International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 5 131- 135

A Review Paper on Big data & Hadoop

Rupali Jagadale

MCA Department, Modern College of Engg.
Modern College of Engginering
Pune,India
rupalijagadale02@gmail.com

Pratibha Adkar

MCA Department, Modern College of Engg.
Modern College of Engginering
Pune,India
pratibhakul@gmail.com

Abstract--Big data is dataset that having the ability to capture, manage & process the data in elapsed time .Managing the data is the big issue. And now days the huge amount of data is produced in the origination so the big data concept is in picture. It is data set that can manage and process the data. For managing the data the big data there are many technique are used .One of this technique is Hadoop. Hadoop can handle the huge amount of data, it is very cost effective, and it can handle huge amount of data so processing speed is very fast, and also it can create a duplicate copy of data in case of system failure or to prevent the loss of data.This paper contains the Introduction of big data and Hadoop, characteristics of big data ,problem associated with big data, architecture of big data and Hadoop, other component of hadoop, advantages, disadvantages and applications of Hadoop and also the conclusion.

Keywords: *Bigdata, Hadoop, Mapreduce, Hive, pig, HBAS*

*****

I) INTRODUCTION:

Big data definition:

Big data is dataset that having the ability to capture, manage& process the data in elapsed time. Big data includes the unstructured data, semi structured data, & structured data but it mainly focus on unstructured data. Big data size is vary from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte).

Characteristics of big data:

Big data is having mainly 3 V's of characteristics.

1) Volume:

The volume contains the amount of data generated in the enterprise.The size ofdata defines the whether the data is considered as the big data.

2) variety:

The variety of data defines the type of data, source of data & nature of data. This helps the people who analyze the data in to structured, unstructured & semi structured data.

3) Velocity:

Velocity of data means the speed at which data is generated and processed to the demand.in other words its means the speed of data

Big data Architecture:

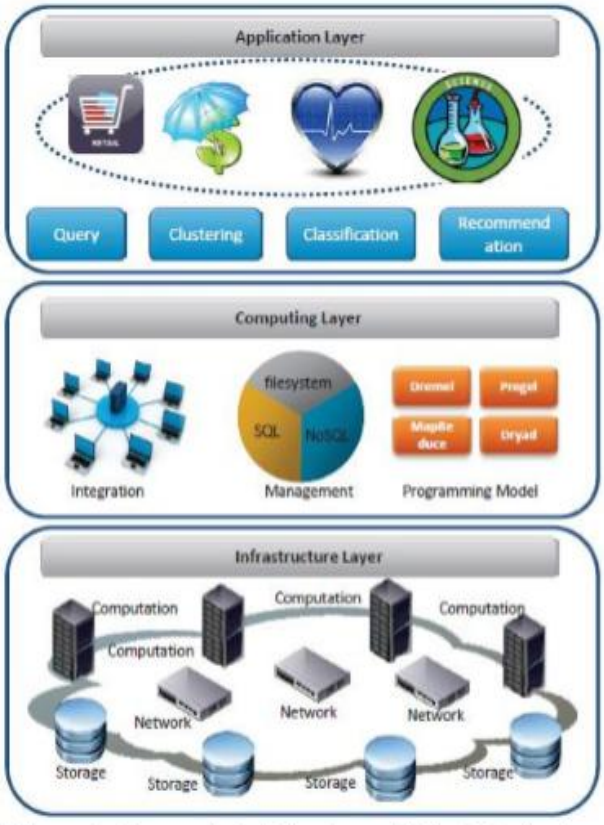

Figure 1: Layered Architecture of Big Data System

Problem associated with big data processing:

1) Information growth:
In big data it's the most important issue that is size.Of course we heard the word big data the first thing we are

having in our mind is size. Managing large and rapidly increasing amount of data is the challenging task. Volume of data is increasing faster and the CPU speed is static.

