 ISSN: 2321-8169
63 – 66

Hadoop - A solution for Big Data Processing

Ravikiran G. Deshmukh
Department of CSE
MGI-COET, Shegaon, India
Email Id: - ravideshmukh611@gmail.com

Vaishali A. Kshirasagar
Department of CSE
MGI-COET, Shegaon, India
Email Id: - vaishalikshirasagar96@gmail.com

Pooja G.Mankar
Department of CSE
MGI-COET, Shegaon, India
Email Id:- mankar96pooja@gmail.com

Nayana G. Tale
Department of CSE
MGI-COET, Shegaon,
Email Id: - talenaina@gmail.com

Mahadeo J. Pathak
Department of CSE
MGI-COET, Shegaon, India
Email Id: - mahadeopathak1@gmail.com

***Abstract*—**The amount of data which is being generated by us is growing day by day. Traditional approaches fails to manage such largevolume of data**.** The term Big Data refers to data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to capture, manage, process or analyzed. To analyze this enormous amount of data, Hadoop can be used.Various Big Data tools and techniques are already in use for managing huge data efficiently and effectively. Among the widely available tools and techniques, Hadoop plays a major role in the IT market .The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications**.**MapReduce programming model is widely used for large scale and one-time data-intensive distributed computing.

Keywords: *Big data,Hadoop,HDFS,MapReduce, Hadoop Components.*

*****

I. INTRODUCTION

Big Data is as a collection of large dataset that cannot be processed using traditional computing techniques.Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework. The need of big data generated from the large companies like Facebook, yahoo, Google, YouTube etc. To manage that huge amount of datawe use Hadoop.

Hadoop is an open source software framework for storing data and running applications on cluster of commodity hardware. It provides massive storage for any kind of data,enormous processing power and the ability to handle virtually limitless concurrent task or jobs.

Hadoop provides distributed file system and a frame work for the analysis and transformation of very large data sets using the MapReduce paradigm. An important characteristic of Hadoop is the partitioning of data by computationacross many (thousands) of hosts, and executing application computations in parallel close to their data.MapReduce is emerging as an important programming model for data-intensive applications. The model proposed by Google is very attractive for ad-hoc parallel processing of arbitrary data. It shows good performance for batch parallel data processing.

II.THE CHALLENGES OF BIG DATA:

1. Volume:

Volume refers to amount of data. Volume represent the size of the data how the data is large. The size of the data is represented in terabytes and petabytes**.**

2. Variety:

Variety makes the data too big. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

3. Velocity:

Velocity refers to the speed of data processing. The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive.

4. Value:

The potential value of Big data is huge. Value is main source for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5. Veracity:

Veracity refers to noise, biases and abnormality. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

III.HADOOP:(SOLUTION FOR BIG DATA PROCESSING)

Hadoop is a faster, cheaper database and analytics tool. Hadoop is increasingly becoming the go-to framework for

 ISSN: 2321-8169
63 – 66

large-scale, data-intensive deployments. Hadoop is built to process large amounts of data from terabytes to petabytes and beyond. The beauty of Hadoop is that it is designed to efficiently process huge amounts of data by connecting many commodity computers together to work in parallel. Hadoop is open-source software that enables:

- **Reliable**: The software is fault tolerant; it expects and handles hardware and software failures.
- **Scalable:** Premeditated for huge size of processors, memory, and neighboring attached storage space.
- **Distributed**: Handles duplication. Offers particularly parallel programming model, Map Reduce.

IV. HADOOP BENEFITS:

- Cost effective – Because it works on commodity hardware.
- Big cluster (1000 nodes on cluster) – Big storage and more processing power due to number of nodes on cluster.
- Parallel processing – In MapReduce framework thousands of nodes processes in parallel and generate results in timely manner.
- Big storage – Thousands of nodes with high capacity (100 GB). So storage capacity increases
- Failover – Automatic failover. If node crashes, framework will identify automatically.
- Data distribution – Hadoop framework handles itself because it has thousands of nodes.
- MapReduce framework – Designed and will work on hadoop framework as map and reduce function.

V.Why Hadoop?

Working with Hadoop is quiet simple with the knowledge of Core Java and few related concepts of Data Warehousing. Hadoop library has been designed in such a way that it automatically identifies and handles failure which makes it more efficient in the way it does not need to depend on any hardware platform to detect failures. Any server can be removed from the cluster or added to the cluster dynamically and Hadoop continues with its operation. Hadoop is designed in such a way that it can work with any platform.

VI.Who are the users of Hadoop?

Amazon uses Hadoop to build their product search indices and also to process their millions of sessions. Adobe uses it internal data storage and processing. Cloud-space is using Hadoop for their client projects. Hadoop is been used by eBay for their search optimization and research. Facebook uses Hadoop for machine learning and to store their copies of internal log. IIT, Hyderabad uses Hadoop for Information Retrieval and Extraction research projects. Last.fm uses it for charts calculation and dataset merging. Twitter is also using it to manage the data that is been generated daily in their website. Apart from these big players, IBM, Rackspace, The New York Times, LinkedIn, University of Freiburg, University of Glasgow and lot more are using Hadoop.

