International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The Big Data analytics with Hadoop

Swamil Singh[1], Viplav Mandal[1], Dr. Sanjay Srivastava[3]
Computer Science Department, MGM's College of Engineering And Technology, Noida

Abstract:- The big data is a huge number of data ,all these data comes from the daily uses of human life like telephone, mobiles, travelling, shopping, computer, organization data which are to be store, manage and executing the data's from the data server. These type of data are so big, if we calculated these type of data goes in the Tb, Pb or in also zettabytes size, so size of the data are so much that's why it's called BIG DATA. The data storation and management are create a big problems and the form of these data's are may be in structured, unstructured, or semi structured form.

To solving this problem, we have Hadoop platform to analysis these data to using map-reduce to get a desire output with the help of programs Or Hadoop is only the way to come out from this problem in a desired output. Hadoop is the only a single or core platform for structuring the analytics the data gets the desired output from these problems. It's an open-source platform or it's freely available to perform these type of operation. The design purpose of it's to be perform millions of data in a single executive server.

Keywords:-Big data, zettabytes, Hadoop, Map-reduce, HDFS, other concepts

I. INTRODUCTION

A. *Big data*

The term big data is which is used to describe the growth and its availability of the data in the IT sector or other sector. Big data analytics refers to the process of collecting the organizing and analyzing large sets of data to discover the pattern and other useful information.

Basically we can say that the big data is a combination of the Big 5' V i.e., volume (size), velocity (time sensitive), variety (structured, semi-structured), veracity (consistency), and the ratio (20% or 80%). of the data that is structured or unstructured. The big data arise from the internet, mobile data, business data or other etc. All these data comes in the Pb or Zb, now its increasing day by day and time to time as growth of the IT sectors. All this data is a huge data which is a challenge for the common or normal software which is already used in the IT sector but the system are not handle very well this kind of large scale of data.

So, it's a challenge for the searches to analyzing the big data and its need special techniques to analyzing. Big data analyzing is the biggest challenge to give more accuracy because of the term which are hidden data or relations with other data that all are need to be uncover.

B. *Challenges*

1) *Size:* The name (Big Data) which is indicate that the data is big in the size which growth rate is high as compare to the last few recent years, that time the increasement of the data are low as compare to current time. The size of the data breaking all boundaries which is reached at top of the peak of the data storage because all these data are stored in the use of the future purpose. The data size goes to the Petabytes or zettabytes, which is a typical work to manage all these data for further use in the machine.
2) *Speed & Time*: The speed is the major challenge to get the desire output from huge of the data(structured or unstructured) and maintain the same speed to obtain output for the whole process in the quick time as compare to the other devices. The organization need the data as quickly as possible because the volume of the data. Hadoop is quick as compare to the other data to get the desire output.

C. *Performance*

The performance of the big data it may be challenging for the other system but in this technology handle the big data very as compare to the other. The performance is totally depend on you tools of the Hadoop which are play crucial role to get the desire output.

D. *Accuracy*

Accuracy is the main thing in any system to obtain desire output from the big data. Without accuracy you can't be major the things which are correct or worst. Accuracy is one of the other factor which play an crucial role. Without accuracy you can't be major

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

anything which gives a valid output or other things.

