Big Data and Hadoop Architecture: A Review Paper

Krish Nandan Das[1,*], Nikita Madaan[2]
Department of Mathematics, Chandigarh University
nandandaskrish1@gmail.com[1], nikita.e12455@cumail.in[2]
*Corresponding Author

Abstract. In the world of information, the term Big Data comes with new opportunity and challenges to deal with massive amount of data. Big data terms as large, diverse sets of information that grow at exponential rates. 90% of data is come within past 2-3 years. So, it is big responsibility to take care of this data and retrieve some useful information from it. To find useful information from huge amount of data, we need to analyze the data. In this review paper presents overview of big data, use of Hadoop architecture in big data, scope for the future research. This paper discusses about the big data types after that we discuss Hadoop and its mechanism.

Keyword: Big Data, Hadoop, MapReduce, HDFS, Yarn

Introduction

Big Data refers to data that is extremely large in size. Normally, we work with data in the MB (WordDoc, Excel) or GB (Movies, Codes) range, but data in the Peta bytes (1015 bytes) range is referred to as Big Data. It is estimated that about 90% of today's data was created in the last three years [1]. The researches says that the global aggregated big data economy has reached US$220.2 billion in 2013, and expected and to occur USD 401.2 billion by 2028. Now to it is time to build a system that can handle this much data for better performance to compute, process and analyze large-scale data [2-3]. There are three key features of Big Data, and any data that meets these criteria will be considered Big Data. It's the result of combining the four V's Volume, Variety, Veracity and Velocity

Volume: Volume means quantity of data or quantity of data generated in fraction of time by user or any organization via online mode or offline mode. Nowadays Machine generate massive amount of data. 97 zettabytes of data generated in 2022 which is more than 600 times in data generated from 2005. According to Survey of Statista Data will take a jump from 97 zettabytes to 180 zettabytes over the next three years [3]. The data should be massive in size. Big Data provides a solution for managing enormous amounts of data in the Terabyte or Petabyte range. We can easily and successfully conduct CRUD (Create, Read, Update, and Delete) operations on BigData.

Velocity: And second feature of Big Data is velocity that's mean the analysis of internet data. Velocity is the speed of data which is produced by user and processed. For example, content upload on YouTube, Facebook and twitter etc. is responsible for faster access to data. For instance, today's social media requires a rapid flow of data in a short period of time, and BigData is the greatest option. As a result, another feature is velocity, which refers to the rate at which data is processed.

Variety: Another important feature of Big Data is Variety. Variety means the type of Data. Data can be any form such as audio, video or social media data, Text, numerical data etc. on youtube 3.7 million video is uploaded per day. youtube has 2.5 billion active users in 2021 and seen rapid grow in data on it.

Veracity: Veracity is defined as data correctness or unpredictability. Due of inconsistencies and incompleteness, data are unclear.

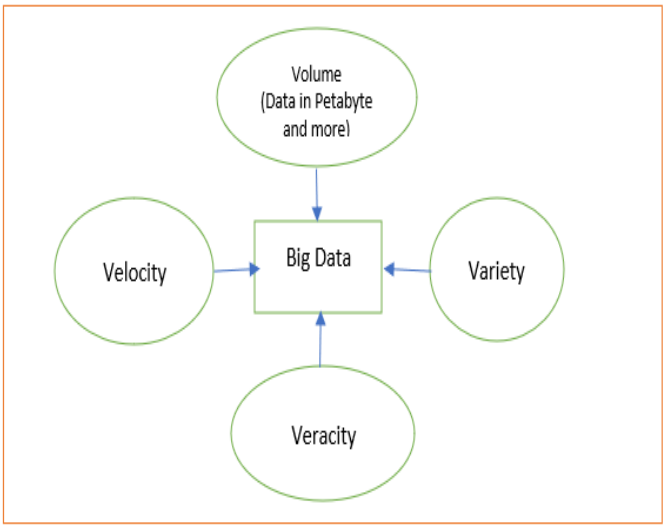

Figure 1 4 V's of Big Data

Challenges with Big Data Processing

Managing massive amounts of data is one of the most difficult aspects of Big Data. Nowadays, data is ingested into a system from a multitude of sources. As a result, correctly managing it is a huge challenge for businesses. To generate a report containing the last 20 years of data, for example, a system must save and retain the previous 20 years of data. Only relevant data should be entered into the system in order to provide an accurate report. It should not contain any useless or unneeded data; otherwise, firms will face a significant issue in managing such a large volume of data. The synchronization of diverse sorts of data is another problem with this technology [4]. As we all know, Big Data supports structured, unstructured, and semi-structured data from a variety of sources, making synchronization and data consistency problematic [5]. The second difficulty that businesses are confronted with is a scarcity of professionals who can assist them and implement solutions to the problems they are encountering in the system. In this industry, there is a significant talent shortage. Managing the compliance component is costly