VII.Who are Hadoop Vendors?

Everybody around the world is gaining knowledge about Big Data. Many vendors today support Hadoop to a greater extent. Notable ones are AWS, Cloudera, Microsoft, MapR, Oracle and IBM.

VIII.Scheduler in Hadoop:

A scheduler plays a very important role in the big data processing. A fair scheduler schedules the jobs in such a way that all the resources share by the command or by the system in equal amount without any over loading on one of the part of the system. Like scheduler, will take care of the resources on each data node. It helps to maintain the load onthe entire data node in the system.

How does scheduling help in the processing of big data?

Take one example, suppose we have ten data node in Hadoop cluster, and our scheduler is not fair it cannot manage resources in a right manner. So what does he do, he give work on 5 data node out of 10 data node in the cluster, and suppose it take around x amount of time to complete that command. Now thinkthat our scheduler is fair enough to distribute work on the entire data node in our cluster, Soaccording to our assumption it will take around x/2 amount of time to complete the whole process.

IX.TOOLS AND COMPONENT

10.1Ambari:

Ambari is the project developed by the Apache Software foundation to support Hadoop by making its management simpler by maintaining the Hadoop

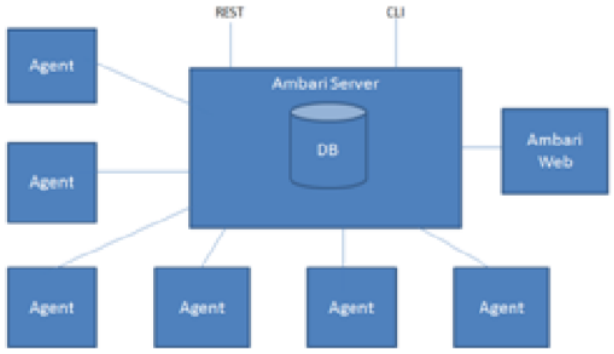

Fig. 1 Ambari Architecture

Clusters. It provides an environment which is easy to maintain using its RESTful APIs. Using Ambari, the system administrators can easily manage, provision and monitor a Hadoop cluster. Various Operating Systems which supports Ambari are OS X, Windows and Linux. Fig. 2 depicts the Ambari architecture.

10.2HBase:

Apache HBase is an open source, distributed database which was built on top of HDFS. It aims at storing millions and billions of data with support to fault-tolerance. HBase is similar to that of the Google's Bigtable. Although it supports the storage of large databases, it is not a direct replacement of the SQL database. It's perform

Once is increasing in the recent days and now it supports the messaging platform of Facebook. HBase is based on Java and is OS independent. The architecture of HBase is depicted in Fig. 2.

 ISSN: 2321-8169
63 – 66

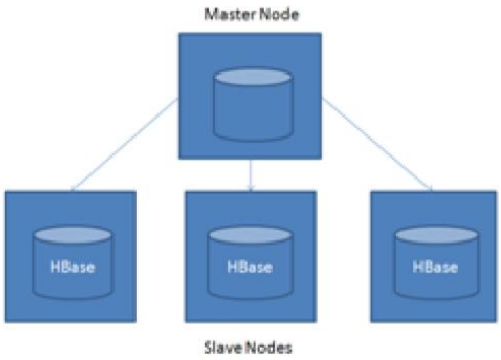

Fig. 2. HBase Architecture

10.3Tajo:

Tajo is the datawarehouse which is distributed for managing the Big Data. It was developed by Apache. Initially, it uses HDFS as the storage layer and the storage gets completed with its own query engine. This query engine will allow direct control of execution and also data flow. This is because it has various evaluation strategies, SQL standardsand also optimization methods. The supported OS of Tajo are Linux and Mac. Fig. 3. depicts the Tajo architecture.

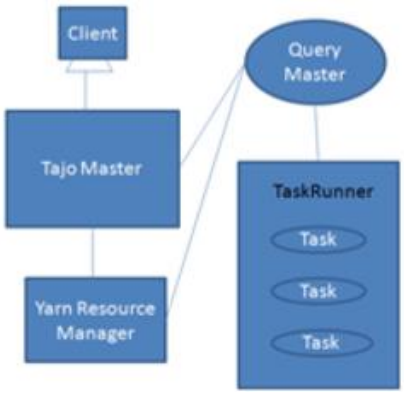

Fig. 3.Tajo Architecture.

10.4MapReduce:

MapReduce is a framework for handling the huge volume of datasets in a cluster. MapReduce was initially found by Google. It consists of a Map() phase and a Reduce() phase. The Map() phase will perform the sorting and filtering operations and the Reduce() phase will then perform a summary operation on the sorted data. It is said to be the heart of Hadoop. It is OS independent. Fig. 5.is the architecture of MapReduce.

