Research Article

BIG DATA CHARACTERISTICS, CLASSIFICATION AND CHALLENGES - A REVIEW

K S Ananda Kumar[1*], Sisay Muleta Hababa[2] Bekele Worku [3] , Gizaw Tadele [4], Yihenew Gebru Mengistu[5] and Prasad A Y[6]

[1,2, 3, 4 &5]School of Computing & Informatics, College of Engineering & Technology, Dilla University, Dilla, Ethiopia
[6]Department of Computer Science & Engineering, GITAM University, Bangalore Campus, Karnataka, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 May 2021

ABSTRACT
The data is an assest of great importance for any organization. Big data is the large collection of data; it includes different types of data collected from banking, e-commerce, insurance, manufacturing, social media and business data etc . Big data Analytics is the examine the bulk amount of data. To deal the large amount of data traditional techniques are inefficient, less accuracy and less performance. Big data technologies are face the large and complex data in efficient manner. Hadoop technology is designed to process the Big data. Hadoop is an open source software used for distributed processing of big data among the servers. Parallelism technique is used to process the large amount of data. Currently the big data analytics is the one of the research area and development. Nowadays the big data has great importance and good choice for new researchers. The paper's main purpose is to discuss the features of big data and the technological challenges.

Key words: Big data analytics, Challenges, Characteristics, Hadoop , HDFS.

1. INTRODUCTION

Now a days, huge amount of data generated at unprecedented rate from various sources like social media, government schemes, health and marketing, because of technology development and spreading of many new smart devices, it leads to growth of big data. Big data efficiently handling the bulk amount of data, which is not possible by using conventional techniques like relational databases etc. Big data used to store large amount of data, retrieve and modify the large data sets [1-4]. The advancements in technology and increase in population, large amount of information is generated that can be structured, semi-structure and unstructured format. The large amount of data is in unstructured form, very difficult to process this type of data because it contains variety of records with different people that consist of videos, audios, images, web data etc. To process the large volume of data in an inexpensive and efficient manner using the concept of parallelism [5-7]. Big data is stored and analyzed in computer databases, using software designed to handle complex and large data sets. The hadoop is a system used to process massive data and to operate in a distributed computing environment [8, 9, 18]. Traditionally, innovations are created and tested in laboratories before being released to the public through press releases and advertisements. The general population then adopts these innovations. The rapid growth and widespread adoption of big data by the general public left little time for the academic domain to mature [20]. Despite the fact that many books and electronic media papers on big data are written by experts and writers, basic research is still lacking in academic publications. Traditional data analysis approaches are ineffective when dealing with large-scale, complex data. As a result, nearly 80% of businesses have little insight into their unstructured data and knowledge of how to manage it. Modern companies need new approaches to analyze different big data because of unstructured data. Data analysis approaches are increasingly incorporating new methods such as artificial intelligence (AI), association rule learning, machine learning, genetic algorithm , classification tree analysis, social network analysis , regression analysis, and sentiment analysis. These approaches also have an effect on how data is analyzed. Furthermore, predictive analytics on security log data aids intelligent security by providing a strategy for predicting, preventing, and mitigating potential

cyber-attacks. This crucial aspect of big data analytics will assist companies in identifying new opportunities. However, the feature is still in its early stages and is open to further development [21].

2. BIG DATA CLASSIFICATION

Big data comes in a variety of formats, each with its own set of characteristics. Understanding the strengths and disadvantages of applications that process broad dataset volumes necessitates the classification of big data. Data storage, content types, and data staging are all categories that can be used to classify big data. Each of these groups has its own set of characteristics and interdependencies. The following is a list of big data classifications [21]:

