Apache Hadoop Distributed File System

Nilesh Vishwasrao Patil
M.E. Computer Engineering
patil.nilesh38@gmail.com

***Abstract*— Apache Hadoop has been playing the vital role in market for storing and processing the big data. Apache Hadoop is uses the Hadoop distributed file system (HDFS) for storing the big data in distributed computing environment. It is used Map reduce to analyse and process such huge amount of data. Hadoop distributed file system is designed for storing large amount of data sets with highest amount of reliability and which stream those data sets at high bandwidth to the user applications. Apache Hadoop is provides the greatly fault tolerance and also it deployed on the low cost hardware.**

***Index Terms*—Hadoop, HDFS, Big Data, MapReduce, Hbase**

I. INTRODUCTION

OW days the each and every business is shifting to the Nelectronic business for providing the service to clients anytime and anywhere (24 x 7) which leads to promote the business and earn huge amount of money. As per IDC survey around 80 per cent of electronic data with respect to available data is generated in last two years which shows how the speed of generation of electronic data.

The buzz word in electronic business is “Big data”. It is very common term in not only in IT industries but also in each and every electronic business. We need a powerful distributed environment to store such big data and also analyse such big data for extracting information from big data for providing statistics, graph etc. Apache Hadoop is framework for providing distributed computing environment for storing and processing the big data. Apache Hadoop is providing the framework and hadoop distributed file system for the analysis and transformation of the large set of data with the help of Map reduce.

An important property of the Hadoop framework is splitting of large data and computation across the thousands of hosts by executing the applications in parallel. The Hadoop cluster consists of thousands of node with one master node and others are slave node. The server or master and also the all slaves which commodity hardware so the industry does not require purchasing expensive server and hardware. It has provided the highly computation capacity, IO bandwidth, and storage capacity by splitting of data into multiple blocks and stores on different slaves parallel. All components of the Apache Hadoop are available with open source Apache license. There are many vendors like Facebook, Yahoo, Microsoft has contributed by providing some applications projects to Hadoop. Following table contains some components of Hadoop.

Table 1 Components of Hadoop (Application Projects) [1]

Name	**Uses**
HDFS	Hadoop Distributed file system, which will discuss in this paper in detail.
MapReduce	It is uses for processing big data. It is computation framework for the big data
Pig	It is framework for analysing the large set of data. It allows user for creating own functions to do special task.
Hbase	It is column oriented table service
Hive	It has provided data ware house infrastructure which is based on SQL like queries
ZooKepper	It is distribution coordination service.

HDFS is the file system component of Hadoop framework and built on Unix file system. HDFS file system stores the metadata and application data is separately. In Hadoop framework the metadata information is stored in one dedicated server called as Namenode while the application data is stored in other nodes (slaves) called as Datanode. It has one Namenode while the multiple Datanode. Master and Slaves are fully connected with each other they are communicated with the help of TCP based protocol.

Hadoop distributed file systems does not uses any protection mechanism like RAID for the data durable. HDFS uses the replication mechanism for replicated application data

Mr. Nilesh Vishwasrao Patil, Author. ME Computer Engineering, Government Polytechnic, Ahmednagar.

to multiple nodes for reliability and availability of application data like Google File system (GFS). The replication techniques is provides many advantages including durability, higher speed to access data (increased bandwidth) by replicating data where needed more by analysing the geographical pattern of application data requirement.

II. Architecture components of HDFS

HDFS cluster consist of namenode and datanode. Namenode is master node which manages the cluster metadata and data node stores the data. The data files and directories are represented on Namenode with the help of inode which is based on Linux operating system. The files contents are divided into blocks which may be 64MB or 128MB and each block is replicated independently on different nodes, for replication of blocks to the different geographical location to take back up in case any loss of data due to natural or artificial climes. The blocks are stored into local file system of datanodes. The namenode is monitoring the different replicas of different blocks. The namenode maintains the namespace tree and the mapping blocks of the datanode. The HDFS architecture is as shown in following figure 1.

