ISSN No: - 2456 - 2165

Sqooping of data from RDBMS to HDFS and Reporting Data into Hive

Anshul Khare, Ayushi Dixit, Shubham Singh
B.Tech. Student, Bharati Vidyapeeth COEP,
Department of Computer Engineering,
Bharati Vidyapeeth College of Engineering,
Pune, India

Ms. Dipti D.Pawar
Professor, Computer Engineering Department,
Bharati Vidyapeeth College of Engineering,
Pune, India

Abstract:-Sqoop came into picture to provide, an interaction between relational database server and Hadoop's HDFS. In this paper we present the comparison of traditional RDBMS with Map Reduce technique and use of HiveQL to implement Map Reduce on large databases. The paper also focuses on use of Sqoop connector to connect SQL and Hadoop.

Keywords:- *Hadoop , Map-Reduce , Sqoop , Hive.*

I. INTRODUCTION

Big Data is described as volumes of data which could be both structured and unstructured as well. Mainer times, it is so huge that it provides a challenge to process using conventional database sand techniques. Three observations can be inferred; 1. The data is impressive in terms of volumes. 2. It moves at a very rapid pace. 3. It outpaces the prevailing capacity. In the year 2012, there were few terabytes of data in a single dataset, which has been interestingly catapulted to many pet bytes today positively. New platform of manipulating "Big Data" are being commissioned. 5 exabytesi.e 1 Exabyte = 1.1529*1018 bytes today this amount of information, is created in two days [8, 16]. It is predicted to double every two year. With an increase in the data, there is a comparable increase in the applications and framework to administer it. This gives boost to new vulnerabilities that need being responded. Hdfshadoop has a distributed file system (H.D.F.S) and a layer that handles parallel computation and rate of flow of workflow and configuration administrations.

II. LITERATURE REVIEW

Ministry departments are replying with an action report on these ideas and policy proposal. The biggest issue for Government department is how to be admissible? If all citizens are treated with dignity and invited to team up, it could be an easier administration [2]

Big Data has attracted many Government sectors. The P.M.O is using different Big Data techniques to process ideas given by citizens on its crowd sourcing platform to generate actionable reports for various departments for consideration and implementation.

The fundamental of Hadoop is to process data in a distributed file system manner. So, if a single file is found it splits into blocks and the blocks are spread into the cluster nodes. Hadoop applications require enormous available distributed file systems with unlimited capacity. Hadoop is issued in Facebook and Yahoo for batch processing purposes.

III. NEED OF HADOOP

Hadoop has the upper hand in scalability, fault tolerance, effortless programming and most important is the storage method of Hadoop which is very suitable for the future projects [4].

The necessity can be analyzed by considering a simple example of development of an application for Global Positioning. Such application contains distributed and heterogeneous data sources including GIS maps, satellite images and temperature measurement stations.

Hadoop provides some tools which helps us in managing large data set easily.

Some of the Tools used in hadoop is as follows:

A. Hbase

Hbase is database management system which is column oriented and runs on top of the HDFS. HBase does not support a structured query language like SQL

B. Hive

Hive provides a structure to project structure onto data and query the data using a SQL-like language called HiveQL. this language can also allow traditional map/reduce programmers to plug in their custom mappers and reducers when it is difficult or ineffective to express this logic in HiveQL.

C. Sqoop

Sqoop is a tool provided to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in HadoopMapReduce, and then export the data back into an RDBMS.

D. PIG

Pig helps analyzing large data sets that contains high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. Pig infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist. Our main focus is Sqoop tool in this project. Henceforth we can combine two different platform Structured and Unstructured

IV. ARCHITECTURE OF HDFS

There are only 4 building blocks inside Environment Setup. Starting from top to bottom.

A. Mapreduce Framework

Map reduce contains two important task namely, map and reduce. Map modifies set of data by converting it into other set of data where all the individual elements are broken into tuples. reduce task will take the output from a map as an input and combines those data tuples into a smaller set of tuples.

B. Yarn Infrastructure

YARN takes the resource management capabilities that were in MapReduce and packages them so they can be used by new engines. This also streamlines MapReduce to do what it does best, process data YARN stands for "Yet Another Resource Negotiator" and was added as a part of Hadoop 2.0.. With YARN, you cannow run multiple applications in Hadoop.

C. HDFS Federation

HDFS Federation is the framework responsible for providing permanent, reliable and distributed storage. This is typically used for storing inputs and output.It mainly providestask management, cluster management,user management and data management.

D. Cluster

Cluster is the set of host machines known as nodes. Nodes can be partitioned into racks. This is the hardware part of the infrastructure.

V. COMPARISON BETWEEN MAP REDUCE AND TRADITIONAL RDBMS

Table 1: Traditional RDBMS compared with MapReduce

	Traditional RDBMS	***MapReduce***
Data Size	Gigabytes size	Petabytes size
Access	Interactive as well as batch	Batch
Updates	Many times Read and write	Read many times but Write once
Structure	Schema is Static	Schema is Dynamic
Scaling	Nonlinear	Linear
Integrity	High	Low

VI. SQOOPING OF DATA FROM RDBMS TO HDFS

RDBMS is a perfect fit for scale in processing structured data and static data sets.***Sqoop*** is a command-line interface application for transferring data between relational databases and Hadoop.

Using Sqoop connector and the Hadoop's Map-Reduce framework, structured data is processed using Hive programming. So that user specific analysis can be carried out easily. The storage of data among the Hadoop cluster would help in easy access and availability of data. Sqoop uses the primary key column to determine how to divide the source data across its mappers. Sqoop also has significant integration with Hive, allowing it to import data from a relational source into either new or existing Hive tables.

