A Study on Hadoop HDFS and MapReduce

M. Sona[1] C. Vinola[2] M. Asia Sahaya Christina[3] N. L. Anto Arokia Mary[4]
[1,4]P.G. Scholar [2,3]Assistant Professor
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Francis Xavier Engineering College, Tirunelveli

Abstract— Big data analysis is emerging as computational archetype. We are in the era of Big Data which is defined as a distinct collection of complicated and huge sets of data. The Data originates from Social Medias, Sensors, Digital Cameras, GPS signals and more that leads to compose big data. Big data is difficult to gather, progress, store, explore and analyze using conventional methods; it requires massively parallel software running on multiple numbers of servers. One of the tools to exploit Big Data is Hadoop which is a non-relational data store to handle variety of data. Hadoop Distributed File System and MapReduce are the main components of Hadoop.

Key words: HDFS, MapReduce, Hadoop

I. INTRODUCTION

The hasty development in networking, storage facility and processing speed of computing devices in last few years have given emerge to innovative applications which entails accessing and storing more than terabytes of data. Nowadays, we are creating approximately quintillion bytes of data with different types in a fraction of seconds from everywhere such as social networks, sensors, digital pictures and images, transaction records etc. This data is called "Big Data". The aspects of Big data are:

- Volume – huge amount of data
- Velocity – rapid movement of data
- Variety – different patterns of data which includes structured, unstructured, semi-structured

Big data also provides an opportunity to know impends for up-coming and default types of data and also provides contents for business in more approachable manner. The main challenges of data processing in big data are parallelization, Distribution, Scheduling, Monitoring, Storage capacity and fault-tolerance. To dig out meaningful value from big data, you need optimal processing system. For processing this data, the traditional method based on relational database system is not well suited to handle. The reason is RDBMS uses SAP Sybase IQ which follows column-store technique to compress the data efficiently. Big data needs parallel processing system. Hadoop acts as a tool to provide a parallel distributed environment. That analysis of big data provides a value which may be the results of expressive, projecting, and authoritarian.

This paper gives an idea about the following contents: Hadoop is an open source java framework used for handling huge sets of different data which is discussed in Section II. Section III provides the description about Hadoop Distributed File System. MapReduce is used to process large sets of data in a parallel manner which is explained in the Section IV. Section V described the architecture of Hadoop.

II. HADOOP

A. History of Hadoop:

In 2002, Doug Cutting and Mike started their works on "Nutch". In 2003, Google publishes the paper of Google File System and MapReduce Paradigm. This GFS and MapReduce are the basis for Hadoop development. In 2004, Doug uses the Distributed File System and MapReduce concepts on Nutch. In 2006, Yahoo appoints Doug to spin out Nutch for Hadoop.

B. Hadoop and its Family:

Hadoop is open source software which is managed by Apache for storage and processing large sets of data. Hadoop is not used for creating Application development or software development; used for only analyzing big data. Hadoop is used to analyze the data which is more than Terabytes. It is developed for data-concerted tasks, for that purpose it follows the concept of moving the software to the data rather moving the data into the computation location. Hadoop is a hub of cloud computing infrastructure and also used many of the companies likes Yahoo, Facebook, LinkedIn etc. Hadoop Ecosystem has the following components:

- HDFS
- MapReduce
- Sqoop
- Flume
- Hive
- Pig
- Zookeeper
- HBase
- Oozie

Hadoop Distributed File System (HDFS) is the heart of Hadoop. HDFS contains both input and output files. Hadoop MapReduce is the mixtures of two functions which is map function and reduce function. Initially, the map function breaks up the data across the different nodes and process it in parallel. Then, reduce will use the output of map function as the input to produce the combined result.

HDFS and Hadoop MapReduce are the two main components of Hadoop. Namenode, Secondary Namenode, Datanodes, JobTracker, TaskTracker are the daemons of Hadoop.

Sqoop is used to bring the structured data to the HDFS whereas Flume is used to bring unstructured and semi-structured data to HDFS. Sqoop and Flume are used to write the data to HDFS. Sqoop used for Data Exchange. Flume used for controlling the logs.

To read the data from the HDFS, Hive and Pig are used. Hive is a Query language used to analyze only structured data whereas Pig is used to analyze both structured and unstructured data. Pig is also called as Pig

Latin. Pig is a high-level platform which allows extracting, transforming and loading the data.

Zookeeper is used for coordinating the logs. HBase is a distributed columnar based database. Oozie is used for workflow and co-ordination between the MapReduce jobs. The advantage of Hadoop is Flexible, Scalable, Reliable, Fault-tolerant etc.

III. HADOOP DISTRIBUTED FILE SYSTEM

Hadoop distributed file system is the module of hadoop which can hold huge volumes of data. Hadoop developed this distributed file system in the brainwave of Google file system. Hadoop Distributed File System holds both input files and output files. It acts as an administrator to perform action of addition, subtraction of nodes to/from the Hadoop Cluster.

A. Hadoop Cluster:

Hadoop Cluster is a collection of racks. Each rack is a collection of *n* number of nodes. Hadoop Distributed File System consists of two kinds of nodes in cluster. They are

- Namenode
- Datanode

These nodes are used for offline processing. Nodes in hadoop used for writing the data once and can be reading the data multiple times. A Hadoop cluster has single Name node and multiple number of Datanode.

