International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

A Study on Big Data Analytics

Souvik Baruah[1], Aritra Kaushik[2]

[1,2]*Students, Dept. of Computer Science and Engineering, Jorhat Engineering College, Jorhat, Assam, India*

Abstract - *With the arrival of new tools and technologies the data produced by mankind is increasing by leaps and bounds. This has marked the beginning of the era of Big Data. This in turn, has put forward some new challenges for input, processing and output of such large-scale data. This paper focuses on the limitations of traditional approaches to handle such data and lays emphasis on the tools which can handle big data effectively. One major tool used in processing big data is the Apache Hadoop Framework. This paper presents the major components of the Hadoop framework and it's working process.*

Key Words: Big Data, Data Analytics, Hadoop Framework

1. INTRODUCTION

Big data is a term which is used for data that goes beyond the capabilities of traditional database systems. The characteristics of big data is that it is too large, the data is generated at a very fast speed and generally the data is not in suitable form for the database system. Data and big data are two different terms, hence the processing power required for them are different as well. The process of digitization is extremely fast and because of that production of data is almost in digital form and the data generated is of huge size even exceeding Exabyte. In accordance, the computing power of the modern systems is significantly faster than the traditional systems, but the analysis of data on such a large scale still remains a very crucial factor.

2. TRADITIONAL APPROACH OF DATA MINING

Data mining is the process of extracting useful and meaningful information from data. Several different algorithms are there for the different phases of the data mining process. These algorithms and tools allow the users to predict future trends. The process also known as the data mining pipeline, is represented in the figure below:

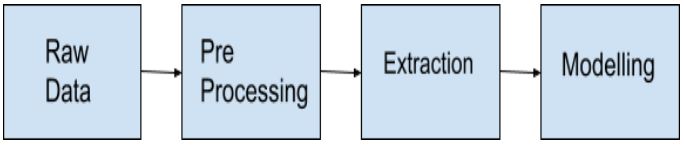

Fig -1: Traditional Data Mining Process

The importance of the approach while implementing the process is on the algorithms processing the data. The information is expected to flow smoothly from the beginning till the end and the algorithms are usually run in the same application.

Data Preprocessing is an often neglected but a very important step in Data Mining. Various steps are involved in preprocessing which are cleaning, integration, transformation and reduction. After the preprocessing of the data, different data mining techniques can be applied for further data analysis.

Most of the traditional data mining methods cannot be applied to big data because of the following reasons:

- The data size involved in big data analysis is huge - in the range of petabytes. Traditional databases find it challenging to process these sorts of data.
- In big data analysis, data is usually generated at very high velocity and the traditional RDBMS lacks high speed data retention because it is designed for steady data instead of very fast growth.
- Most of these traditional methods cannot produce the analysis dynamically based on the input.

After these processes, the output can be analyzed with a number of measures such as errors, result accuracy, speed of computation and memory consumption.

2.1 Big Data

Big data is used to describe large volumes of data, which can be either structured or unstructured. It refers to data so huge, fast or complex, that it is difficult to process using traditional methods. Big data is stored in databases and is analysed using tools designed to operate large, complex data sets. Some areas where big data is generated and used are given below.

- **Social media sites and applications data**: Sites and applications such as Facebook, Twitter, Instagram etc. have a lot of information posted by people worldwide. The data is in petabytes which is almost generated hourly. Big data analysis helps in storing and analysing this data.
- **Healthcare industry data:** The medical industry generates a great amount of data. Big data analysis in this industry helps to reduce cost of treatment and helps reduce the chances of performing unnecessary diagnosis.
- **Stock exchange data**: Stock exchanges contain a great deal of data about the shares of various companies. Big data analysis can help predict and prevent future market crashes, it also levels the

playing field by stabilising the market, and it estimates the outcomes and returns accurately.

- **Flight and aviation data**: A single jet engine of a commercial flight can generate 10+ terabytes of data in only 30 minutes of flight time, with almost thousand flights per day. The data which is generated may reach up to petabytes which is almost impossible to manage using traditional methods, hence big data analysis is used.

