# *Big Data Analytics Using Hadoop*

Monoshi Kumar Roy and Farjana Parvin
Roll:1503067 and Roll:1503068
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology

***Abstract*—This paper is an effort to present the basic understanding of BIG DATA is and its usefulness to an organization from the performance perspective. Along-with the introduction of BIG DATA, the important parameters and attributes that make this emerging concept attractive to organizations has been highlighted. The paper also evaluates the difference in the challenges faced by a small organization as compared to a medium or large scale operation and therefore the differences in their approach and treatment of BIG DATA.**

## I. INTRODUCTION

We live in an age of data. Now a days,most people have acounts in different social networking sites such as Facebook,Intragram, Twitter etc. Facebook alone have 1.393 billion monthly active users. Among of them 890 million are daily active users and 350 million photos are uploaded and 4.75 billion items are shared in Facebook on a daily basis, on an average. Each day Facebook stored 600 petabyte of data takes in an amount of 600 terabyte of data. This data never gets deleted. Instead, it is increasing day by day in such a way that the rate of increase in the data itself gets increased. Such large amount of data are termed as 'Big Data' [1]

**What is Big Data:** "Big Data is the territory where out existing traditional relational database and file system processing capacities are exceeded in high transactional volume, velocity, responsiveness and quantity and on variety of data.The data are too big, move too fast, or don't fit the structures of RDBMS architectures." [4]

**The characteristics of Big Data:** There are three characteristics of Big Data , which specified by three 'V's which stands for volume , velocity and variety. All the challenges regarding Big Data mainly involve dealing with large volume, high velocity and different varieties of data.

i. Volume: When the terabytes amount of data are generated in the social media each day gets added to the existing petabytes of data and soon the data will be zettabytes. Such a large volume of data neither possible nor efficient to handle using traditional database, which made it a challenge.

. ii. Variety: Data are now being available additionally in the form of pictures, vedios, tweets etc. It implies that data that form Big Data which may vary in terms of type as well as source.

iii. Velocity: The use of online space are increasing. As a result the data that was available was rapidly changing and therefore had to be made available and used at the right time to be effective. It specifies the data that is in motion. Another dimension of velocity is by the lifetime of data utility. It specifies how long the data will be valuable. [3]

In today's world Big Data has great importance from healthcare to large scale analytics. For example: Aggregation of all related healthcare information from different sources helps tremendously in the treatment of a patient. The doctor can easily obtain information relating to a disease from different parts of the world and the integration of data from different areas such as administrative data, clinical data, cost involved data, pharmaceutical data, patient behavior and sentiment data, etc helps in efficiently treating a patient. [5]

Although the importance of Big Data is real and significant, there are some technical challenges that should be analyzed to completely realize its potential. These challenges mainly involve dealing with large volume, high velocity and different varieties of data. [6]

Hadoop is an important tool for storing and analyzing Big Data. For this Hadoop mainly consists of a distributed storage unit named Hadoop Distributed File System (HDFS) and a software framework called MapReduce. We can write different MapReduce programs for the analysis of Big Data and we can also edit the source code of Hadoop to enhance the performance of Hadoop. [1]

Our expected benefit from the experiment is to analytics of big data using hadoop. The experiment was performed on a hadoop cluster set up using the Amazon EC2 Servers. The Hadoop cluster was tested using word count and pi calculation problems and the results were obtained efficiently.

The rest of this paper is organized as follows: section 2 provides an complete explanation of the phases of Big Data analytics, section 3 exploits Hadoop, section 4 explores the implementation of Hadoop and section 5 concludes the paper.

## II. HADOOP

Now to store and analyze big data, data scientists have introduced hadoop. Now the question is what is Hadoop? Basically, Hadoop is an open source java framework provided by apache software foundation to store and analyze huge amount of data. [1]

Now, how does hadoop works? In our primary level, we solved some simple mathematics. That problem was like-If a man individually can finish a work in 40 days, then how many days it will take to complete the same task by 10 men? . Without calculating it, we can obviously say that it will take fewer days to complete the task when 10 men are working together. This basic concept is used in hadoop. In hadoop the datasets are

distributed among many nodes to provide ultimate uniformity but not absolute consistency. [7]
Now, one question may arise. Why hadoop is needed? Why not traditional database management system? Database management system works perfectly for the small amount of dataset. The main drawbacks of database management system are that a proper schema must be defined for every dataset and it is only suitable for structured dataset. But when a commercial company is dealing with terabytes of data, it is expected that for this huge amount of data it is very difficult to define a schema because as the size increases so the complexity and relationships among the dataset. Again, hadoop is fine with unstructured, structured and semi-structured datasets.
Let us consider an example, When result of any board examination(JSC,SSC,HSC) is published most of the students want to know their results through the website of education board. But, as thousand of students try to access the website at the same time, the server of the education board doesnt respond. But now compare this to the social networking sites like facebook, twitter or search engine giant Google. Millions of people across the world are browsing facebook at this moment and many people are searching their quest in Google. But their servers are working quite perfectly. This is the magic of big data analytics.
Hadoop mainly consists of two components. They are HDFS (Hadoop Distributed File System) and Mapreduce.

