speedily, so high-speed data means millions of rows/columns of data at every nanosecond. Traditional technologies cannot resolve data velocity issues [4].

- DATA VARIETY

Big data comes in many forms, such as photos in social media, websites, messages, email, and GPS signals from sensors, satellites, cell phones and much more. Most of these figures are erratic, rude and noisy, requiring rigorous techniques to make data in a meaningful form for analyzing/decisions. To analyze such data better algorithms are required to overcome this issue [5].

- DATA VALUE

Mostly data storage by organizations for gain results/outcome uses them for analytical business intelligence. This develops the main bridge between organizational management and ITprofessionals.

B. MANAGEMENT ISSUE

Unstructured data is always treated as unsolicited data because big data produces many various sources of various formats, and representations [6]. So performance required to manage big data through multidimensional management tools otherwise unacceptable results will come out. Big data features are different at their natures so managing data formats and infrastructure in business organizations needs data stores with flexible features and scalability. Mostly big data is not organized and creates complexity in business organizations. It is difficult to evaluate and extract meaningful information, requiring the upgraded deployment of management techniques or tools with subsequent expandable data management tools/techniques to set up a new business improvement.

C. STORAGE ISSUE

Making accurate decision requires more information according to market strategies [7]. A large unstructured big data has a good amount of information for big data professionals [8].

In the above statements and observations, it is important to grow business by understanding the importance of big data but, unfortunately lack of data storage equipment. The amount of data is about our decisions, marketing strategies, and recommendation systems, and data is stored daily in terabytes. However, resources to store data are very limited, making it difficult for business organizations to select data or part of data with its features and datasets. So, in this regard, the constant necessity of tools and methods are helpful in different organizations / firms with identifying features or principal components and thousands of attributes for users to understand in detail.

D. PROCESSING ISSUE

Nowadays real-time results are really important, especially for business organizations. If no accurate and timely results are prepared, they will be used sparingly[8]**.** Maximum of the organizations, they have shifted their business style from 'bricks and mortar' mode to online style for the promotion of consumer with global sales resulting in a data storm. In the existing data model, current data techniques are incapable of processing such a huge amount of data in real time [7] which makes business organizations with disabilities. Some latest indexing schemes are available (such as Fast Bit) [9]and processing methods such as map reduction [10], [11] are available to raise processing speed. In zettabytes, processing (10^{21}) and exabytes (10^{18}) data is still a difficult task. Real time data is required accuracy in its processing. Organizations are also expediting the search. Also, processing traditional data systems must be upgraded for accuracy and fast. Figure 3 shows the augmentation of data from one bit to Yottabyte. According to this, it gave the idea of data and exceeded terabytes/petabytes. For this purpose, supercomputers are only capable of storing and processing data up to petabytes, which need of organization as well. These measurements are discussed in Fig. 3.

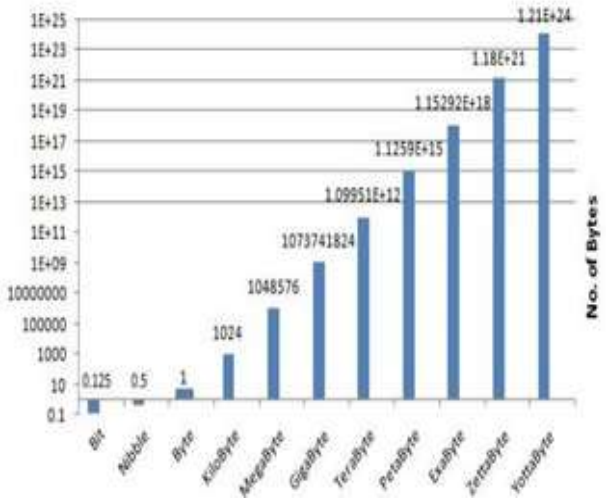

FIGURE 3. Data Scale from Bit to Yottabyte[22]

E. DATA HETEROGENEITY

The data heterogeneity happened due to different standards of organizations regarding data. As a result, researcher faces many difficulties accessing different pools of data and information saved in different formats using different Information Management System. IOT is a big source of generating of data heterogeneity. Experts faces different issues in managing and extract actual information from unstructured data.

F. DATA QUALITY

Big data storage is very expensive and always conflicts between business analysts/organizations and IT experts. The data quality is an essential factor and ensuring sufficient data for a particular result.

G. SCALABILITY

The rapid growth in the volume of big data all over the world devastates the database management system. It needs high level of resource sharing but expensive and brings various challenges regarding performing different jobs with the burden of multiple tasks met successfully. Working with large cluster the system effectively as failures and it is common. The hard disk drives media are being replaced with latest storage technology, so this is a big question around the big data storage issue.

