IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

NCAIT 2017

JSS Academy of Technical Education
Vol. 4, Special Issue 8, May 2017

# Hadoop: A Frame Work for Big Data

**Aditya Priyadarshi[1], Bharath V[2], Malini M Patil[3]**

Dept of ISE, JSSATE, Bangalore[1,2]

Associate Professor, Dept of ISE, JSSATE, Bengaluru[3]

**Abstract:** Due to advent of new technologies, the amount of data produced by mankind has already reached a zetta-byte level. New devices, businesses and social networking sites are a major source for the production of such large amount of data. This data is really huge and collectively called by a well known term "Big data". Due to such huge amount of data being available it becomes very difficult to perform effective analysis using the currently available traditional techniques. From the literature survey it is found that there are a total of 39 tools available for analysis and processing of big data. Survey reveals that but the most influential and established tool for analyzing big data is Apache Hadoop which is an open source framework written in java that allows parallel processing across clusters of computers using basic programming techniques. This paper introduces apache Hadoop, its framework, installation and how it uses map reduce and cluster programming to capture, analyse and process big data."

**Keywords:** big data, Hadoop, mapreduce, HDFS, Clustered.

## I. INTRODUCTION

Grace Hopper stated that during the earlier time's people used oxen for heavy loads, but whenone ox couldn't pull the load, they never tried to growthe ox size instead they increased their numbers. The same logic we're applying in modern days. Today we dont't try for bigger computer's in case of heavy tasks, instead we go for more number of systems.

In the present date, Everybody would have heard about the much hyped term Big Data. But what this "Big data"? Big data is term that describes large amount of structured, semistructured and unstructured data that can be used for obtaining information. Big data is well described by it's 3 dimension called it's 3Vs: firstly, the extreme volume of data, secondly the wide variety of data and at last the velocity at which the data processes. Although this data is increasing day by day, but the term often used to specify it is zetta-bytesof data captured over time.

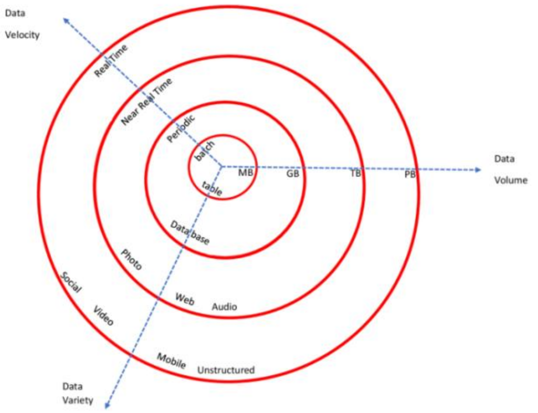

Fig. 1 : Three dimensions of big data

**Where does this data comes from?**
Originally, the source of all this various kinds of huge data was the web, mostly known as internet. But as times changed, data started to get generated from various sources. Although the original source of this data, the web still is considered to be the most data generating source, there are also sources like, Social media data which is all about the data generated daily by sites like facebook, twitter, and all which is considered to be large and increasing on a daily basis. Click stream data, which is the data created and stored when user navigate through a website and their clicks are used for further analysis which can be used for E-commerce and other purposes of advertisement.

Stock exchange data which has the information about the decisions made by the clients for buying and selling of stocks of specific companies.Sensors for temperature, noise and even more also generate huge amount of data. Still there are many more sources which can be taken into consideration for generating such massive amount of data.

**Why do we need Big Data Analytics?**
Big data analytics is needed by organizations as using this they can harness their data in a better way and use it to create new opportunities. Thus overall it leads to smarter business moves, higher profits for organisations, more efficient operations which leads to happy customers who are completely satisfied.

1. **Cost reduction.** Big data tools such as Apache Hadoopcan bring significant cost reduction when it is needed to store such large amount of data and the overall work can be carried out in an efficient way.
2. **Faster, better decision making.** Using the speed at whichHadoopimplements and it's in memory analytics,

organisations are able to analyze information instantly which in turn helps in making better decisions on the basis of what information they have already stored.

