UTILIZING HADOOP BASED DISTRIBUTED FILE SYSTEM FOR RELATIONAL DATA BASE

Arron Mesi.C[1], Gopika Vijayan[2] , Libin Rose.P[3] , Muthu.M[4] and Hema.**R**[5]
1,2,3,4 Final Year B.E, 5 Assistant Professor ,
CSE, Annai Vailankanni College of Engineering,
Email: arronmesi@gmail.com, gopikavijayan94@gmail.com, libinrose10@gmail.com
muthujegatheesh@gmail.com, hema25me@gmail.com

[1] ABSTRACT
Now a days there is an exponential growth of information and data in all the fields such as industrial areas, social networking, business areas, etc.. This paper describes the effective storage methodology to handle big volume of data. Also it deals with the effective utilization of distributed file system for analytical purposes by using specialized tools. HDFS is used to store the large data sets rather than the relational data base. HIVE is used to process those large data sets effectively. The paper can be used for various application area which deals with huge volume of data. The files in the database are stored in HDFS as it is fault tolerant and the file is availed easily for the users. To process the data from HDFS we use a processing technique Hive which is a platform used to develop SQL type scripts to do Map Reduce operations.
Key Words: BigData, HDFS, Hive, Sql, HQL

I. INTRODUCTION

As Big Data requirements shift to the general public, it is important to ensure that the methodologies and techniques which are worked well on small area. In this paper we developed a prototype for utilizing the Hadoop based distributed file system by replacing the relational data storage. It also works well for analytical purposes by retriving data from hdfs by using the tool callled Hive.

To process the data from HDFS we use a processing technique and program model called as map reduce. In order to make the processing easier queries are written to process the data and store it in HDFS HiveQL is used for writing queries which is equivalent to the program written in map reduce. Hive is a platform used to develop SQL type scripts to do Map Reduce operations. The Hive Query Language (HiveQL) is a query language for Hive to process and analyze structured data in a Metastore. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

II. HADOOP

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

1.1 Hadoop Architecture

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provide file system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.

- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.

- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

1.2 HDFS Architecture

Given below is the architecture of a Hadoop File System.

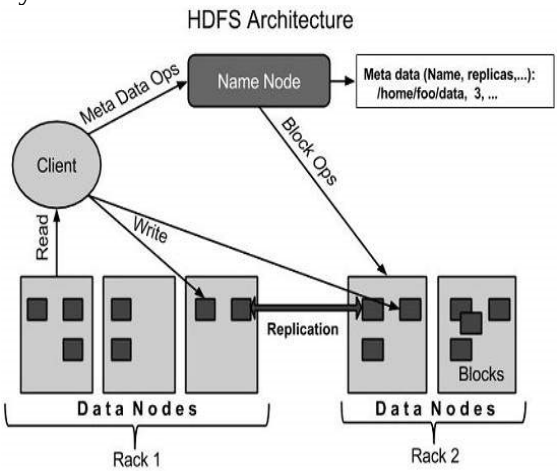


Fig (1.2) – HDFS Architecture

HDFS follows the master-slave architecture and it has the following elements.

- **Name node**

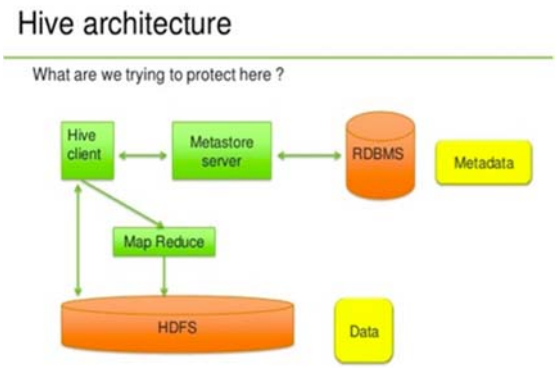
The name node is the commodity hardware that contains the GNU/Linux operating system and the name node

- **Data node**

The data node is a commodity hardware having the GNU/Linux operating system and data node software. For every node (Commodity hardware/System) in a cluster, there will be a data node.

III . HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.



Initially Hive was developed by Face book, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic Map Reduce.

IV. IMPLEMENTATION

In order to implement the proposed methodology, it requires the configuration of Hadoop and Hive. After the congifurations are over, create the table of contents and load the data into the table to the hdfs.

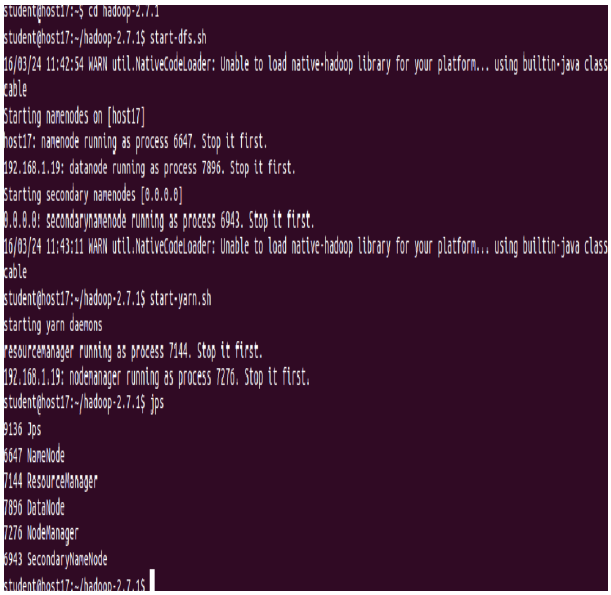


Fig (4.1) Starting Hdfs

Fig (4.2) Starting HIVE

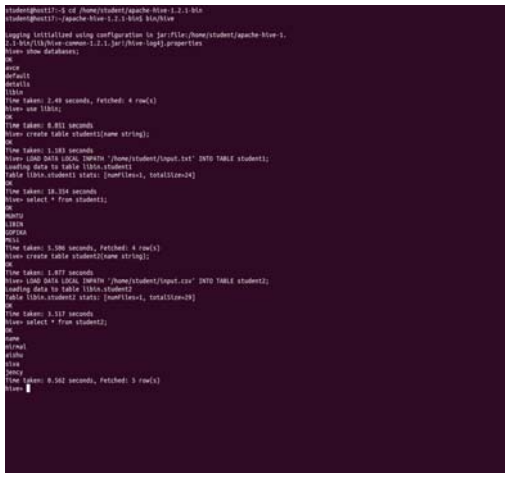


Fig (4.2) Loading data into the table

V. CONCLUSION

This paper describes the effective storage methodology to handle big volume of data. In this paper we proposed and implemnted a prototype for utilizing the hadoop distributed file system for relational data base with analytical tool called Hive. It is best suited for evaluating and analysing the large volume of data.

VI. FUTURE ENHANCEMENT

In our proposed model we have implemted the system with Hadoop and Hive which is a popular SQL interface for batch processing using Hadoop. Until now Mapreduce was the only execution engine in the hadoop ecosystem and hive queries could only run on MapReduce. Now alternative execution engine to mapreduce are available such as Apache Spark.

Spark is relatively new to the Hadoop ecosystem,but its adoption has been meteoric. It is an open-source data analytics cluster cluster computing framework and is built outside of Hadoop's two stage MapReduce paradigm but runs on top of HDFS.So in future we can integrate our proposed with Spark and the data processing can be made even more easier.