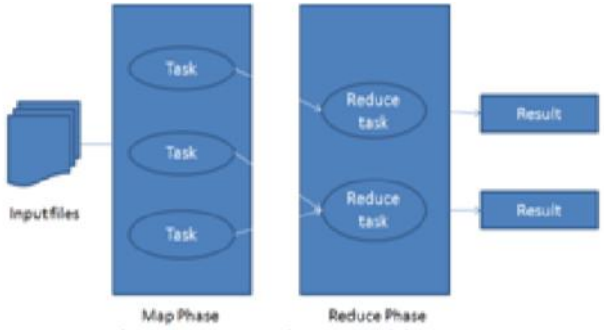

Fig. 4. MapReduce Architecture

10.5HDFS:

HDFS (Hadoop Distributed File System) is the distributed file system which is based on Java for the storage of large datasets. HDFS was initially developed by Apache and now it is the sub-project of Apache Hadoop. HDFS provides high fault-tolerance when compared with the other distributed file systems. It also provides high throughput, scalability and can be deployed on hardware of low-cost. Windows, Linux and OS X are the Operating system which supports HDFS. HDFS architecture is shown in Fig. 6. HDFS architecture is broadly divided into following three nodes which are Name Node, Data Node, HDFS Clients/Edge Node

1. Name Node :

It is centrally placed node, which contains information about Hadoop file system . The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information about thesystem .and provides information which is newly added, modified and removed from data nodes.

2. Data Node

It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance . A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. HDFS Clients/Edge node:

HDFS Clients sometimes also know as Edge node . It acts as linker between name node and data nodes. Hadoop cluster there is only one client but there are also many depending upon performance needs .

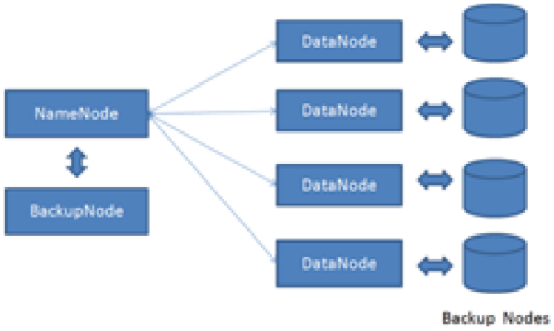

Fig. 5.HDFS Architecture

10.6Hive:

Hive is the open source project which was developed by Facebook. Same like Tajo, Hive is also the data warehouse for managing large datasets which is based on Hadoop. It uses HiveQL which is similar to SQL language, for managing the datasets. Being inconvenient to use HiveQL, it also allows the designers to use MapReduce concept. It is OS independent. The architecture of Hive is depicted in Fig. 6.

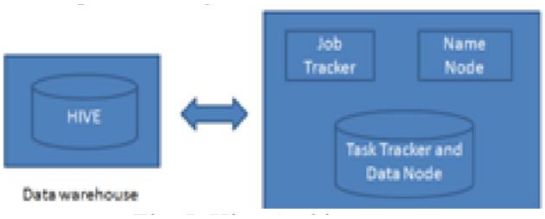

Fig.6. Hive Architecture

XI.CONCLUSION

 ISSN: 2321-8169
63 – 66

To analyze enormous amount of data, various Big Data tools and techniques are already in use for managing huge data efficiently and effectively. Among the widely available tools and techniques, Hadoop plays a major role in the IT market. The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications which makes it very useful while dealing with Big Data.

REFERENCES

[1] Varsha B. Bobade," Survey Paper on Big Data and Hadoop", International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 0Jan-2016.

[2] Andrew Pavlo, "A Comparison of Approaches to Large-Scale Data Analysis", SIGMOD, 2009.

[3] Shilpa Manjit Kaur," BIG Data and Methodology-A review" ,International Journal of Advanced Reseat in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[4] Suman Arora, Dr.Madhu Goel," Survey Paper on Scheduling in Hadoop", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 5, May 2014

[5] Deepika P*, Anantha Raman G R," A Study of Hadoop-Related Tools and Techniques", International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 9, September 2015

[6] Puneet Singh Duggal and Sanchita Paul, "Big Data Analysis: Challenges and Solutions," Nov. 2013

[7] Sangeeta Bansal and Dr.Ajay Rana, "Transitioning from Relational Databases to Big Data," in vol.4, Issue 1, Jan 2014

[8] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, "Big Data: Issues and Challenges Moving Forward," 2012

[9] A. Katal, M. Wazid, R.H. Goudar, "Big Data: Issues, Challenges, tools and Good practices," Aug. 2013

[10] Solanke Poonam G. and B. M. Patil," INCREMENTAL MAPREDUCE FOR BIG DATA PROCESSING", International Journal of Computer Engineering and Applications, Volume X, Issue II, Feb. 16

[11] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu,Member, IEEE, "i2MapReduce: Incremental MapReduce for Mining Evolving Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015.

[12] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler," The Hadoop Distributed File System".