Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 9 (2017) pp. 2835-2840

http://www.ripublication.com

IMPORTING DATA FROM MYSQL TO HADOOP USING SQOOP

Dasari Anantha Reddy[1], Mohd. Jabeed Rihaz[2]

[1] *KITS-Ramtek, Information Technology Department, India.*

[2] *KITS-Ramtek, Information Technology Department, India.*

Abstract

Now-a-days, there is dramatically change in the generation of computerized data, data-sizes are increased from terabytes to petabytes which results into unstructured data generation. Maintaining the large size of data, processing the data, transforming the data, and sending data from one host to another host is big problem. To overcome these issues, concept of Hadoop is used which is open source distributed computing framework based on java and supports large set of distributed data processing. But in Hadoop Distributed File System, Master Name node Failure affects the performance of the Hadoop Cluster. Proposed system presents a Scenario to overcome this failure by selecting new recovery name node with less amount of time which will replicate the data from Name node on the other Data node so that the availability of the metadata is increases and also Decreases the loss and delay of data.

Keywords: Hadoop, HDFS, Master Name node, Name node, Data node.

1. INTRODUCTION

To achieve high performance scalable data has to be transferred on Technologies that able to process the data efficiently. MySQL database is being used by several applications for managing their data. Our experimental setup is a step to transfer data from MySQL database to one of new technology used for analyzing and managing Big Data.

Everyone says it — we are living in the era of "Big Data." Chances are that you have heard this phrase. In today's technology-fueled world where computing power has significantly increased, electronic devices are more commonplace, accessibility to the Internet has improved, and users have been able to transmit and collect more data than ever before. Organizations are producing data at an astounding rate. It is reported that Facebook alone collects 250 terabytes a day [1]. According to Thompson Reuters News Analytics, digital data production has more than doubled from almost 1 million petabytes (equal to about 1 billion terabytes) in 2009 to a projected 7.9 zettabytes (a zettabyte is equal to 1 million petabytes) in 2015, and an estimated 35 zettabytes in 2020 [1]. Other research organizations offer even higher estimates. As organizations have begun to collect and produce massive amounts of data, they have recognized the advantages of data analysis. But they have also struggled to manage the massive amounts of information that they have. This has led to new challenges. How can you effectively store such a massive quantity of data? How can you effectively process it? How can you analyze your data in an efficient manner? Knowing that data will only increase, how can you build a solution that will scale?

These challenges that come with Big Data are not just for academic researchers and data scientists. In a Google+ conversation a few years ago, noted computer book publisher Tim O'Reilly made a point of quoting Alistair Croll, who said that "companies that have massive amounts of data without massive amounts of clue are going to be displaced by startups that have less data but more clue …" In short, what Croll was saying was that unless your business understands the data it has, it will not be able to compete with businesses that do. Businesses realize that tremendous benefits can be gained in analyzing Big Data related to business competition, situational awareness, productivity, science, and innovation. Because competition is driving the analysis of Big Data, most organizations agree with O'Reilly and Croll. These organizations believe that the survival of today's companies will depend on their capability to store, process, and analyze massive amounts of information, and to master the Big Data challenges.

2. BACKGROUND

To transfer data from MYSQL to Hadoop distributed system and implement an environment of data transfer from these systems the following concepts and tools are required: Hadoop, Hdfs, Map reduce, Hive, Sqoop

Hadoop:

It is also an open source framework for distributed storage and distributed processing of very large data sets on computer clusters.in our experimental setup we installed and used a multi node cluster with one machine as Hadoop Master(server) and two client machines as HadoopSlave1 and HadoopSlave2(client) respectively[2, 3]. 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.

HDFS:

Hadoop Distributed File System (HDFS) is a distributed file system designed for storing very large data sets reliably, and stream those data sets with high bandwidth applications.HDFS stores file system's metadata and application data separately, HDFS stores metadata on a dedicated system, called name node [6, 8]. Application data are stored on other nodes called data-nodes. All nodes in cluster are fully connected and communicate with each other using TCP-based protocols. For reliability the file content is replicated on multiple Data-Nodes. This strategy has an advantage that data transfer bandwidth is multiplied. The Name Node executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to data nodes. The Data Nodes are responsible for serving read and write requests from the clients. The Data Nodes also perform block creation, deletion, and replication as instructed by the name node [7].

Map-Reduce:

Map reduce is a processing technique and a program model for distributed computing based on java. The Map reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map reduce implies, the reduce task is always performed after the map job.

Sqoop:

Apache Sqoop is an open source tool used as intermediate between MySQL database and Hadoop system serves for the purpose of transferring data. Sqoop is a tool designed to transfer data between Hadoop and relational database servers [4]. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation.

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. It is a platform used to develop SQL type scripts to do Map Reduce operations. The Hive Query Language (HQL) is like structured query language for map Reduce to process structured data using Hive [5].

3. IMPLEMENTATION

We implemented a multi-node cluster using apache Hadoop, in multi-node cluster three nodes are considered, as follows one node acting as Name Node or Master Node which we considered as HadoopMaster (server), the remaining two nodes acting as Data Nodes or slave nodes which we considered as HadoopSlave1 (client1) and HadoopSlave2 (client2) respectively, all the three nodes are connected through local area network [3, 7]. After successful installation of Hadoop in Name Node i.e. HadoopMaster, we installed MySQL server, after that we have installed apache Hive, finally installed apache Sqoop to act as interface between HDFS and MySQL and as well as MySQL and Hive. Here first we have to start master or name node then it will show the connected data nodes.

