International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

# Survey Paper on "MAP REDUCE PROCESSING USING HADOOP".

**Manisha R Kaila[1], Harshita N Shetty[2], Meet S Bhanushali[3]**

[1]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[2]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*
[3]*Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India*

---

**Abstract -** *With the development of web based applications and mobile computer technology there is a rapid growth of data, their computations and analysis continuously in the recent years. Various fields around the globe are facing a big problem with this large scale data which highly supports in decision making. Even the traditional relational DBMS's and classical data mining are not suitable for handling this big data. Map Reduce is used my many popular and well-known IT companies such as Google, Yahoo, Facebook etc. In Big data world Map Reduce helps in achieving the increasing demands on computing resources affected by voluminous data sets. Hadoop is an approved open-source map-reduce implementation which is being used to store and process extremely large data sets on commodity hardware.*

***Key Words*: *Big Data, Hadoop, MapReduce, Bigdata Analytics***

## 1. INTRODUCTION

Ever since the development of technology, data has also been growing every single day. If you go back in time in 70s or 80s not many people were using computers there were only a fraction of people who were dealing with computers and that's why the data fed into the computer system was also quite less but today everyone owns a gadget like laptops, mobile phones and they are generating data from there every single day. Another factor behind the rise of big data is social media. Billions of people today use social media for posting photos, videos and to interact with others. It has been seen that in Facebook the user generates almost 4 million likes in every 60 seconds and similarly on Twitter there is almost 300 thousand tweets every minute. Now that's a lot of data and it has been rising exponentially over years. Big Data is the term for collection of data sets so large and complex that is becomes difficult to process using on hand database management tools or traditional data processing applications. It becomes very difficult to deal with this kind of data that is being generated with such a high speed and that's why big data becomes an issue because traditional system fails to store and process such big data. The big data problems are identified by 5 V's – Volume, Variety, Velocity, Value, and Veracity.

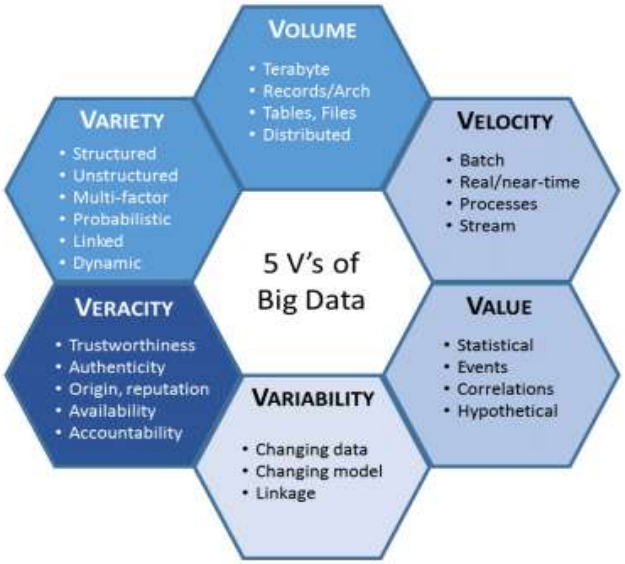
Volume- It implies that the amount of data the client is using is so huge that it becomes increasingly difficult for the client to store the data into the traditional system.

Variety- As there are many sources which lead to Big Data, the type of data they are generating is different. The data generated can be structured, semi-structured or unstructured. Almost 90% of the data are unstructured which is a problem because traditional systems are incapable of processing this unstructured data.

Velocity- Velocity refers to the speed at which data is being generated and processed to meet the demands. 72 hours of YouTube video is uploaded every minute, this is velocity

Value – It is very important to have useful data and you should be able to extract the right information from it because there are many datasets lying around which are unnecessary.

Veracity- Veracity is sparseness of data. Veracity says that you cannot expect data to be always correct or reliable in today's world. There might be data which has missing values and we may have to work with various types of data that is incorrect.