Heterogeneity and Incompleteness of Data: Humans can tolerate a lot of variability when they ingest information. In reality, natural language's complexity and variety can add important depth. Machine analysis algorithms, however, do not comprehend nuance and instead anticipate uniform input. As a result, data must be meticulously organized as the first stage in (or before) data analysis. Computer systems function best when they can store a number of identically sized and constructed things. Effective semi-structured data representation, access, and analysis required further work.

Scale: Big Data, as the name implies, refers to enormous data sets. Managing massive data collections has been a major concern for decades. Earlier, this issue was resolved by faster CPUs, but as data volumes grow nowadays, processing speeds remain unchanged. The world is shifting to use cloud computing, and as a result, a lot of data is being produced [6]. The fast rate of data growth is posing a difficult dilemma for data analysts. To store the Data, hard discs are employed. Their I/O performance is slower. However, solid state drives and other technology have supplanted hard discs in recent years. A new storage system should be developed because these don't operate at a slower rate than Hard disks.

Timelines: Speed is the opposite of big. The length of the analysis will increase with the size of the data set to be processed [7]. A system that can successfully handle size will probably also be designed to process a given size of data set more quickly.

However, when someone uses the word "velocity" in relation to big data, it refers to more than just this speed. Instead, there is a problem with the acquisition rate.

Privacy: Another major issue that is raised in the context of big data is data privacy. There are tight legal restrictions on what can be done with electronic health records. Regulations, notably in the US, are less strict for other types of data. However, there is a lot of public concern about the improper use of personal data, particularly when data from many sources are linked together. In order to fully achieve the promise of big data, managing privacy must be approached from both a technological and a sociological standpoint.

Human collaboration: Despite the enormous progress made in computational analysis, there are still a lot of patterns that people can easily spot but that computer algorithms struggle to find.

Analytics for Big Data should ideally not be entirely computer but rather be specifically designed to involve humans [8-9]. This is what the emerging area of visual analytics is aiming for, at least in terms of the pipeline's modelling and analysis stage. In today's complex world, it frequently requires a team of specialists from many fields to fully comprehend what is happening. A big data analysis system needs to accommodate input from various human experts as well as collaborative results exploration. When it is too expensive to gather the full team in one location, this many expertise may be dispersed throughout time and place. This distributed expert input must be accepted by the data system, which must also facilitate their cooperation.

Opportunities of Big Data

Media: Big data is being used by the media to promote and sell items by focusing on internet users' interests. For instance, when it comes to social media posts, data analysts first count the quantity of posts before analyzing user interest. Obtaining favorable or unfavorable ratings on social media is another method.

Technology: Nearly every prestigious company, including Facebook, IBM, and Yahoo, has embraced big data and is investing in it. Facebook manages 50

billion user photographs. Google processes 100 billion searches per month. These statistics indicate that there are numerous options available on the internet and in social media.

Healthcare sector: Eighty percent of medical data is unstructured, according to IBM Big data for healthcare. Healthcare organizations are implementing big data technology to obtain all of the patient's information. Big data analysis and the adoption of certain technologies are needed to enhance healthcare and save costs [10].

Hadoop & Hadoop Architecture

The Apache Hadoop software library is a framework that uses basic programming principles to enable for the distributed processing of massive data volumes across clusters of machines. It's built to expand from a single server to thousands of devices, each with its own computation and storage capabilities. Rather than relying on hardware to provide high availability, the library is designed to identify and handle problems at the application layer, allowing a highly available service to be delivered on top of an all cluster of computers that may fails [10]. Basically, Hadoop have two major layer named processing/computational layer (MapReduce), and Storage layer (HDFS).