3. **New products and services.** Using Big Data analytics organisations can give the customers what they want by understanding what all they need and expect which fulfills their satisfaction.

## II. BIG DATA TOOLS

Due to a large variety of data that Big data comprises of, it also brings a large number of new challenges related to it's volume and complexity. From a recent survey it can be noted that, 80% of the data created in the world is unstructured. Our first challenge comes as how to convert this unstructured data into structured, before we can understand and capture this important data. Then comes a new challenge as to how to store this data.From the literature survey it has been found that there are a total of 39 tools available for analysis and processing of big data. Here are some of the top tools that are used to store and analyse Big Data.

**1. Apache Hadoop:**
According to surveythe most influential and established tool for analyzation of big data is Apache Hadoop. It's a java based open source software framework which runs applications using the MapReduce algorithm and can store large amount of data in a cluster and parallely process them using basic programming techniques. Hadoop runs in parallel on a cluster of computers with an ability to allow the user to process data parallely across many nodes. Hadoop Distributed File System (HDFS) is a storage system which Hadoop uses to split data and distribute it to different nodes. This method also has a plus point that it replicates the data providing high availability.

**2.Cloudera:**Cloudera is a company which makes commercial version of Hadoop as free version is not easy to use even if it is open source. Thus companies like Cloudera have developed a friendlier version of Hadoop.

**3.MongoDB:** It is a good tool to handle data which is frequently changing orthat is unstructured. Oftenly, it is used to store data in mobile apps, real-time personalization, content management.

**4.Hive:**Hive facilitates in managing large datasets present in distributed storage. This language allows map reduce programmers to plug in their mappers and reducers when it is inconvenient to express it via HiveQL.

**5.Spark:** It is an open source data analytic cluster computing framework. When compared to Hadoop, Spark is considered to be about 100 times faster.

**6.Tableau:**It is a data visualization tool that mainly focuses on business intelligence. With Tableau, we can create bar charts , maps, scatter plots, and even more.

## III. HADOOP ARCHITECTURE

Hadoop framework has the following four modules:

**1.Hadoop Common:** Hadoop common consists of Java libraries and utilities which is required by other Hadoop modules that provides filesystem and OS level abstraction's and consist of Java files and scripts required byHadoop to start itself.

**2.Hadoop Yarn:** Hadoop Yarn is used for job scheduling and cluster management.

**3.Hadoop Distributed File System:** This is a distributed file system that is used for providing access to application data.

**4.Hadoop MapReduce:** This module is based on Yarn and is used parallel processing of huge data sets.

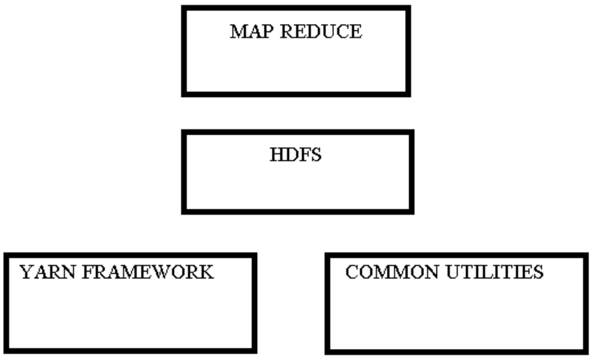

Fig. 2 Hadoop architecture

## IV. HADOOP INSTALLATION

Hadoop can be installed in three different modes of operation:

**1.Stand Alone Mode**: Hadoop is a open source framework which has been designed to run on commodity machines. But, it can also be installed on a single node using stand-alone mode. Under this mode, it runs as a single java process and is used for generally debugging purposes. In addition to this, we can even test run our Map-Reduce application on small data, before we execute it on a cluster with big data.