## 2. APACHE HADOOP

Apache Hadoop is an open-source framework that allow us to store and process large data sets in parallel and distributed fashion. It can handle hardware failure automatically. Several applications like Google, Facebook, Yahoo, YouTube, Amazon etc. uses Hadoop. Hadoop Framework can be divided into two parts- Hadoop Distributed File System (HDFS) which helps to store data across clusters and Map Reduce which deals with processing of data stored in HDFS.HDFS is a distributed file system that

allows you to store large data across the clusters. In Apache Hadoop splits the file into set of blocks and each block length is 64 megabytes or 128 megabytes, these blocks are circulated across multiple nodes of cluster. HDFS follows Master/Slave architecture. It means that you have one master node and remaining all other nodes are slave nodes. In HDFS, Master Node are called Name nodes whereas Slave Nodes are called Data Nodes. Now Data Nodes are actually responsible for storing the actual data and Name Node being the master node manages all the slave nodes or data nodes. Other responsibility of Name Node includes maintaining and managing metadata. Metadata is the information about data that is present inside the data nodes. Along with that Data Node are supposed to send some signal so as to ensure that all the Data Nodes are working properly. If anyone Data Node stops sending signal, then name node will assume that the particular Data Node has been failed and same will be notified to admin. It distributes the file among the nodes and allows the system to continue work in case of any node failure. This approach reduces the risk of catastrophic system failure. Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce, and related projects are zookeeper, HBase, Apache Hive. Hadoop is scalable, fault tolerant and has high availability. Storage and analysis for large scale processing is provided by Hadoop components. Now a day's Hadoop is being used by hundreds of companies.

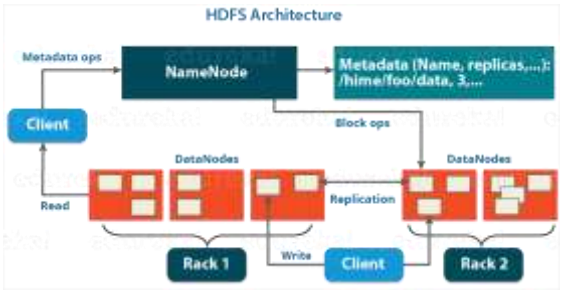



**Fig (a) HDFS**

## 3. MAPREDUCE

Map Reduce is a processing unit of Hadoop using which we can process the Big Data that is stored in Hadoop Distributed File System (HDFS). The Big Data that is stored on HDFS is not stored in a traditional manner. The data gets divided into blocks of data which is stored in respective Data Nodes. There is no complete data that is present in one single location. Hence native client application cannot process that data so we need a special framework that has the capability of processing such data and so Hadoop Map Reduce is used. Map Reduce is used for Indexing and Searching, Classification and it is used as recommendation engine by Amazon, Flipkart and many more. It is also used for Analytics by several companies. Map Reduce has been implemented by Google and it has been adopted by Apache Hadoop for processing data using Pig, Hive and HBase. The Map reduce programming model is based on two functions which are map function and reduce function. It is also called as master and slave architecture. Job tracker is the master process in MapReduce whereas task tracker is the slave in Map Reduce. These are the processes which are used to process the data present in HDFS. Job tracker is running under name node machine whereas task tracker run on data nodes. sub-project of the Apache Hadoop project. The reduce operation combines those data tuples based on the key and accordingly modifies the value of the key. A Map Reduce framework can be classified into two steps such as: Map Task and Reduce Task. The Map Task gets an input and the output yield from map task after processing is given as input to the reduce task and then the final output is given to the client. A Map Reduce works on key value pair. A Map takes key value pair as input and gives an output as list of key value. Now this list of key value goes through a shuffle phase and an input of key and list of values is given to the reducer. Finally, the reducers give a list of key value pair as output.

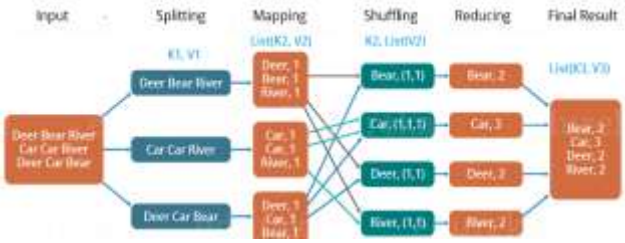



**Fig (b) Map Reduce**

### 3.1 Input reader

The input reader divides the input into applicable size 'splits' (in follow, typically, 64 MB to 128 MB) also the framework assigns one split to every Map perform. The input reader reads information from stable storage (typically, a distributed file system) and generates key/value pairs. A common example can scan a directory filled with text files and return each line as a record.