### A. HDFS

This contains the storage functionality of hadoop. Hadoop is a file system designed for storing very large files with streaming data access patterns running on clusters of commodity hardware. [8] In HDFS the concept of parallel computation is used. Suppose one machine is unable to perform a huge amount of task. Then employing many machines can be a solution of this problem. Similarly when we have to deal with a terabyte size data, it is not practical to store and process that data from a single machine. This amount of data should be distributed among many machines in a single cluster. This system is known as a distributed file system. The distributed file system of hadoop is known as HDFS. [8]
Now what challenges we may face in a distributed file system? It makes sense that if the number of machine increases, the probability of failure rate of any machine will increase too. So in that case, the fear of losing data of the faulty machine is obvious. Hadoop does simply mirror file data to various nodes. This process is known as Replication. Now suppose one or more machines is not working properly. This will cause storage server failure. But because of replication, if there exist at least replica in one server machine, the user will not know the failure of storage. [9]
HDFS provides another interesting feature which is different from other distributed file system. It gives a platform to compute data near the storage of data. [10] HDFS deals with gigabyte to petabyte size data. Now let, someone wants to perform some operations on this huge amount of dataset. It would not be logical if for the computation task to transfer petabyte size data to another platform and then perform operation. Just imagine the time cost to transfer of petabyte size data. So HDFS allow applications to move themselves near the data location. [10] Hadoop services consist of two processors. [9] They are-
(i)Namenode.
(ii) DataNode.

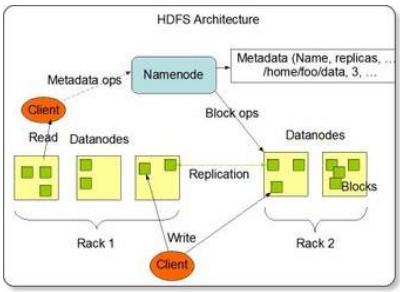


Fig. 1. HDFS Architecture

This is basically a master-slave structure. [10] Namenode is considered as master server and it is responsible for file system name space and operations like open, close and rename. Dataset is divided into many blocks by namenode and after partition into many blocks, this blocks being stored in the datanodes. This is why datanode is known as a slave and it serves the requisite data on the basis of clients request. Data node performs block creation, replication and deletion. [11]

### B. Mapreduce

HDFS is used to store dataset distributed into large clusters. Now, mapreduce is used to process those distributed data form thousand of nodes. Mapreduce consist of two different word. Map and Reduce, this two are phases of a mapreduce. The map function is basically used for tell the data points we want to retrieve from thousands of clusters. Reducer function will take those data from the cluster and aggregate them. [2] The architecture of mapreduce consists of two parts. This is also a master slave structure just like HDFS. The two parts are-
(i)Jobtracker and
(ii) Tasktracker.

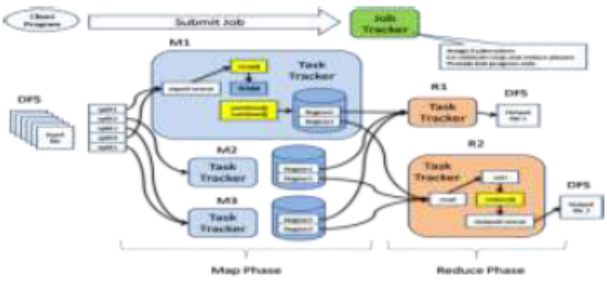


Fig. 2. MapReduce Architecture

Jobtracker is the master node and taskracker is the slave node. [11]

One may think that, a client must know java to retrieve data from cluster(as all programs are Java). But it is not so. Facebook has built a subproject named Hive which is a warehouse infrastructure. It is bit like SQL interpretation. Now SQL is easy to understand and for any general client it is easy to retrieve data using hive rather than java. [2], [9]

## REFERENCES


[1] Cyril, Nikhitha and Soman, Arun."Big Data Analysis using Hadoop," Citeseer, 2015

[2] Dhyani, Bijesh and Barthwal. Anurag,"Big data analytics using Hadoop," in International Journal of Computer Applications, vol. 108, no. 12, December 2014

[3] Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE BigData, pp. 403-410, October 2013.

[4] Raymond Gardiner Goss and Kousikan Veeramuthu,"Heading Towards Big Data Building a Better Data Warehouse for more data, more speed and more users," in IEEE 24th Annual SEMI Advanced Semiconductor Manufacturing Conference, 2013, pp. 220-225.

[5] Sonja Zillner et al, "Towards a Technology Roadmap Big Data Applications in the Healthcare Domain," in IEEE 15th International Conference on Information Reuse and Integration, California, 2014, pp. 291-296.

[6] Divyakant Agrawal et al, "Challenges and Opportunities with Big Data," Cyber Center Technical Reports, Purdue e-Pubs, Purdue University, 2011.

[7] Jagtap, Dinesh D and Patil, BK, "Big Data using Hadoop," in International Journal of Engineering Research and General Science,vol. 2, no. 6, 2014

[8] White and Tom,*Hadoop-The Definitive Guide: Storage and Analysis at Internet Scale (revised and updated)*, 3rd ed. Sebastopol:O'Reilly, 2012,pp. 43-44

[9] Venner and Jason, "Pro Hadoop-Build Scalable, Distributed Applications in the Cloud," in Estados Unidos, 2009, pp. 4-6

[10] Borthakur, Dhruba and others,"HDFS architecture guide," in Hadoop Apache Project,vol. 53,2008

[11] Prajapati and Vignesh,*Big data analytics with R and Hadoop*, UK:Packt Publishing Ltd, 2013,pp. 30-31