IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Review Paper on Big Data with comparative analysis of Hadoop

Priya

[1] Department of Computer Science, BSAITM, Faridabad, India

Available online at: www.ijcseonline.org

Accepted: 21/Jan/2019, Published: 31/Jan/2019

***Abstract*—** Big data is the complete data that is generated by human beings in daily life. Data is gathered from various sources and thus made useful for people by preprocessing it. For processing such huge amount of data (i.e. in terabytes and petabytes), specialized hardware and software is required. Thus, to store, manage, and process the increasing amount of data is really a challenging area of research and development in big data analysis. The objective of this paper is to explore the impact, challenges, architecture of big data and various tools associated with it. This paper also provides a comparative study of Hadoop distributed file system and traditional database and a platform to explore it at various stages

Keywords:-Hadoop, HDFS, Challenges, Big Data Analysis

I. INTRODUCTION

Big data is a potential and budding term that describes a large measurement of data and data is structured, semi-structured and unstructured. It is basically the data sets or combinations of data sets whose size, complexity, and rate of growth make them difficult to be captured, managed, processed or analysed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages. Introduction of Big Data and Hadoop provides an ease and flexibility to store and manage big data. Big Data can also be defined in terms of 4 Vs i: e. Volume, Velocity, Variety and Veracity. These are also considered as the characteristics of big data: -

Volume: The Volume represents the size of the data. Data is generated by machines, networks and human interaction on systems like social media (etc. and so on) and thus the data to be analysed is massive. For ex: 500TB data per day can be considered as big data.

Velocity is referred to the speed of data in terms of data generation, data processing as well as delivery of real-time data. Big data is used for time-sensitive processes, such as considering the example of catching fraud, as it streams into your enterprise in order to maximize its value

Variety: - As data is gathered from various sources, that makes data too big and it is in three forms i: e. structured semi-structured and unstructured forms. Usually the data is in unstructured form; whereas the traditional data which is supported by the RDBMS is only structured. So, its uses are limited and we cannot use this traditional data in big data.

Veracity -it can be defined as the truthfulness of data.

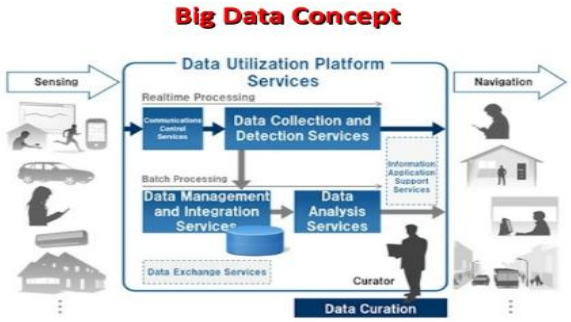

Figure 1: Big Data concept

Different types of big data and its sources are mentioned below in Table 1:

Table 1: Types of Big Data and its sources

Data Types	Sources	Formats
Structured	Business applications such as retail, finance, bioinformatics etc.	RDBMS, OLAP, Data warehousing
Semi Structured	Web Applications such as web logs, email, Webpages	XML, CSV, HTML, RDF
Unstructured	Images, Audio, Video, Sensor data, Blogs, Tweets etc.	User generated text content

© 2019, IJCSE All Rights Reserved

II. CHALLENGES AND OPPORTUNITIES

Millions of web pages on Internet provides information about Big Data. Big Data is the next big thing after Cloud computing. There are plenty of opportunities in Big data that usually deals in the fields of health, education, earth, and businesses. But to deal with the data having large volume using traditional model is very difficult. So, for efficient analysis of data we need to design some computing models.

A. Challenges with Big Data:

1) Heterogeneity and Incompleteness: If we want to analyse the data, it should be structured but data is heterogenous i: e. data may be structured, semi structured and unstructured. So, Heterogeneity is the big challenge in data Analysis and analysts need to cope with it. Consider an example of patient in Hospital. We will make each record for each medical test. And we will also make a record for hospital stay. This will be different for all patients. This design is not well structured. So, we need to manage with the Heterogeneous and incomplete data. A good data analysis should be applied to this.

2) Scale: Big Data consists of large size of data sets and Managing with large data sets is a big problem. Earlier, this problem was solved by the faster processors but now data volumes are becoming very huge and processors are static.
World is moving towards the Cloud technology, due to this shift data is generated at a very high rate. This high rate of increasing data is becoming a challenging problem to the data analysts. Data is stored using the hard disks but they are slow I/O performance. But now Hard Disks are replaced by the solid-state drives and other technologies. These are not in slower rate like Hard disks, so new storage system should be designed.