A. Data Sources

Social data, machine data, and transactional data are the three main sources of big data. Furthermore, businesses must distinguish between data generated internally, or data that remains behind a company's firewall, and data generated externally, or data that must be imported into a system. It's also important to consider whether the information is unstructured or organised. Since it lacks a pre-defined data model, unstructured data takes longer to comprehend. On the world's most prominent social networking sites, Likes, Tweets & Retweets, Comments, Video Uploads, and general media are all sources of social data. This type of data can be extremely useful in marketing analytics because it offers unparalleled insights into customer behavior and sentiment. Information generated by industrial machinery, sensors mounted in machinery, and even web logs that track user behaviour are all examples of machine data. As the internet of things becomes more popular and expands around the globe, this type of data is expected to grow exponentially. Medical devices, smart meters, road cameras, satellites, sports, and the rapidly growing internet are all examples of sensors. Transactional data includes invoices, payment orders, storage records, and delivery receipts, but data on its own is almost worthless, and most companies have no idea what data they're generating or how to use it effectively.

B. Content Format

Big data content formats can also be used for classification. The following are the various forms of big data based on content format:

Structured data – In a database, structured data is typically tabular data expressed by columns and rows. Relational databases are those that store tables in this format. The mathematical term "relation" refers to a table that contains a constructed collection of data. Every row in a table in structured data has the same set of columns. SQL (Structured Query Language) is a structured data programming language.

Semi structured data – Semi-organized data is information that isn't structured (like a relational database), but also has some structure. Documents in the JavaScript Object Notation (JSON) format make up semi-structured data. It also contains graph databases and key-value stores. Such data necessitates dynamic processes for complex laws during data operations. As a result, the challenge of working with such a complex data source is still under investigation.

Unstructured data – Unstructured data is information that does not have a pre-defined structure or a data model. Unstructured data is a collection of text-heavy documents that may also include numerical, chronological, and factual data. There is no guarantee that videos, audio,or binary data files would have a particular structure. Unstructured data is the name given to them. Since the scale of this form of data is constantly growing as a result of the number of mobile phones and social media apps, managing it is a significant challenge.

C. Data staging

Traditionally, we have a range of staging or intermediate data storage areas / structures in our information architectures. Publish directories on source networks, staging areas in data centers, data vaults, and, most notably, data file hubs have all been used in the past. These data file staging methods have two major drawbacks in general: Data retention was normally restricted to a few months due to storage costs. End-users did not have access to staging data for analytics or data discovery because these systems were designed to publish data for system integration.

D. Data Processing

The type of processing that produces the data can be used to classify the data. The following are examples of processing methods:

Batch processing – Batch processing is the simultaneous processing of a large amount of data. For a single day, the data can easily consist of millions of records and can be processed in a number of ways (file, record, etc). In most cases, the jobs are done in a nonstop, sequential order.

Stream processing - Stream processing is the ability to interpret data that is flowing from one computer to another in a near-real-time manner. This type of continuous computation occurs as data flows through the system, with no time constraints on the performance. Systems do not require large volumes of data to be processed due to the near-instant flow. If the events you want to monitor happen regularly and close together in time, stream processing is a great option. It's also the best option if the event needs to be identified quickly and responded to. So, tasks like fraud detection and cybersecurity benefit from stream processing. Fraudulent transactions may be detected and stopped before they are completed if transaction data is streamed.

Real time – Data reactions are commonly referred to as "real-time data processing." A system is real-time if it can guarantee that the reaction will happen in a short amount of time in the real world, usually seconds or milliseconds. One of the best examples of real-time systems is stock exchange systems, which produce real-time data.

E. Data Stores

Clusters of data storage are needed for effective and timely performance from big data analytics. Since conventional relational database models are not suited for very large-scale databases, performance issues arise throughout big data analytics. Because of the ability to horizontally partition data, extensive processing power, and better performance, No-SQL databases are favored over SQL databases for processing. NoSQL is used by companies including Google, Facebook, Amazon, and LinkedIn to handle ever-increasing data streams. Top ten big data stores are Cassandra, Hbase, MongoDB, Neo4j, CouchDB, OrientDB, Terrstore, FlockDB, Hibari and Riak.

3. BIG DATA CHARACTERISTICS

The characteristics of big data are Veracity, Velocity, Volume, Value and Variety [1, 4, 6, 14]. Figure 1 shows the big data characteristics.