III. CHALLENGES FOR BIG DATA

In every field of life or even in organizations or firms, opportunities and challenges always travel parallel. Big data came with various opportunities in I.T. world and in firms/businesses and many challenges there. Now the discussions are given below on some important andpertinent challenges for the concentration of researchers.

A. LACK OF BIG DATA PROFESSIONALS

In the current scenario, there is a lack of I.T professionals in

development for the complexity of big data and its processing technologies need highly skilled professionals. These skilled professionals should be equipped with the latest technologies and tools tailored to an organization's requirements. There is no doubt that many experts exist in data sciences. Still, given the current situation, these inexperienced specialists need special training to become experts in dealing with big data of various scopes, including data modelling, data analysis, and data integration [7]. Therefore, organizations engaged in big data analysis and framework require gigantic demand of professionals for dealing with challenges like data scientists

/ engineers and data architecture. There is also a need of management for effective decision-making and initiatives. According to research, every organization needs to make a specialist in big data having vast analytical skills [12], also big data analysis in business intelligence has been specify is one a big reason for the rapid growth of businesses [13].

B. LOADING AND SYNCHRONIZATION

Big Data synchronization is a very serious issue. Data loading involve retrieving data from a single warehouse of data from several different data bases [15]. Many issues about the loading process need the consideration of researchers and experts. Multiple data sources should be mapped into correct structural framework, tools and infrastructure. Loading issues, and synchronization in different data sources consider main challenges. The data can be loaded from different sources to another source with different speeds of time and is likely to move away from synchronization. The data synchronization source is the process of establishing uniformity in data over time in different data sources. Incorrect or poor mining results are the guarantee of failure of synchronization this needs a greater attention need for synchronization of data sources to help firms / organizations and should avoid risks during analysis processes. Also draw the accurate and appropriate conclusions check it solutions again for clarity / perfections. The inconsistent nature of the data makes it more complicated for businesses to change and clean up before entering them in the warehouse for analysis [16].

C. VISUALIZATION

It's the procedure of representing knowledge in an understandable way for decision-making. Due to the increase of data daily, it becomes very tough to speed up because of unavailability of resources, i.e. volume and scalability. There is no doubt that online marketers (such as eBay) are using great data visual tools such as Tableau for converting large complex data sets into them image shapes to make all the data clearly reasonable [17]. So far, the analysis phase or researchers must focus on the conceptual challenges of big data for considering future. As everyone generates data through online social networks, medical science, geo stationery satellites, sensors, big data fits as "larger data"[14]. The challenges will increase daily, so updating or adopting new technologies and tools is mandatory. Big data visual techniques can be monitored, and outcomes/results by different graphs for decisions. The visual reports give a better ideology than text reports for customer understanding.

IV. TECHNIQUES FOR BIG DATA

For the big data processing delays as a major concern, normally two approaches proposed and used for processing of big data like Hadoop MapReduce and Apache Spark which are discussed below:

A. HADOOP

The Hadoop is an open source tool used as a major data processing framework by researchers and analysts [18]. Hadoop MapReduce has an option for data processing in which all inputs must be read at once but is very slow in the multi-pass calculation. On the other hand, Google developed MapReduce with two components i.e. Map and Reduce used to compute the key and value in pairing for the map combines input and low map function results in scalar using of these tool data analysts can take full advantage of Hadoop MapReduce. This framework's implementation gives benefits like schedules, monitoring tasks and implementation of failed tasks [19]. In the MapReduce framework output data stores and enter in the DFS (Distributed File System) but this slows down processing of

data. It has capabilities large quantities of complex and complex clusters to manage the big data.

B. APACHE SPARK

The Spark use for cluster computing framework, which integrates APIs and parallels across language operators. Apache project was developed by AMP Lab Berkeley approximately 6-years ago in AMP Lab Berkeley since 2010 [18]. The Sparks are very useful other than Hadoop. A unified approach for managing processing activity required speed up application in Hadoop cluster. Also, it runs 100 times faster in memory and 10 times fast on disk. This also supports in writing applications like more than 75 applications in which Scala, Java or Python are the top-level operators. Direct Cyclic Graph (DCG) in complex development through sparks which support multi-step data pipelines. Also, this supports for memory sharing across multiple jobs and tasks.