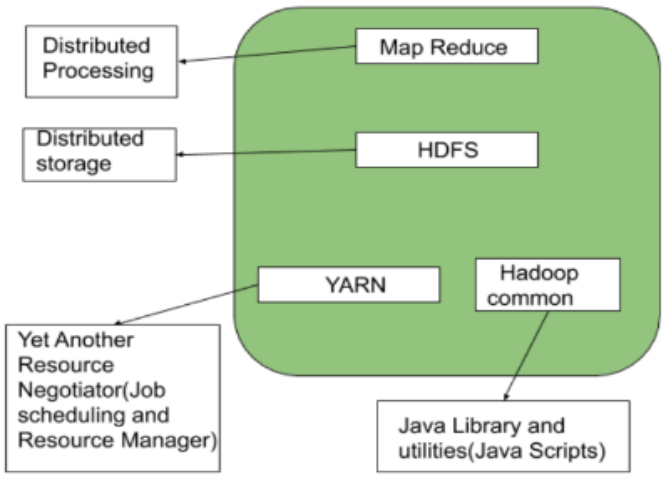

Figure 2: Hadoop Architecture

Hadoop Distributed file system (HDFS)

HDFS is one of the major components of Hadoop. HDFS is a distributed file system that runs on commodity hardware and can handle massive data collections. It is used to scale an Apache Hadoop cluster from a few nodes to hundreds (or even thousands) of nodes [11].

Consider a file containing the phone numbers of everyone in the United States; the numbers for persons with surnames beginning with A, B, and so on might be placed on server 1, B, and so on. With Hadoop, bits of this phonebook would be stored across the cluster, and your software would need the blocks from every server in the cluster to reconstruct the whole phonebook. By default, HDFS replicates these smaller parts onto two more servers to provide availability in the event of a server failure. (Data redundancy can be enhanced or decreased per-file or for the entire environment; for example, a development Hadoop cluster usually doesn't require any [12].

NameNode: In a Hadoop cluster, NameNode serves as a Master and directs the Datanode (Slaves). The primary purpose of Namenode is to store metadata, or information about metadata. Transaction logs that record user activity in a Hadoop cluster can serve as meta data.

In order to locate the nearest DataNode for faster communication, Namenode keeps information about the location (Block number, Block ids) of DataNodes as part of Meta Data [13]. DataNodes are given instructions by Namenodes for actions like create, replicate, and remove.

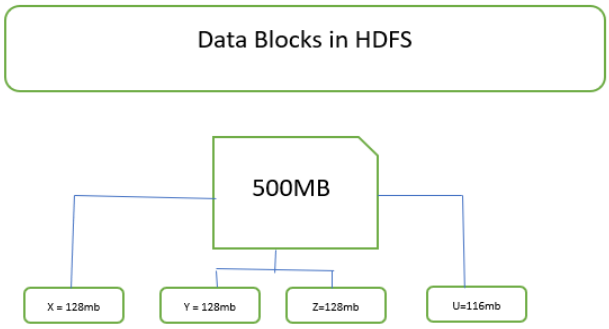

Figure 3 File Hadoop file distribution system

DataNode: Working as a Slave is DataNodes, which can range in number from 1 to 500 or even more, are mostly used for storing data in a Hadoop cluster. The Hadoop cluster can hold more data the more DataNodes it has. Therefore, it is recommended that the DataNode have a high storage capacity in order to store a lot of file blocks.

MapReduce: MapReduce is a distributed computing processing technique and programme model based on Java. Map and reduce are two important tasks in the MapReduce algorithm [15]. Map turns a set of data into another set of data by breaking down individual pieces into tuples (key/value pairs). Second, there's the reduction job, which takes the result of a map as an input and merges the data tuples into a smaller set. The reduction work is always executed after the map job, as the name MapReduce suggests.

MapReduce's main advantage is that it's simple to expand data processing over several computing nodes. Mappers and reducers are the data processing primitives in the MapReduce model. It can be difficult to break down a data processing application into mappers and reducers [16-17]. Scaling an application to run over hundreds, thousands, or even tens of thousands of servers in a cluster is only a configuration modification once we build it in MapReduce form. Many programmers have been drawn to the MapReduce approach because of its simple scalability.