E. Hadoop

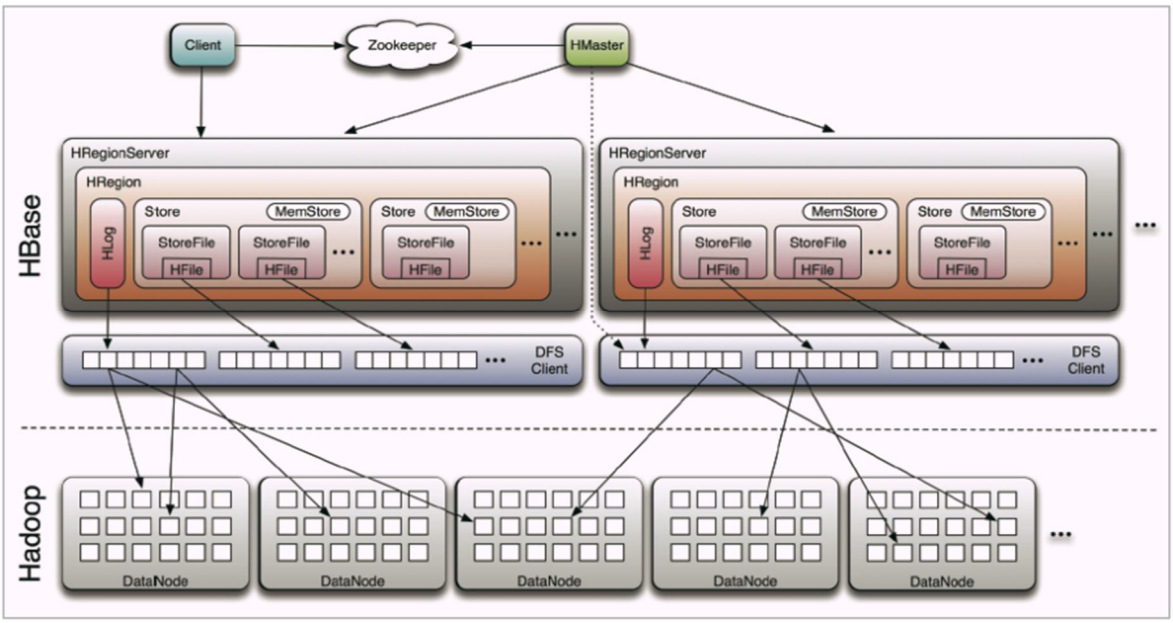
Hadoop is a software technology which is used for storing and processing large amount of distributed data stored in different cluster. The Hadoop ecosystem consist few tools i.e. Map Reduce, Hive, Flame, HDFS which are very important to execute the processing the data. Hadoop enables businesses to quickly gain insight from massive amounts of structured and unstructured data. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce).

Here some tools which are playing an important role in the Hadoop environment,

1) Hadoop platform
2) Map reduce
3) HDFS

1) *Hadoop platform:* The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (Map Reduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop Map Reduce transfers packaged code for nodes to process in parallel, based on the data node needs to process. This approach takes advantage of data locality—nodes manipulating the data that they have on hand to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

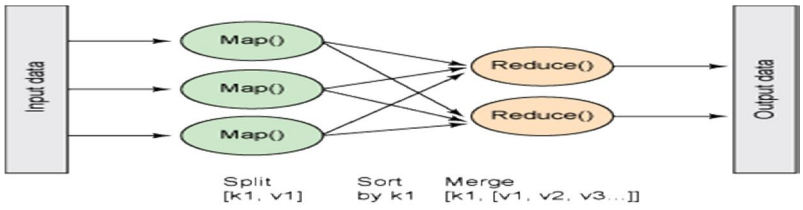
a) *Hadoop architecture:* Hadoop is an open source Apache project started in 2005 by engineers at Yahoo, based on Google's earlier research papers. Hadoop then consisted of a distributed file system, called HDFS, and a data processing and execution model called Map Reduce. The Apache Hadoop architecture consists of the Hadoop common package, which provides file system and operating system (OS)-level abstractions, a MapReduce engine and the Hadoop Distributed File System (HDFS).To store a large file on the HDFS, the input file is split into smaller data sets and sent over to different nodes (servers) for parallel processing of data and the nodes hold the processed data. The framework, which is used for overall processing of data, is called MapReduce.