2) Speed:

Size matters the speed. If there is the larger the dataset having the large information the then it will take more time to response.

3) Privacy And Security:

Privacy of data is the huge issue arises in big data. In US there is great fear regarding the inappropriate use of the personal data.

Hadoop-solution to the big data:

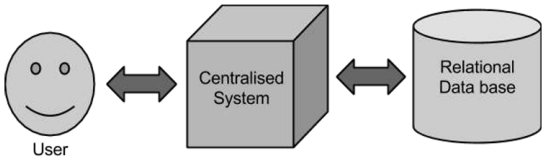

Figure 2.Actual working of hadoop

The above approach is worked fine with that application which processing the small amount of data. But it is not suitable for the application which is having the large amount of data to process.

To solve this problem the Google invented the technique of Hadoop. In this technique the hadoop uses the mapReduce algorithm, in that it divides the task into the small parts and assigns it to separate computer and collect the result in the dataset. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

Hadoop Architecture:

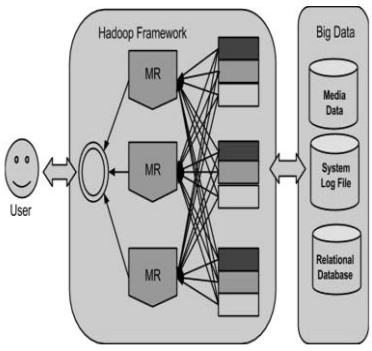

Figure 3.Hadoop Architecture

Hadoop Architecture:

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts

Hadoop has two major:

(a) Storage layer (Hadoop Distributed File System).
(b) (a) Processing layer (MapReduce)

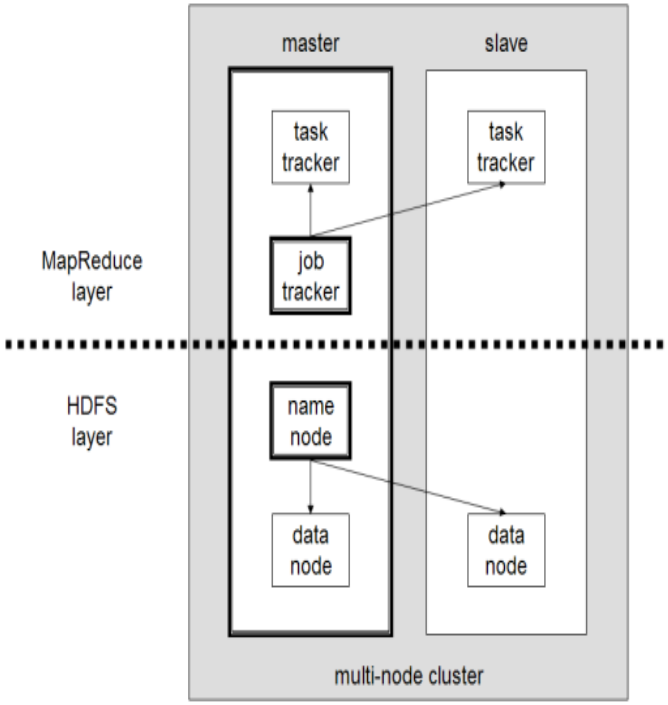

Figure 4 .Hadoop Architecture

a) HDFS(Hadoop Distributed File System):

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. HDFS holds very large amount of data and provides easier access.