**2.Pseudo Distributed Mode**: Hadoop software can also be installed under this mode on a single node. Several daemons of Hadoop such as NameNode, DataNode, JobTracker will be running on the same machine as separate java processes

**3.Fully Distributed Mode**: In this mode, the daemons namely NameNode, JobTracker, SecondaryNameNode run on a Master Node. Other daemons like DataNode and TaskTracker run on the Slave Node.

## V. ADVANTAGES OF HADOOP

1.Using Hadoop framework user can quickly write and test distributed systems. It is very efficient as it does automatic distribution of taskacross number of machines and utilizes parallelism of the CPU cores.

2. At application layer, Hadoop can detect and handle failures.

3. InHadoop framework servers can be dynamically added or removedfrom the cluster without interpreting any of the Hadoop's operation.
4. Hadoop is compatible on a number of platforms because it is java based and open source.

## VI. MAP REDUCE

Map Reduce is the most essential programming framework of Big data .Using Map Reduce framework applications can be written to process huge amounts of data parallely on large clusters of commodity hardware in an efficient manner.

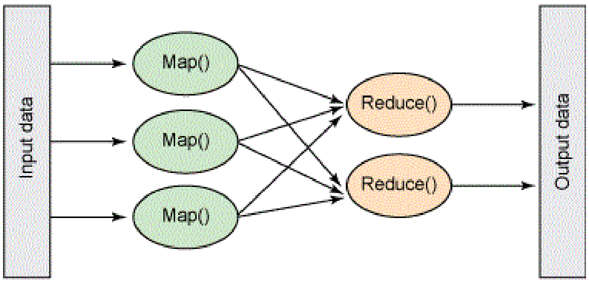

Fig. 3 Map reduce

MapReduce is a programming and processing model for parallel computing based on java. It's algorithm mainly consists of two tasks, that is Map and Reduce.

Map is used for conversion of a set of data into another set of data, in which individual elements arebroken down into tuples. The second algorithm is reduce, which gets the output from a map as its input and breaks the data tuples into a even more smaller set of tuples. As per the naming reduce task is followed after the map. The advantage of Map Reduce is that it becomes easy to scale data processing over multiple nodes. The data processing primitives under this model are called mappers and reducers. As we write an application in the Map Reduce form, making the application to run on any number of machines in a cluster is just a configuration change. This simple configuration change drew the attention of everyone to use this model as this way was more time efficient than the traditional model.

### The Algorithm

MapReduce algorithm is based on specifying the computer current location of data. MapReduce has three stages, that is map stage, shuffle stage, and reduce stage.

**Map stage** : This stage is to process the input data, which is either in the form of a file or a directory and is stored in the HDFS. The mapper processes this data from the input file line by line and creates small sets of data.

**Reduce stage** : It combines of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process whatever it gets from the mapper. After processing of this data, a new output is created which is stored in the HDFS. This framework manages all the details of data passing such as verifying task completion, and copying of data around the different clusters.

1. Most of the computing tasks takes place on nodes where data is on the local disks which in turn reduces network traffic.
2. After the completion of given tasks, the cluster collects and reduces this data to form an appropriate result, and at last sends it back to the Hadoop server.

## VII.HDFS

HDFS stands for Hadoop Distributed File System as it was developed using distributed file system design. It is also a commodity hardware. It is designed using low-cost hardware. It can hold huge amount of data and thus provides easier access. This huge amount of data is stored across multiple machines. These files are stored in such a way that possible data losses in case of failure can't take place from the system. HDFS also helps in parallel processing.

### HDFS Architecture

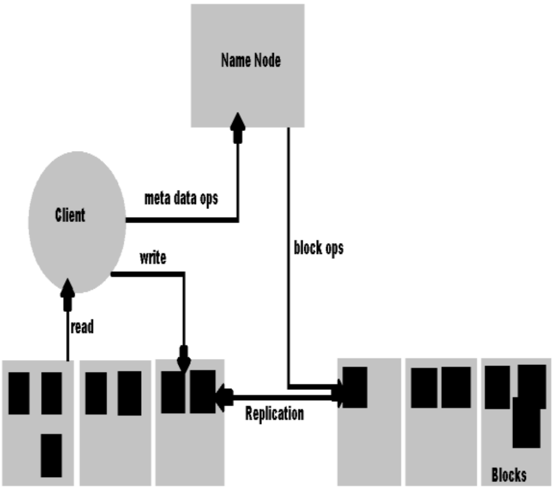

Fig.4 HDFS Architecture

HDFS has the master-slave architecture with the following elements.