2.2 Characteristics of Big Data

Big data is very difficult to handle as it is emerging as one of the most rapidly growing technologies. Several applications are dependent on these data and hence, if anything goes wrong, it may be fatal for an organisation and its partners. Big data can be described by the three basic characteristics also known as the 3Vs i.e., velocity, volume and variety, which implies that the data is generated very quickly, is huge in size and the data is not in a homogeneous format. The details of the 3Vs are given below:

1. Velocity

Velocity in big data analysis, refers to the speed at which data is being created. In general terms, we say that it comprises the rate of change, linking of incoming data sets at varying speeds to the databases, and activity bursts. For instance, nearly 350 million photos are uploaded to Facebook daily and almost 500 million tweets are posted on Twitter. This is a massive explosion of data on these platforms at a very quick speed. Big data helps these corporations to withhold this explosion and accept the coming flow of the data. At the same time, it also helps to process the data quickly such that there is no bottleneck.

2. Volume

Volume describes the amount of data generated by organizations or individuals. The volume hands out the major challenge for the traditional approaches of data processing. It encourages the use of distributed computing. Size of data plays a very important part in determining value out of data. Also, whether the data can be determined as Big data or not is determined by its volume.

3. Variety

The third element of Big Data is its variety. The data can be stored in different formats like databases, images, csv, or in a simple text file or document file as well. Sometimes the data is not in the traditional format. It may be in the form of video, SMS, pdf or something we might have not thought about. The data is primarily in the unordered and unstructured format which requires preprocessing which will turn the data into meaningful resources either for an individual or an organization. The data can be generated from a variety of sources. While generating processed data from the source data there may sometimes occur loss of information.

Keeping the 3Vs aside, different characteristics were added to explain big data[1][2] such as value, venue, variability, vocabulary and validity. The report published by Wikibon Research [3] indicates that the marketing of big data was valued at around $169 billion in 2018. The same report indicates that it will grow to $ 274 billion in 2022.

On analysing the information given above it becomes necessary to have an insight on big data along with the tools required to handle big data efficiently.

3. TOOL SELECTION

Big data not only focuses on data alone but it also includes the tools and technologies which are used to handle large scale data in an effective manner. The vast amount of data needs new technologies and mechanisms for storage, processing, management and analysis. It is accepted that big data is too large for traditional relational database management systems (RDBMS), hence new tools are required that includes a wide range of database systems, file systems, programming languages (*see Fig. Adaptive Big Data Value Chain*). There are various tools required for handling big data, each of which comes with its different characteristics and qualities. The selection of the software tool [4] depends on the tool's capability and various dimensions listed on the table given below (*Table 1*).

Table -1: Dimensions of Big Data

Dimension	Description	Issues in selection of dimension
Data Structurization	It involves data acquisition and cleaning the data on our own or making it available to the marketplace.	• Cost of data cleaning • Time required for data cleaning • Data quality
Data Transfer	It involves processing either by transporting the data or without transporting it.	• The size of data to be transferred • Time required for transferring • Cost of transfer • Data locality
Big Data Solution	It is generally a software, an appliance or a cloud-based implementation. It can also be implemented as	• Privacy • Regulation • Data locality • Human resources

	a combination of all of the above.	

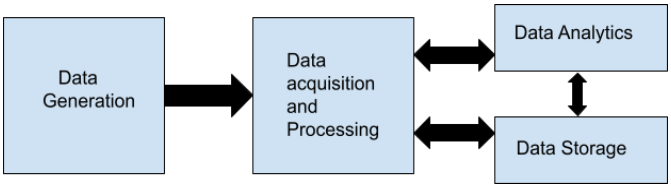

Fig -2: Adaptive Big Data Value Chain [5]

Big data analysis is a special indicator for the huge quantity of data, which increases extremely in size and rapidly with time. Big Data analysis tools can be defined as software tools for analysing, processing, and extracting data from a large and complex data set with which traditional management tools can never deal with. These tools and technologies can be used in storing the data and later analysing it.