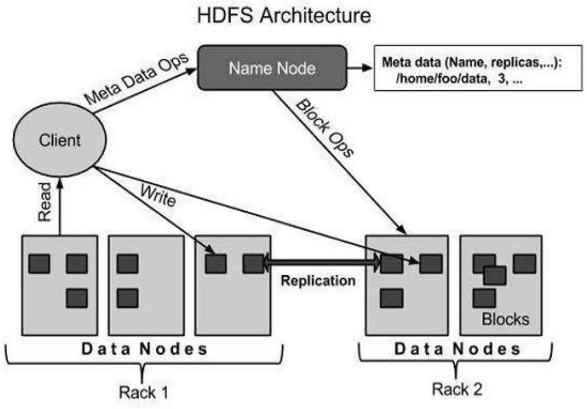

Figure 5. HDFC Architecture

b) MapReduce Architecture:

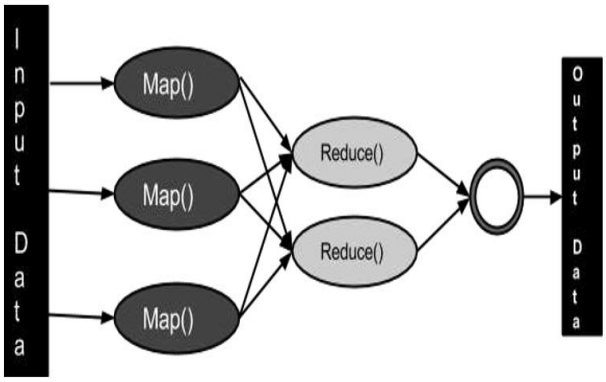

Figure 6. MapReduce Architecture

MapReduce framework is the processing pillar of hadoop. The framework is applied on the huge amount of data divided in part and run parallel. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

1) Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

2) Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Other components of Hadoop:

Hive

Hive is data warehousing software that addresses how data is structured and queried in distributed Hadoop. Hive is also a popular development environment that is used to write queries for data in the Hadoop environment. Hive is a declarative language that is used to develop applications for the Hadoop environment; however it does not support real-time queries. Hive is a technology developed at Facebook that turns Hadoop into a data warehouse complete with a dialect of SQL for querying. Being a SQL dialect, Hive is a declarative language. you specify the data flow, but in Hive we describe the result we want and Hive figures out how to build a data flow to achieve that result. Unlike Pig, in Hive a schema is required, but you are not limited to only one schema. [2]Like PigLatin and the SQL, Hiveitself is a relationally complete language but it is not a Turing complete language. It can also be extended through UDFs just like Piglatin to be a Turing complete. Hive is a technology for turning the Hadoop into a data warehouse, complete with SQL dialect for querying it. Hive works in terms of tables. There are two kinds of tables you can create: managed tables whose data is managed by Hive and external tables whose data is managed outside of Hive

Pig

Pig is a procedural language for developing parallel processing applications for large data sets in the Hadoop environment. Pig is an alternative to MapReduce, and automatically generates MapReduce functions. Pig includes Pig Latin, which is a scripting language. Pig translates Pig Latin scripts into MapReduce. Pig consists of a language and an execution environment. Pig's language, called as PigLatin, is a data flow language - this is the kind of language in which you program by connecting things together. Pig can operate on complex data structures, even those that can have levels of nesting. Unlike SQL, Pig does not require that the data must have a schema, so it is well suited to process the unstructured data. But, Pig can still leverage the value of a schema if you want to supply one. PigLatin is relationally complete like SQL, which means it is at least as powerful as a relational algebra. Turing completeness requires conditional constructs, an infinite memory model, and looping constructs. PigLatin is not Turing complete on itself, but it can be Turing complete when extended with User--Defined Function

HBase

HBase is a scalable, distributed; NoSQL database .It was designed to store structured data in tables that could have many of rows and many of columns. HBase is not a relational database and wasn't designed to support transactional and other real-time applications. Apache HBase is distributed column based database like layer built on Hadoop designed to support billions of messages per day, HBase is massively scalable and delivers fast random writes as well as random and streaming reads. It also provides row-level atomicity guarantees, but no native cross-row

transactional support. From a data model perspective, column-orientation gives extreme flexibility in storing data and wide rows allow the creation of billions of indexed values within a single table. HBase is ideal for workloads that are write-intensive, need to maintain a large amount of data, large indices, and maintain the flexibility to scale out quickly.