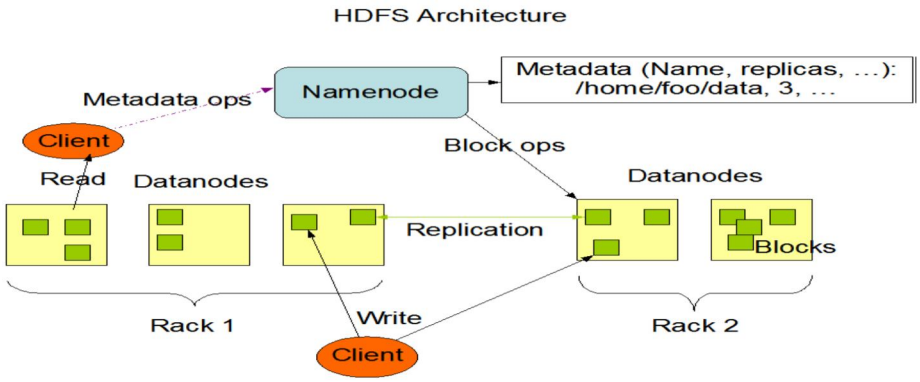
2) *Map reduce:* Map reduce (map + reduce) is a framework for writing applications that process large amounts of structured and unstructured data stored in the Hadoop Distributed File System (HDFS). Map reduce is a tool which are very crucial to analysis of the data which are structured or unstructured. In which the data are splits in the small parts of the cluster to get the original form of the data. The MapReduce concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions. For people new to this topic, it can be somewhat difficult to grasp, because it's not typically something people have been exposed to previously. Map reduce is 100 times faster than the Spark and 1000 times faster than the Drill, now Drill is also working platform.

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

For e.g.- CPU is the brain of computer like that Map reduce is the heart of the Hadoop.



3) *HDFS:* HDFS (Hadoop distributed file system) is a file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers. HDFS has demonstrated production scalability of up to 200 PB of storage and a single cluster of 4500 servers, supporting close to a billion files and blocks. HDFS take different types of file which are likes context, images, videos and etc.



II. FUTURE JOBS

The future of the Hadoop (big data) is so bright in the upcoming years. Most of the IT companies stored their relevant data and handling of these data's are a typical task for management. A report says that the upcoming 2017 comes with the great opportunity for the big data in IT sectors. Now its become so popular in the IT sector for the management of the data. Its becomes so popular as compare to other software for the need of speed, accuracy and other things in a single software.

III. CONCLUSION

The huge demand of this technology in the market. There are huge volume of data is lying in the organizations but there is no other technologies like this one who can easily handle every type of data in the low cost of the hardware & used by the large set of audience. We studies in this paper, Hadoop is software which is easily handles the data that is structured or unstructured in large amount of the data. Big data analytics with Hadoop, in which Hadoop analysis the data very quick and view the analysis in excel sheet and the graphical form and the view of map representation.

REFERENCES

[1] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M "Efficient Analysis of Big Data Using Map Reduce Framework".
[2] Chen He, Derek weitzel, David Swanson, Ying lu "HOG: Distributed Hadoop Mapreduce on the Grid".
[3] Jeffrey Shafer, Scott Rixner, and Alan l. Cox "The Hadoop Distributed Filesystem: Balancing Portability and Performance".
[4] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler "The Hadoop Distributed File System".
[5] Suman Arora, Dr.Madhu Goel "Survey Paper on Scheduling in Hadoop".
[6] Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simpli_ed Data Processing on Large Clusters"
[7] Hadoop,"PoweredbyHadoop,"http://wiki.apache.org/hadoop/Powere dBy.
[8] Hadoop Distributed File System (HDFS), http://hortonworks.com/hadoop/hdfs/Hadoop http://hadoop.apache.org/mapreduce/.
[9] www-304.ibm.com/easyaccess/fileserve?contentid=217007Foundation,"Yarn,"https://hadoop.apache.org/docs/r0.23.0/hadoopyarn/hadoop-yarn-site/YARN.html.
[10] Hadoop Tutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html
[11] Bakshi, K.,(2012)," Considerations for big data: Architecture and approach"
[12] Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review"