### 3.2. Map function

The Map function takes a series of key/value pairs, processes each, and generates zero or more output key/value pairs. The input and output forms of the map usually can be (and often are) completely different from one other. If the appliance is doing a word count, the map function would break the line into words and output a key/value combine for every word. Each output combine would contain the word as the key and also the variety of instances of that word within the line as the value.

### 3.3. Partition function

Each Map function output is allocated to a particular reducer the application partition for sharing purpose. The partition perform is given the key, and also the variety of reducers and returns the index of the required reducer. A typical default is to hash the key and use the hash price modulo the number of reducers. It is necessary to choose a partition perform that

provides AN around uniform distribution of knowledge per fragment for load-balancing functions, otherwise the Map Reduce operation are often delayed waiting for slow reducers to finish (i.e. The reducers assigned the larger shares of the non-uniformly divided data). Between the maps and cut back stages, the data are shuffled (parallel-sorted / exchanged between nodes) to move the data from the map node that produced them to the shard in which they will be reduced. The shuffle will generally take longer than the computation time counting on network information measure, CPU speeds, data produced, and time taken by map and reduce computations.

### 3.4. Comparison function

The input for every cut back is forced from the machine where the Map ran and sorted using the application's comparison perform.

### 3.5. Reduce function

The framework calls the application's Reduce function once for each unique key in the sorted order. The cut back will retell through the values that are related to that key and manufacture zero, or additional outputs. In the word count example, the reduce perform takes the input values, adds them and generates one output of the word and also the final sum.

### 3.6. Output writer

The output writer writes the output of the Reduce to the stable storage.

## IV. ADVANTAGES

### 4.1 Scalability:

Hadoop is highly scalable. This is because of its ability to store as well as distribute large data sets across plenty of servers. It provides a model in which data can be processed by spreading the data over multiple inexpensive machines which can be increased or decreased as per the enterprise requirement

### 4.2 Cost-effective solution:

Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are staggering: rather than cost accounting thousands to tens of thousands of pounds per computer memory unit, Hadoop offers computing and storage capabilities for many pounds per computer memory unit.

### 4.3 Flexibility:

Hadoop allows businesses to simply access new knowledge sources and faucet into differing of information (both structured and unstructured) to come up with worth from that data. This means businesses will use Hadoop to derive valuable business insight from knowledge sources like social media, email conversations or click stream knowledge .In addition, Hadoop may be used for a large type of functions, like log process, recommendation systems, knowledge deposit, market campaign analysis and fraud detection.

### 4.4 Parallel processing:

One of the first aspects of the operating of Map Reduce programming is that it divides tasks in a very manner that enables their execution in parallel. Parallel processing permits multiple processors to take on these divided tasks, such that they run entire programs in less time.

## V. CONCLUSION

This paper analyses the concept of big data analysis and how it can be simplified as from existing traditional relational database technologies. This paper clearly specifies the Hadoop environment, its architecture and how it can be implemented using Map Reduce along with various functions. Hadoop map reduce programming overview and HDFS are progressively being utilized for processing vast and unstructured data sets.

## VI. REFERENCES

[1] http://hadoop.apache.org.2010

[2] https://en.wikipedia.org/wiki/MapReduce

[3] T. White, Pro Hadoop: The Definitive Guide. O'Reilly Media, Yahoo!Press, June 5,2009.

[4]http://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm

[5]V. Patil, V.B.Nikam, "study of mining algorithm in cloud computing using MapReduce Framework", Journal of engineering, computers & applied sciences (JEC&AS) vol.2,No.7,July 2013

## BIOGRAPHIES

Manisha R Kaila is pursuing Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.

Harshita N Shetty is pursuing Bachelor's degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.

Meet S Bhanushali is pursuing Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering, Navi Mumbai.