Advantages and disadvantages of Hadoop:

Advantages:

1) Range of data sources:

The data collected from various sources will be of structured or unstructured form. The sources can be social media or even email conversations. A lot of time would need to be allotted in order to convert all the collected data into a single format. Hadoop saves this time as it can derive valuable data from any form of data. It also has a variety of functions such as data warehousing, fraud detection etc.

2) Cost effective:

The companies had to spend lots of amount of their benefits into storing large amounts of data. In some cases they had to delete large sets of raw data in order to make space for new data. There was a possibility of losing valuable information in such cases. By using Hadoop, this problem was completely solved. It is a cost-effective solution for data storage purposes.

3) Speed:

Every organization uses a platform to get the work done at a faster rate. Hadoop enables the company to do just that with its data storage needs. It uses a storage system wherein the data is stored on a distributed file system.

4) Multiple copies:

Hadoop automatically duplicates the data that is stored in it and creates multiple copies. This is done to ensure that in case there is a failure, data is not lost. Hadoop understands that the data stored by the company is important and should not be lost unless the company discards it.

Disadvantages:

1) Lack of preventive measures:

When handling sensitive data collected by a company, it is mandatory to provide the necessary security measures. In Hadoop, the security measures are disabled by default. The person responsible for data should be aware of this fact and take the required measures to secure the data.

2) Small Data concerns:

There are a few big data platforms in the market that are not fit for small data functions. Hadoop is one such platform wherein only large business that generates big data can utilize its functions. It cannot efficiently perform in small data environments.

3) Risky functioning

Java is one of the most widely used programming languages. It has also been connected to various community because cyber criminals can easily exploit the frameworks that are built on Java. Hadoop is one such framework that is built entirely on Java. Therefore, the platform is vulnerable and can cause unforeseen damages.

Applications:

1) Amazon:

To build Amazon's product search indices; process millions of sessions daily for analytics, using both the Java and streaming APIs; clusters vary from 1 to 100 nodes.

2) Yahoo! :

More than 100,000 CPUs in ~20,000 computers running Hadoop; biggest cluster: 2000 nodes (2*4cpu boxes with 4TB disk each); used to support research for Ad Systems and Web Search

3) Facebook:

To store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning; 320 machine cluster with 2,560 cores and about 1.3 PB raw storage.

II) CONCLUSION:

Managing the data is the big issue. And now days the huge amount of data is produced in the origination so the big data concept is in picture. It is data set that can manage and process the data. For managing the data the big data technique is used i.e.hadoop. Hadoop can handle the huge amount of data, it is very cost effective, and it can handle huge amount of data so processing speed is very fast,and also it can create a duplicate copy of data in case of system failure or to prevent the loss of data.

REFERENCES

[1] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar," A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, 10, October 2014.

[2] Yashika Verma, Sumit Hooda," A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications,Feb,2015.

[3] IqbaldeepKaur, NavneetKaur,AmandeepUmmat, JaspreetKaur,NavjotKaur," Research Paper on Big Data and Hadoop", International Journal Of Computer Science And Technology publicatios,Oct-Dec 2016.

[4] Tom White,*"Hadoop: The Definitive Guide"*,O'Reilly Media Publication ,3rd Edition
[5] Garry Turkington,"Hadoop: Beginner's Guide",Packt Publishing,2013
[6] http://www.tutorialpoints.com/hadoop _overview.pdf
[7] http://blogs.mindsmapped.com/bigdatahadoop
[8]]https://www.knowledgehut.com/blog/bigdata-hadoop/top-pros-and-cons-of-hadoop
[9] https://readwrite.com/2013/05/23/hadoop/applications
[10] https://datajobs.com/what-is-hadoop-and-nosql
[11] https://www.quora.com/What-is-the-difference-between-HBase-and-Hadoop
[12] https://www.sciencedirect.com/science/article/pii/S221457961730014X#se0160