3) Timeliness: Another challenge is speed. As discussed, earlier data sets are large in size, as a result for the analysis of data longer the time it will take. Any system which deals effectively with the size is likely to perform well in term of speed. Sometimes the analysis results are needed immediately. For example, if there is any fraud transaction, it should be analysed before the transaction is completed. So, some new system should be designed to meet this challenge in data analysis.

4) Privacy: Privacy of data is another problem with big data. For example, in social media we cannot get the private posts of users for sentiment analysis.

5) Human Collaborations: Today we have many computational models, but there are many patterns that a computer cannot detect. A new method of harnessing human ingenuity to solve problem is crowd-sourcing. Wikipedia is the best example. We are reliable on the information given by the strangers, however most of the time they are correct. But there can be other people with other motives as well as like providing false information. We need technological model to cope with this. As humans, we can look the review of book and find that some are positive and some are negative and come up with a decision to whether buy or not. We need systems to be that intelligent to decide.

B. Opportunities to Big Data:

Current era is of Data Revolution and Big Data is giving so many opportunities to business organizations to grow their business to higher profit level. Big data is playing an important role in technological fields as well as many other fields like health, economics, banking, and corporates as well as in government. Some examples are listed below:

1) **Technology**: Many organizations like Facebook, IBM, and yahoo have adopted Big Data and are investing on big data. Facebook handles. Every month Google handles 100 billion searches. From these stats we can say that there are many opportunities on internet, social media.

2) **Government**: Big data can be used to handle the problems faced by the government. Obama government announced big data research and development initiative in 2012. Big data analysis played an important role of BJP winning the elections in 2014 and Indian government is applying big data analysis in Indian electorate.

3) Healthcare: According to IBM Big data for Healthcare, 80% of medical data is unstructured. Healthcare organizations are adapting big data technology to get the complete information about a patient. To low down the cost and to improve the healthcare big data analysis is required and some technology should be adapted.

4) Science and Research: Big data is a latest topic of research. Many researchers are working on big data. NASA centre for climate simulation stores 32 petabytes of observations.

5) Media: Foreknowing the interest of the user on internet, Media is using big data for the promotions and selling of products. For example, social media posts, data analysts get the number of posts and then analyse the interest of user.

III. HADOOP FRAMEWORK

Hadoop is an open-source software framework for storing and processing the big data. It provides massive storage for

© 2019, IJCSE All Rights Reserved

any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
Hadoop is influenced by Google's architecture i: e. Google File System and MapReduce. Hadoop processes the large data sets in a distributed computing environment. An Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and other components like Apache Hive, Pig, HBase and Zookeeper. The Hadoop architecture can now be concluded and the place of HDFS, Map Reduce and YARN can be seen in the fig. There are two more tools in Hadoop ecosystem that are important are FLUME and SQOOP known as the data ingestion tools.

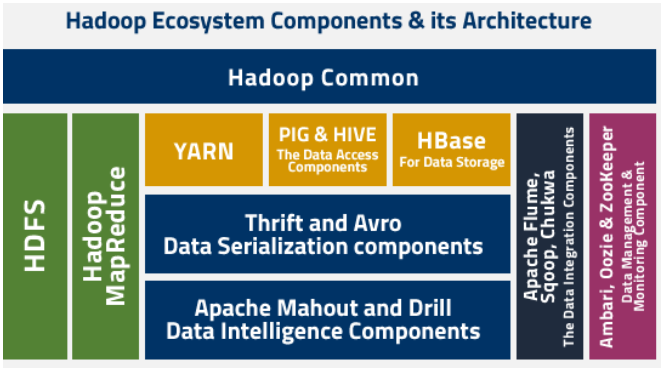

Figure 2: Hadoop Ecosystem

Hadoop has two main components: HDFS and MapReduce.

a) HDFS: - A fault- tolerant storage system is HDFS and it is able to store huge amount of information, scale up incrementally and survive the failure of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," (size of block is 64 megabytes) and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.
The architecture of HDFS is considered as master/slave relationship. It consists of a Single NameNode, a master server that manages the file system namespace and regulates access to files by clients. The data node is usually responsible for management of the storage attached to the nodes. Due to the replication of data on different data nodes Hadoop distributed file system is called as highly fault tolerant. The important feature of HDFS is that it is the storage system for the Map reduce jobs, both for input and output.