**Namenode**

It is a software that can run on commodity hardware. The system having the Namenode is the master server and performs the following tasks:
1.Manages file system namespace.
2.Provides access to files for clients.
3.Performs basic operations like renaming, opening, closing of files and directories.

**Datanode**

There will be a Datanode for every node in a cluster. Datanode manages the data storage of the system.
1.They perform read and write operations on the file.

2.They perform operations like block creation, replication and deletion.

**Block**
Data mostly is stored in the files of HDFS for the user. Files are be divided into segments and stored in individual data nodes under this file segments. These file segments are known as blocks. The default size of these blocks are 64MB, but it can be changed as per the change in HDFS configuration.

**Goals of HDFS**
**1.Fault detection and Recovery** :
A large number of commodity hardware are included in HDFS, therefore failures will be more frequent for components. For such cases, HDFS should have mechanisms for automatic and quick fault detection and recovery.
**2.Huge datasets** : HDFS should have a large number of nodes per cluster to manage applications having huge datasets.
**3.Hardware at data** : A task which is requested can be done efficiently, only if the computation takes place near the data. Especially when huge datasets are involved, it reduces the network traffic and increases the processing speed.

## VIII. HADOOP CHALLENGES

**1.Mixed workload environments cause jobs to fight for resources.**
While Hadoop scheduling has been improved over the past few years, they still pre-allocate resources when a job starts. The problem here is that each job requires different hardware resources during the span of their lifetime. Moreover, some hardware resources aren't limited in Hadoop. Thus these all factors lead to competition for resources.

**2.Troubleshooting is difficult and cantake hours.**
In case of Hadoop there are a large number of tools that allows to monitor the clusters, administrators are in most cases left unknown with the factors that affect the cluster health. It isdifficult to identify the root cause of these problems as because of lack of granular tools which causes a lot of inefficient behaviour, such as restarting and asking users about jobs they submitted. As with the increase in cluster size businesses heavily rely on Hadoop, therefore such methods will be unsustainable.

**3.Buying more hardware than needed.**
Organizations usually are unknown about the amount of hardware needed, so they size their clusters on the basis of preassumed peak loads. Their goal is such as to make sure that jobs don't overload the cluster and in turn cause degraded performance and job failures. In Hadoops up-front allocation of resources we can't know the amount of hardware required, so this technique becomes expensive and leaves capacity unused most of the time, in addition to it, still fails to prevent workloads as they are often unpredictable.

## IX CONCLUSION

The present technological advancement have contributed for the massive data collections. Few sources of these data sets are from weather forecasting data, sensor data, internet web click streams, ecommerce data. Traditional software's are unable to handle such a massive collection of data. An effort is made in this paper to have an insight on the tools & packages which enable the user to handle the massive data set. Thus user can also create their own programming environment and can handle Big Data easily.

## REFERENCES

[1] https://www.tutorialspoint.com/Hadoop/Hadoop_hdfs_overview.html
[2] http://www.edupristine.com/blog/Hadoop-installation
[3] https://www.sas.com/en_us/insights/analytics/big-data-analytics.html
[4] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
[5] http://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/
[6] https://www.ibm.com/developerworks/library/wa-introhdfs/
[7] http://bigdata-madesimple.com/top-6-big-data-tools-to-master-in-2017/
[8] OReilly Hadoop The Definitive Guide 3rd Edition May 2012
[9] Hadoop Illuminatedby Mark Kerzner and Sujee Maniyam