Impact Factor Value 4.046 **e-ISSN: 2456-3463**
International Journal of Innovations in Engineering and Science, Vol 4, No. 10, 2019
National Level Technical Paper Presentation- PHOENIX-19
Organized by Godavari College of Engineering, Jalgaon - 425003
www.ijies.net

Hive: A Literature Review

Kavita D. Mahajan [1], Vijay D. Chaudhari [2]

[1] *Asst. Prof. MIT ACS College, Dept. of Computer Applications, Alandi – Pune, Maharashtra, India*
[2] *Asstt. Prof. & M.Tech.(VLSI & Embedded System Design) course coordinator, GF'S Godavari COE , Jalgaon-425003.*

Abstract – *The size of data sets being collected and analyzed in the industry for business intelligence is growing rapidly, making traditional warehousing solutions prohibitively expensive. Hadoop is a popular open-source map-reduce implementation which is being used in companies like Yahoo, Facebook etc. to store and process extremely large data sets on commodity hardware. However, the map-reduce programming model is very low level and requires developers to write custom programs which are hard to maintain and reuse. In this paper, we present Hive, an open-source data warehousing solution built on top of Hadoop. Hive architecture, working, advantages and disadvantages. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.*

Keywords*-Hive, HiveQL, Hadoop, map-reduce, HDFS, HBASE.*

1. INTRODUCTION

1.1 Prerequisites:

Before proceeding with Hive, you need a basic knowledge of Core Java, Database concepts of SQL, Hadoop File system, and any of Linux operating system flavors.

Hadoop:

Hadoop is an open-source framework to store and process Big Data in a distributed environment. It contains two modules, one is MapReduce and another is Hadoop Distributed File System (HDFS).

i. **MapReduce:** It is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.
ii. **HDFS:** Hadoop Distributed File System is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware.

The Hadoop ecosystem contains different sub-projects (tools) such as Sqoop, Pig, and Hive that are used to help Hadoop modules.

i. **Sqoop:** It is used to import and export data to and from between HDFS and RDBMS.
ii. **Pig:** It is a procedural language platform used to develop a script for MapReduce operations.
iii. **Hive:** It is a platform used to develop SQL type scripts to do MapReduce operations.

1.2 What is Hive?

i. 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. Initially Hive was developed by Facebook, 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 MapReduce.

1.3 Hive is not:

i. A relational database
iii. A design for On-line Transaction Processing (OLTP)
iv. A language for real-time queries and row-level updates

2. LITERATURE SURVEY

To handle the data is increasing data in volume, variety and velocity , we have to use databases with massively parallel software running on tens, hundreds, or more than thousands of servers. So Big data platforms are

used to acquire, organize and analyze these types of data. In this paper [1], first of all, authors will acquire data from social media using Flume. Flume can take log files as source and after collecting data, it can store it directly to file system like HDFS or GFS. Then, organize this data by using different distributed file system such as Google file system or Hadoop file system. At last, data will be analyzed using mapreducers in Pig, Hive and Jaql. Components like Pig, Hive and Jaql do the analysis on data so that it can be access faster and easily, and query responses also become faster.

Big data analytics is the process of examining large amounts of data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable [2]. Two most important technologies mapreduce and hive, are for handling big data for solving the problems in hand and to deal the massive data. The size of data sets being collected and analyzed in the industry for business intelligence is growing rapidly, making traditional warehousing solutions prohibitively expensive. Hadoop is a popular open-source map-reduce implementation which is being used in companies like Yahoo, Facebook etc. to store and process extremely large data sets on commodity hardware. However, the map-reduce programming model is very low level and requires developers to write custom programs which are hard to maintain and reuse. Hive, an open-source data warehousing solution built on top of Hadoop. Hive supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into mapreduce jobs that are executed using Hadoop. In addition, HiveQL enables users to plug in custom map-reduce scripts into queries.