3.1 Technologies for Handling Big Data

1. NoSQL Database

NoSQL includes a wide variety of different big data technologies in the database, which are developed to design modern applications. It highlights a non-SQL or non-relational database providing a method for data acquisition and recovery. It stores unstructured data and offers faster performance and flexibility while addressing various data types. Examples of NoSQL databases are MongoDB, CouchDB and Cassandra. It provides the facility to generate patterns and trends without need for additional infrastructure. It also provides design integrity, easier horizontal scaling and control over opportunities in a range of devices. It uses data structures that are not the same as concerning databases by default, which speeds up NoSQL calculations. Facebook, Google, Twitter, and similar companies store user data in terabytes daily.

2. Artificial Intelligence

Artificial Intelligence is a type of technology that deals with the development of intelligent computable machines capable of doing different tasks typically requiring human intelligence. It is a multidisciplinary branch of computer science; it considers a number of approaches like Machine Learning and Deep Learning to make a great curve in most technological industries. AI is changing the way existing Big Data Technologies are implemented. Some common examples of AI are Apple's Siri and Amazon's Alexa.

3. Massively Parallel Processing and MapReduce

Massively Parallel Processing (MPP) database is a type of database where the data and processing capability are split up among several different nodes, with one major node and one or many compute nodes. MPP processors or nodes communicate with one another using a messaging interface connected via common data paths. Generally, the partitioning of this database and allocating work to different processors is a very hectic and complex process.

MapReduce is a framework and programming paradigm which is used for the processing of big data. MapReduce as suggested by its name works in two phases i.e, Map and Reduce. In the first phase, Map tasks deal with splitting and mapping the data whereas Reduce tasks sort and reduce the data or, in layman terms it breaks the large data into smaller volumes.

In MPP, just like in MapReduce, processing of data is distributed across many different processors. These processors further process data parallelly, and later all the outputs are stitched together to give the final result set.
Even though MPP and MapReduce are very similar as mentioned above but, they are used for different purposes because they have many different characteristics as given in Table 2.

Table -2: MapReduce and MPP Characteristics

MapReduce	MPP
Data loading is faster.	Data loading is slower
The control mechanism of MapReduce is based on Java code.	Querying is done using Structured Query Language (SQL).
Querying is more complex compared to MPP.	Querying is easier.
It is deployed on clusters of commodity servers that use commodity disks.	Required expensive specialized hardware.
It performs good on structured and semi-structured data. It can capture, store and refine such data in native format.	It performs good on huge amount of structured data. It can easily perform fast, iterative queries and analytics on such data.

4. Storage

For storage of data in big data analysis, Amazon Simple Storage Service (Amazon S3) and Hadoop Distributed File System (HDFS) are two of the most used frameworks. Amazon S3 is a storage service for the world wide web or more commonly, the internet. It is a service provided by the Amazon Web Services (AWS). Some benefits of AWS S3 are it's durability, low cost of setup, and its ability to scale to larger forms of operations.

On the other hand, Hadoop Distributed Files System(HDFS) is a Java-based file system. It ensures scalable and reliable

data storage and it was specifically designed to incorporate large clusters of commodity servers.

If one wishes to get involved with big data analytics, he/she must possess sound knowledge of these two frameworks. Apart from these two there are many other frameworks such as Hadoop which are available to process big data. In the next section, the Hadoop framework is being explained in detail.

4. THE HADOOP FRAMEWORK AT A GLANCE

Hadoop is an open-source framework written in Java which is developed by Apache Software Foundation. It is used to create data processing applications which are executed in a distributed computing environment.[6]. Such applications constructed using Hadoop permits the appropriate handling and processing of huge datasets across clusters of computers using very basic programming models.