Veracity – Veracity means accuracy of of collected data, when we are working with large volume of data, its not possible to get 100% correct data. Data is uncertain because of incompleteness. The quality of the captured data varies greatly. The accuracy is analyzed based on the source of data.

Velocity – Velocity is the speed at which information generated and proceed for processed, stored and analyzing using databases, like data generated with very speed in social media, its spread around the world in less time.

Volume – Volume means bulk amount of data generated from all the machines and smart devices, its stored in databases from megabytes to petabytes. The volume shows the challenge to traditional IT companies, large amount of data available to process the data. The large amount of data process by only Big data analytics only.

Value - The data value shows the business value of the big data. The Big data is very valuable and its very costly to apply for IT industry to store and access big data. Big data 's value for business is high and it gives good returns to investors.

Variety – Variety is a key aspect of big data, data is available in different forms like audio, images, videos, social networking data, text, tables etc. Big data contains structured and unstructured data, most of the data generated in unstructured form, cannot use directly to analyzing the data using big data. For storing and analyzing the data it becomes complex process.

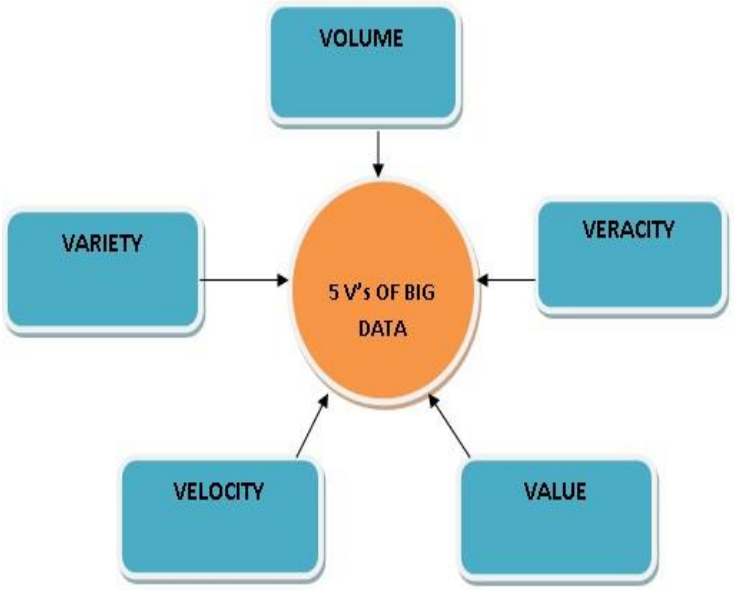

Figure1 5 V's of big data

4. CHALLENGES IN BIG DATA ANALYTICS

There are some of the challenges of big data [15-17] .

DATA ACQUISITION AND RECORDING

Big data's greatest obstacle is the data storage. The quantity of data stored in the IT systems nowadays doubles every two years. Most of the data is in unstructured form, it is not possible to store in databases. Text, images, audios, videos and other types of data can be difficult to analyze. To deal with data growth, the number of companies introduces a number of different technologies. To reduce the amount of storage space and costs, use the technologies such as tiering, compression etc. The organizations are using tools such as Spark, NoSQL databases etc, helping in searching the data used for their business need.

KNOWLEDGE DISCOVERY AND CLEANING

To extract the new information from available data, used for suitable analysis. Some times data may be very poor quality and incomplete, this data not able to use for analysis. Data cleaning is very critical step in big data, after cleaning the data its used for Analysis.

DATA SCALABILITY AND REPRESENTATION OF DATA

The most serious challenge of big data is ability to grow i.e data scalability. In the last decade most of the researchers are given importance to increase the data analysis and speed up processors. Incremental technique is the one of the scalability property of big data analytics. The main objective of representation of data quickly understands the data. Graphical representation gives the relation between data and proper interpretation. The online shopping applications like amazon, flipkart etc have large number of users and more number of items sold every month. Current visualization tools have poor performance in scalability, functionality and time response. In this process bulk amount of data is generated, some organizations uses a Tableau tool for data visualization. This tool converts the large data into suitable images. This helps the company's employees to visualize, Check significance, track the latest feedback from customers and evaluate their perceptions.