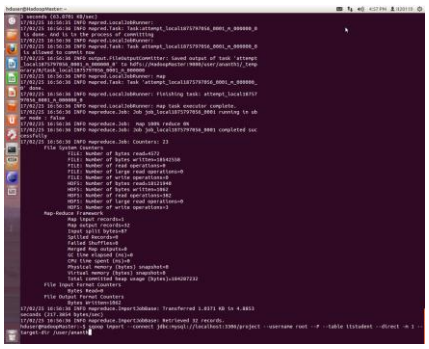

Fig-1 map reduce example-word count

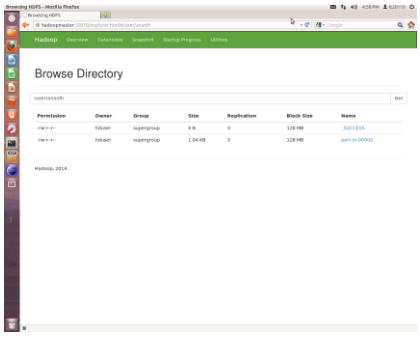

Fig-2 output file in master node

In our sample application we have considered word count example to run this we have started Server and Clients, later we have compiled and created a jar file in our server machine, after which we tried to execute the map reduce program by specifying input and output file, the first fig-1 shows the successful execution of word count example in our server machine, fig-2 shows output file with count of all words in the file at one of the client machines.

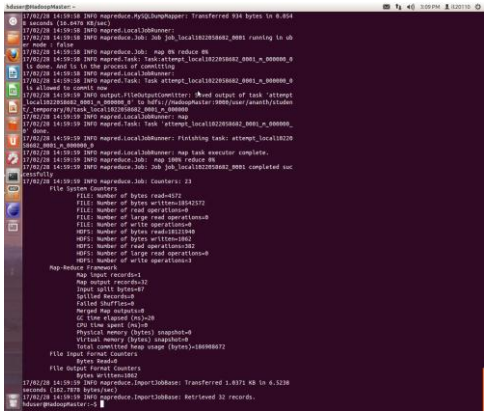

Fig-3 Importing data from MySQL to hadoop

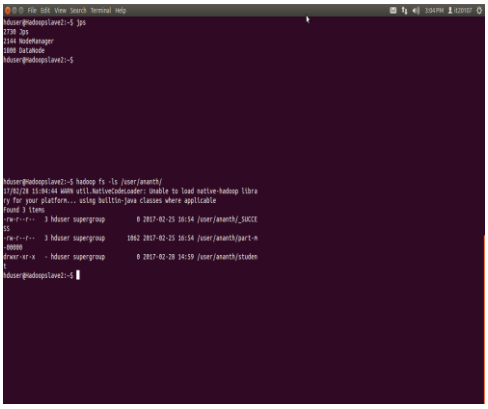

Fig-4 File is imported to hadoop

The above two figures Fig 3 and Fig 4 shows how the importing of data from MySQL to hadoop file system has been done. Later we have started MySQL and created the database, then we have tried to import the same data into hadoop distributed file system using JDBC driver, after successful execution of importing data, data file were loaded into Hadoop Distributed File System and can be accessible at all connected data nodes.

4. CONCLUSION AND FEATURE SCOPE

The system we implemented is Hadoop multi node cluster which gives privilege to access data from name node to the data nodes, by using this we can process huge amount of data with less time, in the same time we can process structured data as well as unstructured and semi structured. We can solve the data mining features like clustering and classification of the data. In feature work we will apply this system to real time problems like weather forecast prediction, customer churn analysis.

REFERENCES

[1] Boris Lublinsky Kevin T. Smith Alexey Yakubovich, Published by John Wiley & Sons, Inc. - 2013, Professional Hadoop® Solutions.

[2] Apache Hadoop single node cluster installation link https://hadoop.apache.org/docs/r2.7.2/

[3] Hadoop multi-node cluster installation link https://www.tutorialspoint.com/hadoop/hadoop_multi_node_cluster.htm

[4] Hadoop sqoop installation link https://www.tutorialspoint.com/sqoop/sqoop _installation.htm

[5] Hadoop hive installation link https://www.tutorialspoint.com/hive

/hive_installation.htm

[6] Lija Mohan, Sudheep Elayidom M., "A Novel Big Data Approach to Classify Bank Customers - Solution by Combining PIG, R and Hadoop", International Journal of Information Technology and Computer Science (IJITCS), Vol.8, No.9, pp.81-90, 2016. DOI: 10.5815/ijitcs.2016.09.10.

[7] https://www.eduonix.com/blog/bigdata-and-hadoop/learn-import-data-mysql-hadoop-using-sqoop/

[8] Robert Chansler, Hairong Kuang, Sanjay Radia, Konstantin Shvachko, and Suresh Srinivas,"Hadoop Distributed File System", http://www.aosabook.org/en/hdfs.html