The traditional approach of handling Big Data makes use of a computer to store and process the data. A Relational Database Management System (RDBMS) such as Oracle Database, DB2 or MS SQL Server is used for data storage purposes. Certain software can then be used to interact with the database which processes the required data and gives it as output to the users for data analysis.

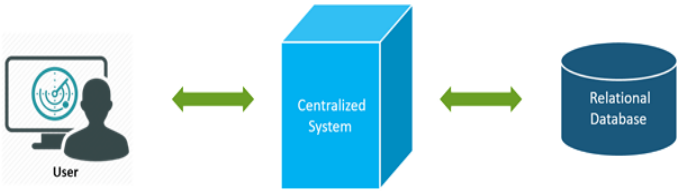

Fig -3: Traditional Approach to Store and Process Big Data.[12]

This approach works fine when the data to be handled is in small amounts. When an application requires low data storage and processing, the traditional approach is a good match. However, when the data to be dealt with is very huge, this approach does not hold good. It becomes a very difficult and time-consuming task to process such huge amounts of data through a traditional database server.

Google solved this problem by introducing a new algorithm known as MapReduce. This algorithm splits the input into a number of smaller parts and then it assigns them to multiple computers which are connected over the network. It then collects and integrates the respective results and forms the final output result. Hadoop utilizes this MapReduce algorithm to process Big Data in parallel on multiple CPU nodes which is being illustrated in the figure below:

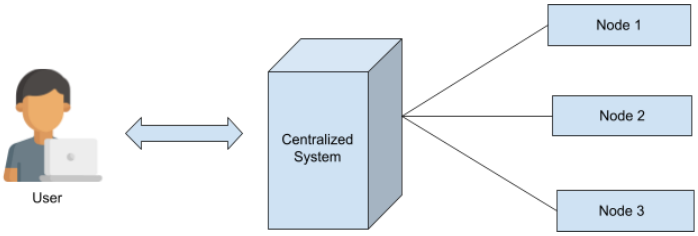

Fig -4: Working of MapReduce Algorithm to Store and Process Data in Parallel

4.1 Components of Hadoop

Hadoop is the most commonly used software framework to process Big Data. It uses parallel processing and distributed storage to manage Big Data. There are four main components of the Hadoop framework: Hadoop HDFS, Hadoop MapReduce, Hadoop YARN and Hadoop Common.

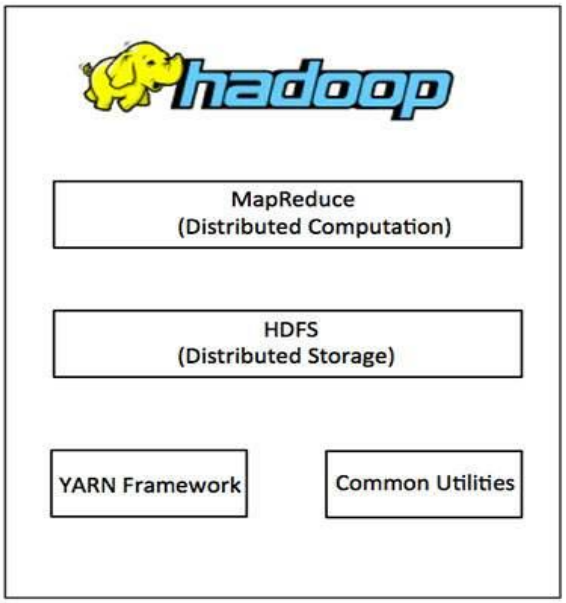

Fig -5: Core Components of Hadoop.[14]

1. Hadoop Distributed File System (HDFS)

HDFS basically serves as the storage unit of Hadoop. In HDFS, the data is stored in a distributed manner. Name Node and Data Node are the two main components of the Hadoop HDFS.[7]. The name node is called Master and the data nodes are called Slaves. Although there can be many data nodes, there exists only one name node which is always an enterprise server. HDFS is particularly designed to store large datasets in commodity hardware. Hadoop allows the use of commodity machines as data nodes.[7]