DATA VALIDATION AND INTEGRATION

Organizations are provided with similar pieces of information from different systems, that does not agree for all the time. Big data integration ensure that data availability to the consumers in right time. The process of getting the records with accurate, secure and usable is called data governance. The IT Organizations create a data verification group of people, and write a set of policies and procedures. That will ensure that big data stores are accurate. Data comes from many different places in data integration — company applications, social media, emails, employee-generated documents, etc. Vendors provide a

range of data integration tools designed to make the process simpler but several businesses agree that the data integration issue has not yet been solved.

QUERY PROCESSING AND ANALYSIS

Big data suitable approaches must be capable of tackling noisy, chaotic, heterogeneous, untrustworthy data and complex relationship-defined data. Despite these challenges, big data may be more useful to detect more accurate hidden trends and information, good data compared with small samples, even though it is noisy and unpredictable. The computing infrastructure and scalable mining methods are used for query processing and analysis.

INTERPRETATION

Decision-makers need to understand the results of the analyzes derived from big data, and this will require users to be able to test the findings at each data processing point and likely to retrace the analysis. The findings of the study will be provided for explanation to decision-makers. End user has to understand and check the results created by computer systems, and the computer system to make the work easier for the user. Consumers would have the tools required for both understanding the findings of the analysis and redoing the analysis with different assumptions, parameters and data sets.

INFORMATION SECURITY

Many companies, however, tend to think that their current data protection approaches are still ideal for their big data needs. The most of the organizations facing the problem with data security. This is a really major organizational problem. Security is one of the big concerns for big data stores. Big data stores are targets for hackers. Big data technologies are gradually increasing their importance but security features are continue to be overlooked. Data security is both a question of loss prevention and of privacy. Because of Data Criticality, A minor accident may result in enormous losses and therefore companies are bound to introduce the best safety practices in their systems [2,15-17].

PRIVACY

Another major issue is data protection, and one that is growing in the perspective of big data. However, there is considerable public concern over inappropriate use of personal data, by integrating data from different sources in particular. Privacy management is indeed both a technical and a sociological issue that needs In order to deliver on the promise of big data, to be addressed jointly from both perspectives [16].

5. TRADIOTIONAL DATA Vs BIGDATA

Traditional information is standardized information that is used by a wide range of organizations, from small organizations to large organizations. In a conventional database system, data was stored and maintained in permanent formats in a file using a centralized database architecture. Big data is deals with large or huge data sets that are difficult to handle using traditional data-processing tools. It manages vast amounts of structured, semi-structured, and unstructured data. Big data does not only refer to a vast volume of data; it also refers to extract useful data from complex data sets [19].

	Traditional data	Big data
1	Impossible to store large amount of data	Store huge volume of data easily
2	Based on a fixed schema that is static	Here uses a dynamic schema
3	Deals with structured data only	Deals with Unstructured, Semi- structured and Structured data
4	Centralized database architecture	Distributed architecture
5	Sources of data were fairly limited	Sources of data were unlimited
6	Data size is small	Data size is large
7	Data handling is easy	Data handling is difficult
8	Normal function are enough for manipulate the data	Special functions are required for manipulate the data

9	Data generated per hour or per day	Data generated frequently per seconds
10	Easily integrate the data	Difficult to integrate the data

6. HADOOP FRAME WORK

Hadoop is an open source program used for the big data analysis. This is implemented by Apache software foundation. It contains of many small sub programs belongs to the distributed computing. Hadoop is very popular and used by researchers to analyze the big data [04]. Apache hadoop system consists of hadoop kernel, HDFS, Map reduce and other elements etc. Map reduce is developed by Google, it is part of Apache hadoop. MapReduce is a programming framework model where the application breaks into the different parts [1, 6,11].