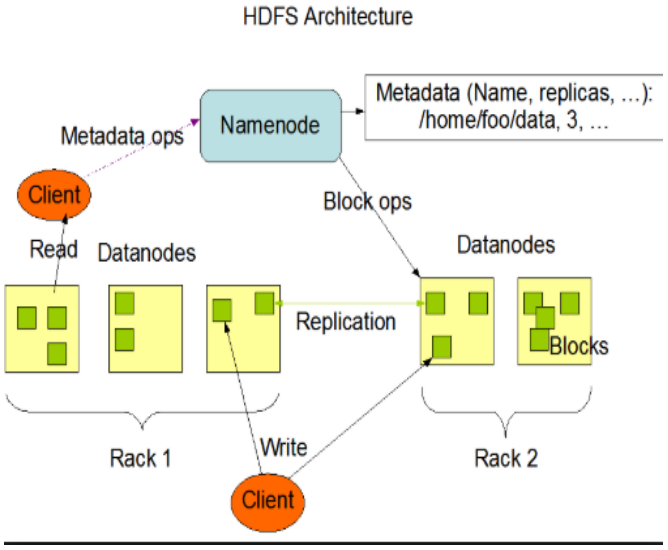

Figure 3: HDFS Architecture

There are three modes of Hadoop configuration:

1. Standalone mode: -in this mode, all Hadoop services runs in a single JVM on a single machine.

2. Pseudo-distributed mode: - in pseudo-distributed mode, each Hadoop runs on its own JVM, but on a single machine.

3. Fully Distributed mode: -in Hadoop distributed mode, Hadoop services runs on individual JVM, but these reside in separate commodity machine in single cluster.

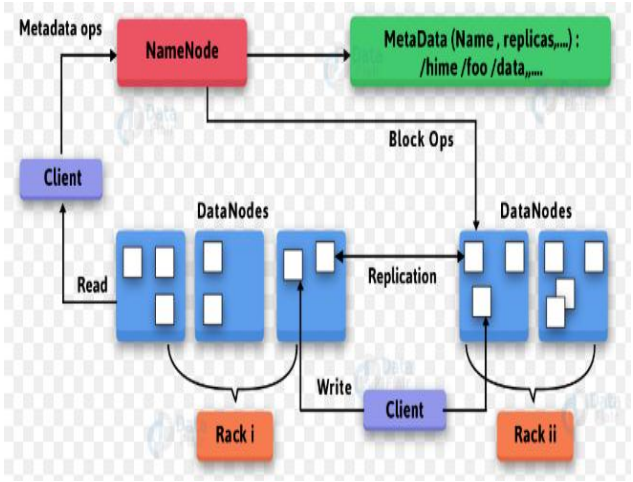

Figure 4: Representation of data nodes

Hadoop services- The main services of Hadoop involve the following: -

© 2019, IJCSE All Rights Reserved

Data node: -data nodes in Hadoop distributed file system (HDFS) are the slaves that are responsible for storing blocks of data.
Name node: -It is the master node that is responsible for the management of data blocks that resides in data node. It is centrally placed node, which contains information about Hadoop file system.

Secondary name node: -the secondary name node is a especially dedicated node in HDFS to take a checkpoint of the file system metadata that is present in name node. It keeps track of the data that it is alive.it cannot be considered as the replacement for name node but can be considered as the helper node. if the name node fails, the data can be recovered from secondary name node's logs.

b) MapReduce: Another component of Hadoop apart from HDFS is map reduce. It is the programming model and an implementation for processing and generating large data sets with parallel and distributed algorithms on a cluster. Map reduce has become a ubiquitous framework for large-scale data processing. It is an initial ingestion and transformation step, where individual input records can be processed in parallel. Reduce process is the aggregation or summation step, in which all associated records must be processed together in a group. Task tracker keeps track of individual map tasks, and can run parallel. A map reduce job runs on a particular task tracker slave node. Jobs of a map reducer:

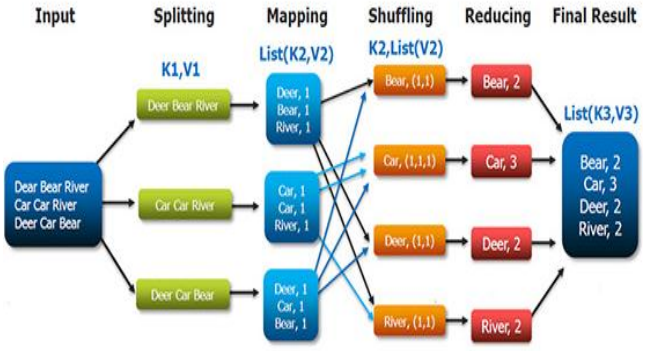

Figure 5: Map Reduce

The steps involved in map reducer job are: -

1. Input
2. Split
3. Map
4. Shuffle
5. Reduction

IV. COMPARISON OF BIG DATA WITH TRADITIONAL DATABASE

Table 2: Traditional Databases Vs. Hadoop

	Traditional Databases	Hadoop
Processing	Traditional RDBMS cannot be used to process and store large amount of data or big data. Traditional database supports OLTP.	Hadoop has two main components HDFS that is responsible for storage of big data and Map Reduce that is responsible for processing large data by splitting it into several blocks of data and then distributing these blocks across the nodes on different machines. It supports OLAP.
Throughput	Throughput means the total volume of data processed in a particular period of time so that the output is maximum. RDBMS fails to achieve a higher throughput as compared to the Apache Hadoop Framework.	This is one of the reasons behind the heavy usage of Hadoop than the traditional Relational Database Management System.
Data Variety (Type of data)	RDBMS can only be used for either structured (data in tabular form) or semi structured data (e.g., JSON data)	But because of the variety feature explained above in Hadoop, it can be used for either structured, semi structured or unstructured data .
Cost	RDBMS is a licensed software, you have to pay in order to buy the complete software license.	Hadoop is a free and open source software framework; you don't have to pay in order to buy the license of the software.

© 2019, IJCSE All Rights Reserved

V. CONCLUSION

Today we have entered in an era of data revolution where billion or trillions of data is called as Big Data. The paper describes the concept of Big Data based on concept of 4 Vs that stands for volume, velocity, variety and veracity of Big Data. The technology associated to deal with big data is Hadoop (composition of HDFS and Map Reduce). The challenges and opportunities of Hadoop in comparison with traditional database approach. The paper also describes Hadoop ecosystem which is open source software used for processing of Big Data. Hadoop is the need of today for processing, managing and dealing with big data.

REFERENCES

[1] International Journal of Engineering Technology, Management and Applied Sciences www.ijetmas.com March 2017, Volume 5 Issue 3, ISSN 2349-4476 19 AnjanaRaviprolu The Big Data and Market Research Anjana Raviprolu, Dr.Lankapalli Bullayya.

[2] International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153 www.ijsrp.org A Review Paper on Big Data and Hadoop Harshawardhan S. Bhosale1 , Prof. Devendra P. Gadekar.

[3] V.K. Gujare, P. Malviya, "*Big Data Clustering Using Data Mining Technique*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.9-13, 2017

[4] Shilpa Manjit Kaur," BIG Data and Methodology- A review" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[5] D. P. Acharjya, S. Dehuri and S. Sanyal Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.

[6] Jyoti Kumari, Mr. Surender, Statically Analysis on Big Data Using Hadoop,IJCSMC, Vol. 6, Issue. 6, June 2017, pg.259 – 265

[7] C.L. Philip Chen, Chun-Yang Zhang, "Data intensive applications, challenges, techniques and technologies: A survey on Big Data" Information Science 0020-0255 (2014), PP 341-347, elsevier.

[8] K. Parimala, G. Rajkumar, A. Ruba, S. Vijayalakshmi, "*Challenges and Opportunities with Big Data*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.5, pp.16-20, 2017

[9] Abdelladim Hadioui!!", Nour-eddine El Faddouli, Yassine Benjelloun Touimi, and Samir Bennani Machine Learning Based On Big Data Extraction of Massive Educational Knowledge,iJET – Vol. 12, No. 11, 2017.

[10] Mantripatjit Kaur, Anjum Mohd Aslam, "Big Data Analytics on IOT: Challenges, Open Research Issues and Tools", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.81-85, 2018

[11]]Kache, F., Kache, F., Seuring, S., Seuring, S.,Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. International Journal of Operations & Production Management 37, 10–36, (2017)https://doi.org/10.1108/IJOPM-02-2015-0078

[12] Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., Machine Learning on Big Data: Opportunities and Challenges. Neurocomputing, (2017).

© 2019, IJCSE All Rights Reserved