The human beings are now living in an enormously developed, technical era, where internet is becoming fundamental need of all individual. Today, our social, personal as well as professional life are revolving around world wide web. Thus, giving birth to Big Data at an incredible momentum. Traditional management tools and frameworks are proved, un-fair while dealing with Big Data. The paper [3] emphasize on the challenges increasing from the use of big data. Authors tried to uncover correct approach to retrieve valuable information from the pile of Big Data.

Big data came into existence when the traditional relational database systems were not able to handle the unstructured data (weblogs, videos, photos, social updates, human behaviour) generated today by organisation, social media, or from any other data generating source. Data that is so large in volume, so diverse in variety or moving with such velocity is called Big data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC, Overflow. These technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB. In this research paper [4], authors has mentioned various technologies for handling big data along with the pros and cons of each technology for catering the problems in hand to deal the massive data.

3. HIVE Architecture

3.1 Architecture of Hive

Figure 1 below shows the component diagram of Hive architecture.

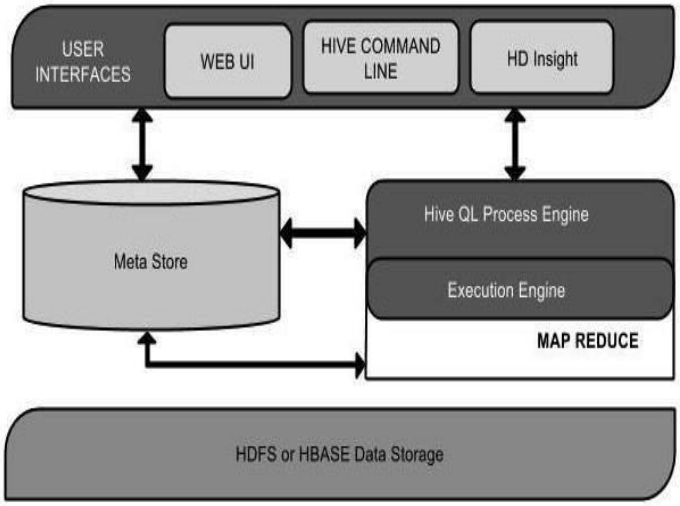

Figure 1. Hive architecture

And Table 1 below shows the operation of each component [7]-[14].

Table 1: Operation of each unit in Hive

Unit Name	Operation
User Interface	Hive is a data warehouse infrastructure software that can create interaction between user and HDFS. The user interfaces that

	Hive supports are Hive Web UI, Hive command line, and Hive HD Insight (In Windows server).
Meta Store	Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping.
HiveQL Process Engine	HiveQL is similar to SQL for querying on schema info on the Metastore. It is one of the replacements of traditional approach for MapReduce program Instead of writing MapReduce program in Java, we can write a query for MapReduce job and process it.
Execution Engine	The conjunction part of HiveQL process Engine and MapReduce is Hive Execution Engine. Execution engine processes the query and generates results as same as MapReduce results. It uses the flavor of MapReduce.
HDFS or HBASE	Hadoop distributed file system or HBASE are the data storage techniques to store data into file system.

3.2 Interaction of Hive with Hadoop framework

Figure 2 shows the workflow between Hive & Hadoop. All Hadoop sub-projects such as Hive, Pig, and HBase support Linux operating system. Therefore, you need to install any Linux flavored OS [5]-[14].

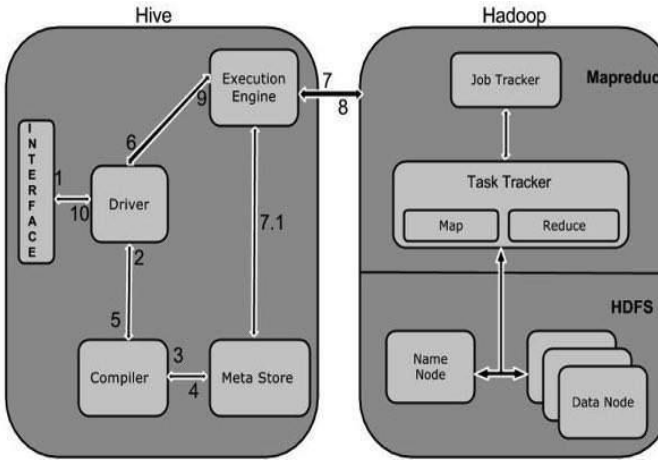

Figure 2. Workflow between Hive & Hadoop

Table 2 below shows how Hadoop interacts with Hadoop framework:

Step No.	Operation
1	**Execute Query** The Hive interface such as Command Line or Web UI sends query to Driver (any database driver such as JDBC, ODBC, etc.) to execute.
2	**Get Plan** The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query.
3	**Get Metadata** The compiler sends metadata request to Metastore (any database).
4	**Send Metadata** Metastore sends metadata as a response to the compiler.
5	**Send Plan** The compiler checks the requirement and resends the plan to the driver. Up to here, the parsing and compiling of a query is complete.
6	**Execute Plan** The driver sends the execute plan to the execution engine.
7	**Execute Job** Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Here, the query executes MapReduce job.
7.1	**Metadata Ops** Meanwhile in execution, the execution engine can execute metadata operations with Metastore.
8	**Fetch Result** The execution engine receives the results from Data nodes.
9	**Send Results** The execution engine sends those resultant values to the driver.

10	**Send Results** The driver sends the results to Hive Interfaces.

4. CONCLUSION

From this comprehensive literature review, we found that Hive is a higher level query language that simplifies working with large amounts of data. Also it has lower learning curve than Pig or MapReduce, the HiveQL is much closer to SQL than pig and Less trial and error than pig. Inspite of these advance features Hive has some setbacks too like u pdating data is complicated mainly because of using HDFS, it can add records and overwrite partitions too. Hive can access real time access to data and need to use other means like Hbase or Impala. Hive has high latency too.

REFERENCES

[1] *Munesh Kataria, Ms. Pooja Mittal, "Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql", International Journal of Computer Science and Mobile Computing (ISSN: 2320-088X) , Vol.3 Issue.7, July- 2014, pg. 759-765.*

[2] *Tripti Mehta, Neha Mangla, "A Survey Paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework", International Journal of Recent Advances in Engineering & Technology (IJRAET)", (ISSN: 2347-2812) , Vol.4 Issue.2, Feb-2016, pg. 112-118.*

[3] *Prity Vijay, Bright Keshwani, " Emergence of Big Data with Hadoop: A Review", International organization of Scientific Research", (ISSN: 2250-3021) , Vol.6 Issue32, Mar-2016, pg. 50-54.*

[4] *A. Antony Prakash, Dr. A. Aloysius, " Architecture Design for Hadoop No-SQL and Hive", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)", (ISSN: 2456-3307) , Vol.3 Issue 1, Feb-2018, pg. 1069-1077.*

[5] *R. Chaiken, et. al. Scope: Easy and Efficient Parallel Processing of Massive Data Sets. In Proc. of VLDB, 2008.*

[6] A. Pavlo et. al. A Comparison of Approaches to Large-Scale Data Analysis. In Proc. of ACM SIGMOD, 2009.

[7] *Apache Hadoop. Available at http://wiki.apache.org/hadoop*

[8] Hadoop Map-Reduce Tutorial at http://hadoop.apache.org/common/docs/current/mapred_tutoral.html.

[9] *Hadoop HDFS User Guide at http://hadoop.apache.org/common/docs/current/hdfs_user_guide.html.*

[10] *Mysql list partitioning at http://dev.mysql.com/doc/refman/5.1/en/partitioning-list.html.*

[11] *Hive Performance Benchmark. Available at http://issues.apache.org/jira/browse/HIVE-396 .*

[12] *Running TPC-H queries on Hive. Available at http://issues.apache.org/jira/browse/HIVE-600*

[13] *Hive wiki at http://www.apache.org/hadoop/hive.*

[14] *DataNucleus .Available at http://www.datanucleus.org.*