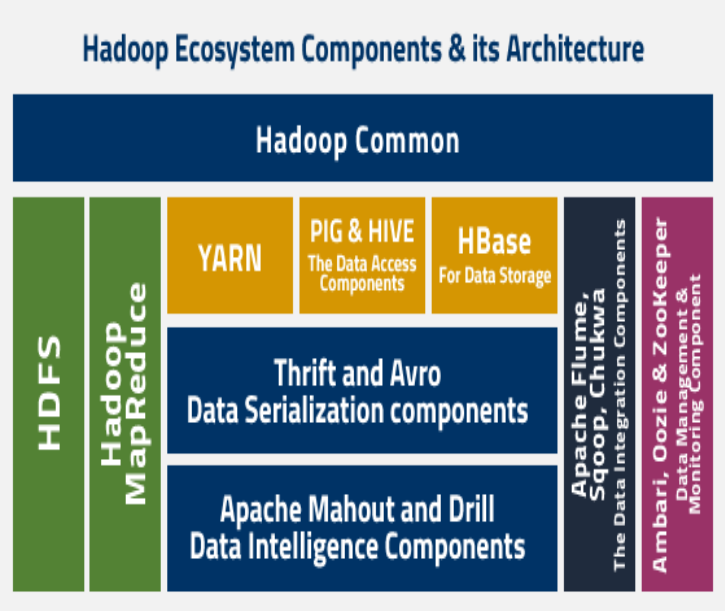

Figure 2 Architecture of Hadoop Ecosystem

Hadoop Ecosystem mainly consists of HDFS and Map reduce programming. Figure 2 shows the Architecture of Hadoop Ecosystem. Along with HDFS and MapReduce it contains Yarn, Pig, Hive, Hbase, Ambari, Oozie and ZooKeeper.

HADOOP DISTRIBUTED FILE SYSTEM

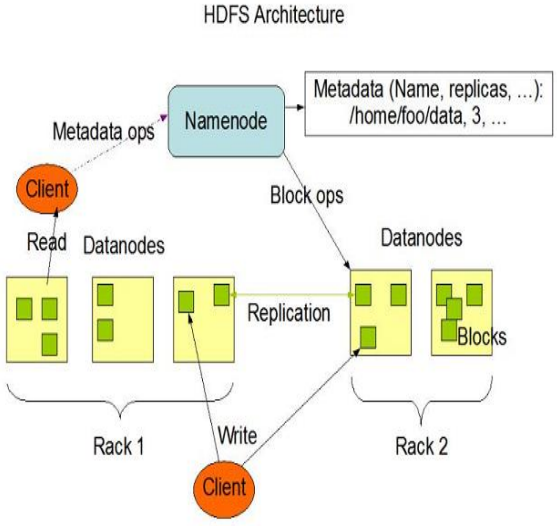

Figure 3 Architecture of HDFS

Figure 3 explains the HDFS architecture and distributed file system, storing a huge number of files by using commodity hardware. HDFS store data on thousands of servers. It has master-slave architecture. The HDFS stores a each file in three copies on different hosts. Each block has 64MB. The HDFS is mainly divided into three modes- Name node, Data node and Client node [6, 11].

MapReduce is programming model introduced by Google and it's managed by Apache Hadoop. MapReduce is a tool which is used to analysis the data it may structured or unstructured form. In this data is split in to small pieces and then get a original data. This is faster than other tools such as spark, drill etc [6, 11].

7. CONCLUSION

Big Data is nowadays in great demand on the marketplace. In recent years data is generated very fast, to analyze this data is a challenging to common man. In this paper discussed the big data characteristics, big data challenges and brief about hadoop technology. In terms of variety, Value, veracity, volume, velocity this paper revolves around the big data and their functionality. A big data technique provides the more privacy and security for data compared to traditional techniques. A big data platform are process the analytical tools and examine the data and shows the visualization patterns, these results are useful for business development and new trends in market etc. The complexities of big data cannot only be calculated in storage units, work on the social effects of machine learning and reflects a modern reformulation of priorities. This review will be useful for further progress and Improvement of Big Data Analytics from diverse research perspectives.

ACKNOWLEDGEMENTS

The authors would like to express sincere thanks for the encouragement and constant support provided by the School of Computing and Informatics, COET, Dilla University, Dilla, Ethiopia during this work.