B. Name Node:

Namenode acts as master node which provides the metadata. Namenode is used for file name operations. It handles the blocks which are in DataNode. The secondary namenode in hadoop which acts as backup namenode temporarily for namenode while its fail.

C. Data Node:

Datanodes acts as slaves. Datanodes are used to perform read, write operations by the clients. Hadoop has multiple number of datanode. To tolerate the node failures, all the data are splits into blocks (default size of 64MB) and replicated it three times and stores the replicated copies in three different nodes. A single HDFS has a block size of multiple Operating System block size.

D. Backup Node:

Backup node is also known as Edge node. It is responsible for identifying the failure of Namenode using periodic checkpoints. The advantages of HDFS is simple, scalable, provide streaming read performance.

IV. MAPREDUCE

MapReduce framework was developed by Google. For processing large sets of data, Google developed MapReduce and Google File System. Google used MapReduce to index the content in web. However the Google's deployment is not an open source, there was an open source platform called Hadoop. Hadoop is used by many cloud computing industries.

Other than Hadoop, some commercial versions are also available such as Microsoft Daytona, Amazon Elastic MapReduce service. According to Dean and Ghemwat [1], MapReduce is used in many real-world applications. So, this programming prototype should provide efficient performance.

MapReduce framework involves in two functions: Map function and Reduce function. First, each input files are processed by map function in parallel. Then, sorts the result comes from output of the map function. Finally, the output of the map function is processed by reduce function.

MapReduce layer contains the following components:

- Jobtracker
- Tasktracker

When the job submitted to the Jobtracker from the Jobbclient, the Jobtracker assigns the tasks to the respective tasktracker. Tasktracker is used to perform the tasks. The JobTracker and TaskTracker involve in reading the data from HDFS but the JobTracker only perform writing data into the HDFS.

A. Jobtracker:

Jobtracker present in the master node. The jobtracker provides the progressing details of their node i.e., namenode. The JobTracker is responsible for assigning the tasks to the TaskTracker.

B. Tasktracker:

The tasktracker is present in slave nodes. The tasktracker is responsible for perform the assigned tasks. The TaskTracker sends the heartbeat messages to the Jobtracker every 10 secs which indicates their activeness. When the tasktracker slows in processing the tasks, the same job is assigned to another tasktracker as backup. This execution is called speculative execution. Due to faults which may be communication failure or process failure in the tasktracker, the jobtracker will reassign the same tasks to another tasktracker.

C. Job History Server:

Job History Server is also one of the daemon which provides the accomplished applications details.

V. HADOOP ARCHITECTURE

Hadoop Architecture has master/slave architecture. Hadoop Architecture is shown below which describes the daemons such as JobTracker, TaskTracker, Namenode, and Datanodes. Namenode acts as master node which provides the metadata. Namenode is used for file name operations. The secondary namenode in hadoop which acts as backup namenode temporarily for namenode while its fail. Datanodes acts as slaves. Datanodes are used to perform read, write operations. Jobtracker acts as master whereas the tasktracker acts as slave. The jobtracker provides the progressing details of their node i.e., namenode. The tasktracker sends the heartbeat messages to the Jobtracker every 10 secs which indicates their activeness. In the figure, green box represents the HDFS layer and blue color box indicates the MapReduce layer.

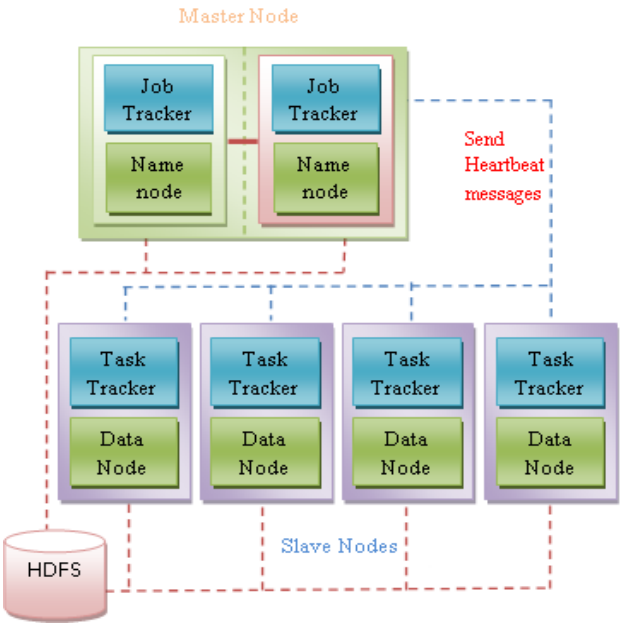

Fig. 1: Hadoop Architecture

VI. CONCLUSION

In this paper, we have studied Big data, distinctiveness of big data, traditional examine method for big data and their drawbacks. We have also discussed about Hadoop and its advantages. Hadoop is an open source java framework used for handling huge volumes of data. Hadoop uses the MapReduce concept to process the large sets of data in a parallel manner.

REFERENCES

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. Sixth Symp. Operating Systems Design and Implementation, Dec. 2004.
[2] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google File System," Proc. 19th ACM Symp. Operating Systems Principles, pp. 29-43, 2003.
[3] T.White, " The Definitive Guide", O'Reilly, 2009
[4] http://en.wikipedia.org/wiki/Apache_Hadoop
[5] http://en.wikipedia.org/wiki/Big_data