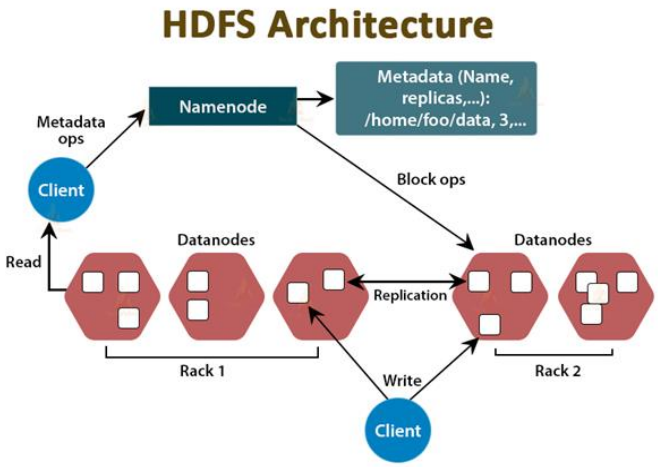

Fig -6: HDFS Architecture.[13]

2. Hadoop MapReduce

Hadoop MapReduce is the processing unit of Hadoop. It is a Java-based programming model created by Google that enables parallel processing of very large data sets. The MapReduce framework also uses the master-slave architecture like HDFS. The working of Hadoop MapReduce can be divided into two phases: the Map Phase and the Reduce Phase. At first, in the Map Phase the Map Function takes a set of data as input and transforms that into another set of data where elements are broken down into tuples or key/value pairs. After this, in the Reduce Phase the Reduce Function takes the output of a map task as input and then combines those tuples into a smaller set of records based on the key and changes the value of the key accordingly. The assignment of tasks of MapReduce is mainly handled by two daemons namely, Job Tracker and Task Tracker which are being shown in the figure below:

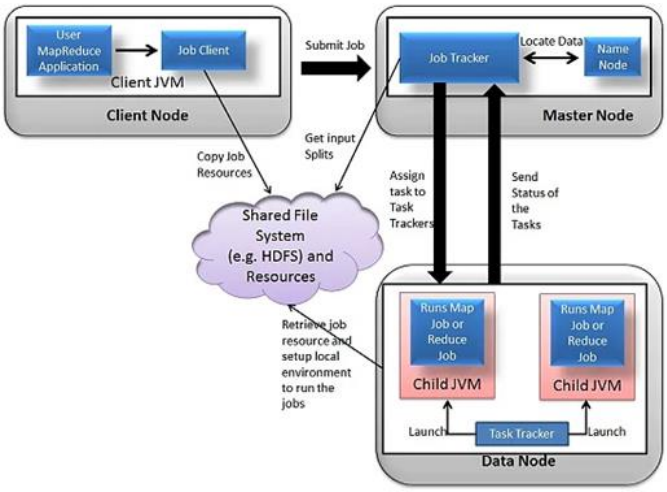

Fig -7: Delegation of Tasks of MapReduce.[8]

3. Hadoop YARN

YARN is the abbreviated form of Yet Another Resource Negotiator. It is basically a job scheduling and cluster resource management unit of Hadoop. It is viewed as the next generation of Hadoop's computing platform. It can be described as a large-scale, distributed operating system for big data applications.[9]. YARN improves use over more static MapReduce rules, that were delivered in early forms of Hadoop, through dynamic assignment of cluster resources.