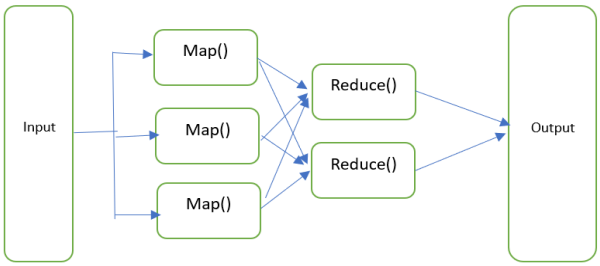

Figure 4: MapReduce

YARN:

MapReduce runs on a framework called YARN. The two tasks that YARN carries out are resource management and job scheduling. The goal of job scheduling is to break large tasks down into smaller ones so that each job can be distributed across different slaves in a Hadoop cluster, maximising processing [18]. The job scheduler also keeps track of the jobs' priorities, dependencies on one another, importance levels, and other details like job timing. To manage all the resources made available for running a Hadoop cluster, Resource Manager is used.

Common Utilities

Hadoop common, often known as the "common utilities," is nothing more than our Java library, java files, or the Java scripts that we require for all the other components found in a Hadoop cluster. For the cluster to function, HDFS, YARN, and MapReduce use these tools. Hadoop Common confirms that hardware failure in a Hadoop cluster is frequent, necessitating an automatic software solution by the Hadoop Framework.

Advantages and Disadvantages of Hadoop

Advantages

Range of Data Source: Both organised and unstructured data will be gathered from diverse sources. Social media sites or even email exchanges could be the sources. The process of converting all the gathered data into a single format would take a lot of time. Hadoop reduces this time since it can extract useful information from any type of data [19]. Additionally, it performs a wide range of tasks, including fraud detection and data warehousing.

Cost Effective: Large amounts of money from the companies' profits had to be spent on data storage. To make room for fresh data, they occasionally had to erase huge amounts of raw data. In such situations, it was possible to lose crucial information. This issue was totally resolved using Hadoop. It is an affordable option for data storage needs.

Speed: Every organisation makes use of a platform to do tasks more quickly. The company's data storage requirements are met by Hadoop. It utilises a system of storage where A distributed file system is used to store data.

No chance of loss Data: Hadoop automatically makes numerous copies of the data that is stored there. This is done to make sure that data is not lost in the event of a failure. Hadoop is aware that the data kept by the business is crucial and shouldn't be lost unless the business decides to trash it [20].

Disadvantages

Lack of preventive measures: It is required to take the essential security precautions while managing sensitive data that a corporation has gathered. The security precautions in Hadoop are by default turned off. The person in charge of the data should be aware of this and take the necessary precautions to protect the data.

Risk Functioning: One of the most popular programming languages is Java. Due to the ease with which cybercriminals can exploit Java-based frameworks, it has also been linked to numerous communities. One such framework that is totally Java-based is Hadoop [21]. As a result, the platform is weak and capable of causing unexpected harm.

Application

Both business organisations and researchers can benefit greatly from big data in their efforts to identify data trends in large data sets. Data mining is the process of obtaining meaningful information from enormous amounts of big data. On the internet, there is a vast amount of information in the form of text, numbers, social media posts, photographs, and videos. 97 zettabytes data is created by 2022, which is 90 times more than the year 2005 [22]. We must implement a new, efficient data mining system in order to evaluate this data and obtain pertinent information for security, health, education, etc. Big data can be mined using a variety of ways, some of which include:

Classification Analysis: It is a methodical procedure for gathering crucial data and metadata information. The data can also be clustered using classification.

Cluster Analysis: It is a methodical procedure for gathering crucial data and metadata information. The data can also be clustered using classification.

Evolution Analysis: The basic purpose of genetic data mining is to extract information from DNA sequences. However, it can be utilised in banking to forecast stock exchange prices using historical time series Data [23].

Outlier Analysis: There are some observations and item identifications that are made that do not reveal a pattern in the data set. This is employed in financial and medical issues.

Conclusion

A new era of big data has begun. The four Vs, volume, velocity, veracity and variety of big data, are discussed in the paper along with the idea of big data. The report also focuses on issues with big data processing. To process Big Data effectively and quickly, several technical issues must be solved. At all phases of the analysis pipeline, from data gathering to result interpretation, the obstacles include not just the obvious ones of scale but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualisation. Since these technical difficulties are prevalent across a wide range of application domains, it would not be cost-effective to address them in the context of a single domain. The article discusses Hadoop, an open-source programme used to process Big Data.