REFERENCES

[01] Anuradha, J. "A brief introduction on Big Data 5Vs characteristics and Hadoop technology." Procedia computer science 48 (2015): 319-324.

[02] Acharjya, Debi Prasanna, and K. Ahmed. "A survey on big data analytics: challenges, open research issues and tools." International Journal of Advanced Computer Science and Applications 7, no. 2 (2016): 511-518.

[03] Priyanka, K., and Nagarathna Kulennavar. "A survey on big data analytics in health care." International Journal of Computer Science and Information Technologies 5, no. 4 (2014): 5865-5868.

[04] Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat ,"Research Paper on Big Data and Hadoop";International Journal of Computer Science And Technology, Vol. 7, s sue 4, Oct – Dec 2016.

[05] Oussous, Ahmed, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. "Big Data technologies: A survey." Journal of King Saud University-Computer and Information Sciences 30, no. 4 (2018): 431-448.

[06] Beakta, Rahul. "Big Data And Hadoop: A Review Paper." International Journal of Computer Science & Information Technology 2, no. 2 (2015): 13-15.

[07] Nagdive, Ashlesha S., and R. M. Tugnayat. "A review of Hadoop ecosystem for bigdata." Int. J. Comput. Appl 180, no. 14 (2018): 35-40.

[08] Bhagavatula, VS Narayana, S. Srinadh Raju, S. Sudhir Varma, and G. Jose Moses. "A Survey Of Hadoop Ecosystem As A Handler Of Bigdata." International Journal of Advanced Technology in Engineering and Science, Vol. No.4, Issue NO.08, August 2016.

[09] Rucha S Sheloadkar, Himanshu U Joshi,Survey Paper On Big Data Analytics And Hadoop, International Journal OF Engineering Sciences & Management Research, ISSN 2349-6193, January, 2017.

[10] Yadav, Dharminder, and Umesh Chandra. "Modern Technologies of Big Data Analytics: Case study on Hadoop Platform." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 6, no. 4 (2018): 044-050.

[11] Revathi.V, Rakshitha.K.R, Sruthi.K, Guruprasaath;Big Data With Hadoop – For Data Management, Processing And Storing; International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 08, Aug -2017

[12] Bharti Kalra, Anuranjan Misra,D. K. Chauhan; Analysis of Data Using Hadoop and Mapreduce; International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015

[13] Urmila R. Pol,Big Data Analysis Using Hadoop Mapreduce, American Journal of Engineering Research (AJER), Volume-5, Issue-6, pp-146-151, 2016

[14] Gayatri Kapil, Alka Agrawal, and R. A. Khan; A Study of Big Data Characteristics, DOI: 10.1109/CESYS.2016.7889917

[15] Elisa Bertino,Big Data – Opportunities and Challenges, 2013 IEEE 37th Annual Computer Software and Applications Conference, DOI 10.1109/COMPSAC.2013.143, 2013

[16] Nasser T, Tariq RS; Big Data Challenges, Journal of Computer Engineering & Information Technology, doi:http://dx.doi.org/10.4172/2324-9307.1000135, Volume 4, Issue 3, 1000135.

[17] Bolón-Canedo, Verónica, Beatriz Remeseiro, Konstantinos Sechidis, David Martinez-Rego, and Amparo Alonso-Betanzos. "Algorithmic challenges in big data analytics." ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017.

[18] K S Ananda Kumar.2020.A Kernel Oriented Controller Modelling System to Optimize the Convergence Performance in Big Data Analytics, IPI, Application No.202041010279,13/03/2020.

[19] Rajendran, Praveen Kumar, A. Asbern, K. Manoj Kumar, M. Rajesh, and R. Abhilash. "Implementation and analysis of MapReduce on biomedical big data." Indian Journal of Science and Technology (2016): 31.

[20] Alabdullah, Bayan, Natalia Beloff, and Martin White. "Rise of Big Data–Issues and Challenges." In 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1-6. IEEE, 2018.

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