V. CONCLUSION

Data Science is the grouping of programming skillswith an analytical approach, domain expertise and knowledge of mathematics with statistical techniques for gathering/retrieving meaning / required data. Data science is all about discovering knowledge from very large data sets that are semi-structured and unstructured for the interpretation or search of data. In the revolutionary period, the frequently changing world, it is implemented/used by different industries like finance, retail, healthcare, manufacturing, sports and communications. Due to global change in IT sector daily, the data is increasing e.g. (Facebook generates 1.00million data in one minute, YouTube generates 4.5 million in one minute etc.) and it

needs modernization with high-performance and capable equipment. For this purpose, challenges and issues come with algorithms and scalable techniques in big data sets. Analysis of big data is much more important for a successful business organization.

On the other hand, the organizations lagging behind in analyzing their big data are weak and will face financial loss in terms of their futuristic customers. In this case, data mining can play a good role that will benefit firms/ organizations by avoiding financial loss or big unhandled data. Cloud computing is also an example of predicting future trends and business operation behaviours with the support of knowledge-based decisions. A detailed introduction comparing the most popular big data processing frameworks, i.e. Hadoop MapReduce / Apache Spark, was introduced. These are helpful for data scientists and researchers in analyzing big data and unknown patterns/uncovering hidden. Researchers will have to work hard on eachside means hardware and software to overcome current challenges/handling of big data.

FUNDING STATEMENT

The authors received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] Albert Bifet, "Mining Big data in Real time", Informatica 37, pp15- 20.

[2] J. M. M. Hansen1, T. Miron-Shatz2, A. Y. S. Lau3, C. Paton, "Big Data in Science and Healthcare A Review of Recent Literature and Perspectives" Year Med Inform: 1-6.

[3] Elkay Altintas, Amarnath Gupta, "Introduction to Big Data", University of California, San Diego

[4] Harsh Kishore Mishra, "Big data issues and challenges", CUPB\MTech-CS\SET\CST\2013-14\01

[5] Andrew McAfee,Erik Brynjolfsson, "Big Data: The management revolution", Harvard Business Review.vol. 90, no. 6, pp. 60-68,2022.

[6] X. Wu, X. Zhu, G. -Q. Wu and W. Ding, "Data mining with big data,"in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014, doe: 10.1109/TKDE.2013.109.

[7] Katal, Avita et al. , "Big data: Issues, challenges, tools and Good practices." *2013 Sixth International Conference on Contemporary Computing (IC3)* 2013, 404-409.

[8] Che, D., Safran, M., Peng, Z, "From big data to big data mining: challenges, issues, and opportunities," In: Database Systems for Advanced Applications. pp. 1-15. Springer 2013.

[9] Wu, K, "Fast bit: an efficient indexing technology for accelerating data-intensive science," In Journal of Physics, Conference Series, vol. 16, p. 556. IOP Publishing 2005.

[10] Dittrich, J., Quiane-Ruiz, J.A., “Efficient big data processing in Hadoop MapReduce,” Proceedings of the VLDB Endowment vol. 5, no. 12, pp. 2014-2015, 2012.

[11] Triguero, I., Peralta, D., Bacardit, J., Garc a, S., Herrera, F.," MRPR: A MapReduce solution for prototype reduction in big data classification," neurocomputing vol. 150, pp.331-345, 2015.

[12] Kim, G. H., Trimi, S., & Chung, J. H. "Big-data applications in the government sector," Communications of the ACM, vol. 57, no. 3, pp.78-85, 2014.

[13] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., "Big data: The next frontier for innovation, competition, and productivity," 2011.

[14] Kaisler, S., Armour, F., Espinosa, J.A., Money, W., "Big data: Issues and challenges moving forward," In: System Sciences (HICSS), 2013 46th Hawaii International Conference on. pp. 995-1004. IEEE (2013).

[15] Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., "Spark: cluster computing with working sets," In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. vol. 10, p. 10, 2010.

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

[17] Sawant, N., & Shah, H., "Big Data Visualization Patterns. In Big Data Application Architecture," Q & A, pp. 79-90, 2013.

[18] Apache Software Foundation, "The Apache Software Foundation Blog," (August 2014).

[19] Apache Software Foundation, Hadoop MapReduce Tutorial, version 3.3.1, 2021.

[20] Hadi,Hiba Jasim,Shnain, Ammar Hameed,Hadishaheed,Sarah, "Big Data and Five v's Characteristics," International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835, vol.2, no. I, Jan-2015.

[21] Ishwarappa,J.,Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," International Conference on Intelligent Computing, Communication and Convergence (ICCC 2015), Bhubaneswar,Odisha,India

[22] Wani, Mudasir & Jubin, Suraiya," Big Data: Issues, Challenges, and Techniques in Business Intelligence", 10.1007/978-981-10-6620-7_59, 2018.