Here I have considered a student database with table name as student info. Steps involved in importing of data from RDBMS(MySQL)to HDFS and also reporting it into HIVE.

- [root@master Desktop]# start Mysql
- [root@master Desktop]# sudo service mysqld start
- [root@master Desktop]# ifconfig
- [root@master Desktop]# vi /etc/hosts
- [root@master Desktop]# jps
- [root@master Desktop]# sudo service hadoop-hdfs-namenode start
- [root@master Desktop]# sudo service hadoop-hdfs-datanode start
- [root@master Desktop]# sudo service hive-serve2 start
- [root@master Desktop]# sudo service hive-metastore start
- [root@master Desktop]# sqoop import --connect jdbc:mysql://master/studentinfo --username root --m 1 --table studentFees --create-hive-table --hive-import --target-dir /user/a

VII. HIVEQL TO RUN MAPREDUCE JOBS

Following table shows comparison of SQL and HiveQL considering operations they perform.

Table 2: SQL vs Hive Query Language

Sr. No.	Operations & Functions	SQL	Hive Query Language
1.	Select	SQL-92 supports it.	Single table or view in the FROM clause. For partial ordering SORT BY is used. To limit number of rows returned LIMIT operations is used. HAVING clause is not supported.

2.	Updates	UPDATE,
		INSERT,
		DELETE
			INSERT
			OVERWRITE
			TABLE(It populates complete table or partition)

3.	Data types present	Integral, floating point, fixed point, text and binary strings, temporal	Integral, floating point, boolean, string, array, map, struct
4.	Default Join Types	Inner Join	Equi Join
5.	Built-In Functions	Built-in functions are in Hundreds.	Dozens of built-In Functions present.
6.	Multiple table inserts	Not supported in SQL	Supported in HiveQL
7.	Create table as select	Not valid in SQL but may be found in some databases	Supported by HiveQL
8.	Extension points	User-defined functions and Stored procedures.	User-defined functions and Map-Reduce scripts.

A programmer happy with SQL language will prefer to express data operations with SQL language only. Hive is Hadoop's data warehouse system that provides a platform to project structure onto data stored in HDFS or a compatible file system [8].

SQL is an informative programming language and not ancompulsory programming language like C, for accessing and manipulating database systems.

Hive enables users to plug in conventional map-reduce jobs into queries. The language includes a type system with support for tables containing primitive types, collections like maps, and nested compositions of maps.

A compiler translates HiveQL statements into a directed acyclic graph of MapReduce (MR) jobs, which are sent to Hadoop for execution purposes. [9] Hive maintains metadata in a metastore, which is stored in a relational database, as well as this metadata contains information about tables.

Hive is one of the easiest to use high-level MapReduce (MR) frameworks. Particular strength of HiveQL is in offering ad-hoc querying of data, in contrast to the compilation requirement of Pig. Hive is an initial point forfull featured BI (business intelligence) systems which offer a user friendly interface for common users.

VIII. CONCLUSION

Hadoop is one of the best distributed massive data processing framework and it has high performance on distributed computing and distributed storage. Connection between RDBMS and HDFS is possible by SQOOP. Using Hadoop the massive data storage and processing by MapReduce in a big organization proposes the use of Hive architecture and its components that have advantage over traditional methods of data processing. Also Hive can be useful for treatment of small files in the system in near future.

References

[1] Rajagopalan M.R and Solaimuruganvellaipandiyan "*Big data framework for national E-governance plan*" ICT and Knowledge Engineering (ICT&KE), 2013, 11th InternationalConferenceon DOI:10.1109/ICTKE.2013.67 56283 Publication Year: 2013, Page(s): 1–5 IEEE CONFERENCE PUBLICATIONS

[2] http://analyticsindiamag.com/pmo-using-big-datatechniques-mygov-translate-popular-mood-governmentaction/

[3] A. Thusoo, Z. Shao, S. Anthony *et al.*, "Data warehousing and analytics infrastructure at facebook," in SIGMOD 2010, International conference on Management of data, pp. 1013-1020.

[4] Fu Chang-Jian, LengZhihua. A Framework for Recommender Systemsin E-Commerce Based on Distributed Storage and Data-Mining. International Conference on E-Business and E-Government (lCEE20 I 0), Volume I, pp.3502 - 3505.

[5] D. Borthakur, J. Gray, J. S. Sarma*et al.*, "Apache hadoop goes realtime at Facebook," in ACM SIGMOD International Conference on Management of Data, 2011, pp. 1071-1080.

[6] Tom White (2013). Inkling for Web [Online]. Available: https://www.inkling.com/read/hadoop-definitive-guide- tom-white-3rd/chapter-1/comparison-with-other-systems.

[7] Katal, A.; Wazid, M.; Goudar, R.H., "Big data: Issues, challenges, tools and Good practices," Contemporary Computing (IC3), 2013 Sixth International Conference on , vol., no., pp.404,409, 8-10 Aug. 2013

[8] "Welcome to Hive!" December, 2012; http://hive.apache.org/

[9] Ivan Tomašić, Aleksandra Rashkovska, MatjažDepolli and Roman Trobec "A Comparison of Hadoop Tools for Analyzing Tabular Data**",** Jožef Stefan Institute, Slovenia, Informatica 37(2013) 131–138

[10] Thusoo, A.;Sarma, J.S.;Jain, N.;Zheng Shao;Chakka, P.;Ning Zhang;Antony, S.;Hao Liu;Murthy, R."Hive - PetabytescaledatawarehouseusingHadoop" in Data Engineering (ICDE), 2010 IEEE 26th International conference on DOI: 10.1109/ICDE.2010.5447738 Publication Year:2010, Page(s):996-1005