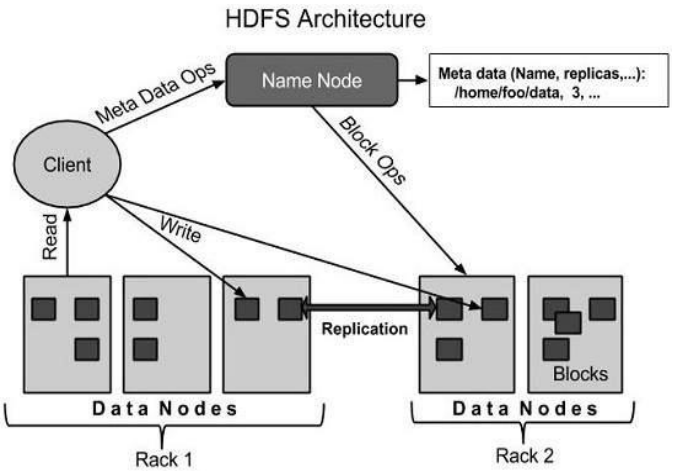

Figure 1: HDFS Architecture

The following figure 2 described, the HDFS client send data file to the Namenode by providing path for metadata information about the file. The data file is divided into multiple blocks, for each block of data file; the NameNode gives the list of DataNodes to store its replicas. The HDFS client creates the pipelines for DataNodes to replicates data blocks, which eventually confirm the creation of the block replicas to the NameNode.

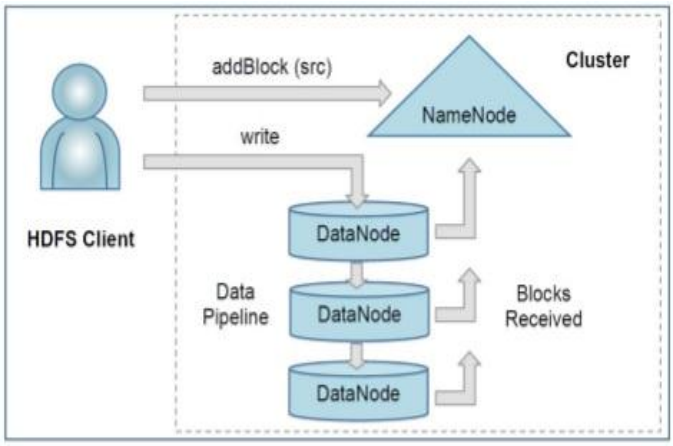

Figure 2: Data blocks moving to Datanode[1].

III. HDFS characteristics

It is java and Linux based file system which provides scalable and reliable data storage. It was designed to store big data with large cluster and on commodity hardware. Following table 2 shows the important features of the HDFS file system.

Table 1 Important Characteristic of HDFS[3]

Feature	Description
Rack awareness of server	Need to consider following physical location of rack/server when allocating storage and scheduling tasks.
Minimal data motion	Instead of moving data towards the master and process it, instead of that passing program to the nodes and process it.
Utilities	Detect the health of rebalance and file system the data on different slaves by dynamically.
Rollback	Hadoop file system allows users to get back the earlier version of HDFS after an upgrade, in case errors
Provide Secondary Name node	In case of failure of namenode, can take backup of primary name node.

IV. Conclusion

Apache Hadoop framework is widely used by industries and research communities. Hadoop Distributed File systems based on Java and Linux file system. It gives highest amount of reliability and fault tolerance on commodity hardware.

References

[1] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. (2010). The Hadoop Distributed File System, *IEEE, 978, pp.12-16*

[2] Ivanilton Polatoa, Reginaldo R´e, Alfredo Goldman, Fabio Kon (2014). A Comprehensive View of Hadoop Research - A Systematic Literature Review. *Journal of Network and Computer Applications*, 46 pp. 1-28.
[3] Web site URL: http://hortonworks.com/hadoop/hdfs/#section_1

Mr. Nilesh Vishwasrao Patil. He is working as System Analyst in Government Polytechnic, Ahmednagar. He has completed Master of Engineering in Computer Engineering from SPPU, Pune. He has published six papers in International Journal and presented two papers in national journal.