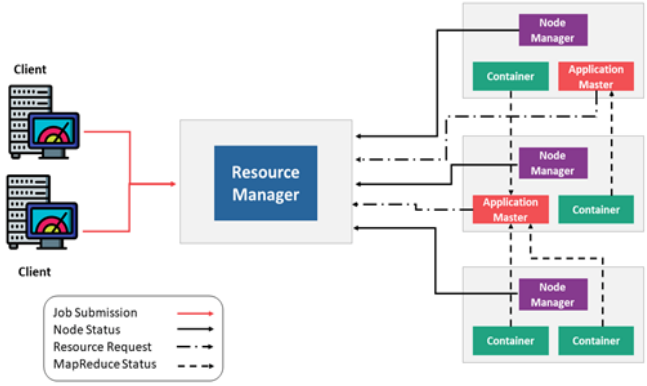

Fig -8: The YARN Architecture.[11]

4. Hadoop Common

Hadoop Common constitutes the Java libraries and the common utilities that are needed by other modules of Hadoop.[10]. It is a fundamental component of the Hadoop framework, alongside the Hadoop Distributed File System (HDFS), Hadoop MapReduce and Hadoop YARN. The libraries contain the essential scripts and files which are needed to launch Hadoop. The libraries also provide OS level abstractions.

4.2 Working of Hadoop

The working of Hadoop takes place in two phases. In the first phase, Hadoop job clients take the job which is submitted by an application or a user. The job submitted includes specification of output files in the distributed file system, input location, jar files that contain map and reduce functions and various configuration parameters related to the job. In the second phase, the job and configuration are submitted by a Hadoop job client to the MapReduce master known as JobTracker which then distributes the jar files to the slaves. In the last phase, tasks are executed according to MapReduce implementation by the slaves on various nodes. After this, the output of the reduce function is stored into the output files of the file system.

4.3 Benefits of Using Hadoop for Big Data

- **Scalable**: Hadoop functions in a distributed computing environment. This makes the addition of more servers very easy.
- **Diversity**: Hadoop HDFS has the capacity to store different formats of data which includes structured, unstructured or even semi-structured.
- **Speed**: Parallel processing, HDFS and MapReduce Model of Hadoop allows users to execute complex queries in only a couple seconds.
- **Resilient**: Data stored in a node is duplicated to other cluster nodes. This ensures fault tolerance.
- **Cost-effective**: Hadoop is an open-source software framework.
- **Compatibility**: Hadoop is compatible on all the platforms.

5. CONCLUSIONS

In this paper, we have performed a brief survey on Big Data Analytics. We presented the significance of Big Data and also explored the various tools and technologies that are used in analyzing Big Data. We have discussed the limitations of the traditional approach of processing Big Data and how the Apache Hadoop Framework can be used to handle Big Data effectively while overcoming the limitations of the traditional method. We have explored the Hadoop Framework in detail along with the working of its components and the benefits that it offers when used in handling Big Data.

REFERENCES

[1] Van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups, Tech. Rep. 2013. [Online]. Available: https://www.bigdata-startups.com/3vs-sufficient-describe-big-data/
[2] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online]. Available: https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.
[3] Peter B. Wikibon's 2018 Big Data and Analytics market share report. 2018. [Online]. Available: https://wikibon.com/wikibons-2018-big-data-analytics-market-share-report/
[4] Agneeswaran VS, Tonpay P, Tiwary J (2013) Paradigms for realizing machine learning algorithms. Big Data 1(4):207–214.
[5] Dennis L, David F, Gottfried V, Technology selection for big data and analytical applications. 2016. [Online]. Available: https://www.ercis.org/sites/ercis/files/structure/network/research/ercis-working-papers/ercis_wp_27.pdf
[6] https://www.guru99.com/learn-hadoop-in-10-minutes.html
[7] https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop
[8] https://www.dezyre.com/article/hadoop-ecosystem-components-and-its-architecture/114
[9] https://bigdata-madesimple.com/basic-components-of-hadoop-architecture-frameworks-used-for-data-science/
[10] https://www.techopedia.com/definition/30427/hadoop-common
[11] https://www.edureka.co/blog/hadoop-yarn-tutorial/
[12] https://intellipaat.com/blog/tutorial/hadoop-tutorial/big-data-solutions/
[13] https://techvidvan.com/tutorials/hadoop-architecture/
[14] https://www.tutorialspoint.com/hadoop/hadoop_introduction.htm