References

[1] Pujari, V., Sharma, Y.K. and Rane, R., 2016. A Review Paper on Big Data and Hadoop.

[2] Duggal, Puneet Singh, and Sanchita Paul. "Big data analysis: challenges and solutions." In *International conference on cloud, big data and trust*, vol. 15, pp. 269-276. 2013.

[3] Agrawal, Divyakant, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Johannes Gehrke et al. "Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States." *Computing Research Association, Washington* (2012).

[4] Beakta, Rahul. "Big data and hadoop: A review paper." *International Journal of Computer Science & Information Technology* 2, no. 2 (2015): 13-15.

[5] Mathew, Prabha Susy, and Anitha S. Pillai. "Big Data solutions in Healthcare: Problems and perspectives." In *2015 International conference on innovations in information, embedded and communication systems (ICIIECS)*, pp. 1-6. IEEE, 2015.

[6] Thakur, Vishesh Kumar, V. Harshit, R. Utkarsh, J. Parmod, and C. Amit. "Review Paper on Big Data Analytics." *International Journal for Research in Applied Science and Engineering Technology* 8, no. 6 (2020): 785-788.

[7] Garg, Naveen, Sanjay Singla, and Surender Jangra. "Challenges and techniques for testing of big data." *Procedia Computer Science* 85 (2016): 940-948.

[8] Gudipati, Mahesh, Shanthi Rao, Naju D. Mohan, and Naveen Kumar Gajja. "Big data: Testing approach to overcome quality challenges." *Big Data: Challenges and Opportunities* 11, no. 1 (2013): 65-72.

[9] Smitha, T., and V. Suresh Kumar. "Applications of big data in data mining." *International journal of emerging technology and advanced engineering* 7, no. 3 (2013).

[10] Mridul, Mrigank, Akashdeep Khajuria, and Snehasish Dutta. "Kumar N "Analysis of Bidgata using Apache Hadoop and Map Reduce"." *International Journal of Advanced Research in Computer Science and Software Engineering* 4, no. 5 (2014).

[11] Lakshkar, Sushma, Geet Kalani, and Vinod Todwal. "A Catholic Research on Big Data and Hadoop Environment." *International Journal of Computer Applications* 975 (2015): 8887.

[12] BalaAnand, M., N. Karthikeyan, S. Karthik, and C. B. Sivaparthipan. "A survey on BigData with various V's on comparison of apache hadoop and apache spark." Advances in Natural and Applied Sciences 11, no. 4 (2017): 362-370

[13] D. Goldston, Big Data: Data Wrangling, Nature, Vol. 455, No. 7209, pp. 15, September, 2008.

[14] A. Oguntimilehin, E. O. Ademola, A Review of Big Data Management, Benefits and Challenges, Journal of Emerging Trends in Computing and Information Sciences, Vol. 5, No. 6, pp. 433-438, June, 2014

[15] J. Liu, E. Pacitti, P. Valduriez, A Survey of Scheduling Frameworks in Big Data Systems, International Journal of Cloud Computing, Vol. 7, No. 2, pp. 103-128, January, 2018.

[16] Y. Chen, M. Zhou, Z. Zheng, Learning Sequence-Based Fingerprint for Magnetic Indoor Positioning System, IEEE Access, Vol. 7, pp. 163231-163244, November, 2019.

[17] Ramesh, B. (2015). Big data architecture. *Big Data: A Primer*, 29-59.

[18] Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big data service architecture: a survey. *Journal of Internet Technology*, *21*(2), 393-405.

[19] Demchenko, Y., Oprescu, A., Ngo, C., Grosso, P., & de Laat, C. Towards Defining Big Data Architecture Framework. *University Van Amsterdam*.

[20] Ivanov, T., & Singhal, R. (2018, April). Abench: Big data architecture stack benchmark. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering* (pp. 13-16).

[21] Petrillo, A., Picariello, A., Santini, S., Scarciello, B., & Sperli, G. (2020). Model-based vehicular prognostics framework using Big Data architecture. *Computers in Industry*, *115*, 103177.

[22] Tan, C., Sun, L., & Liu, K. (2015). Big data architecture for pervasive healthcare: a literature review. ECIS.

[23] Inmon, W. H., & Linstedt, D. (2014). *Data architecture: a primer for the data scientist: big data, data warehouse and data vault*. Morgan Kaufmann.