SACRAMENTO STATE

CSC 244

SPRING 2019

A Project Report on

"APACHE PIG AND HIVE"

Instructor Name:

Dr. Ying Jin

Submitted By:

Navya Alapati

Pujitha Reddy Koppurapu

NAME	CONTRIBUTION
Navya Alapati	50%
Pujitha Reddy Koppurapu	50%

TABLE OF CONTENTS

ABSTRACT

Big Data is used for huge datasets which is difficult to be processed using traditional devices. This platform is used to organize, analyze and acquire huge amount of data running on servers. One of the best option is Apache Hadoop, which is modeled to scale up from single server to thousands of devices each having storage and local computation.

Apache Pig and Apache Hive are core components of Hadoop ecosystem that provides specification and higher latency. They run on MapReduce algorithm which process bigdata sets on a cluster.

Apache Hive data warehouse software facilitates writing, reading, and managing large volumes of datasets residing in distributed environment using SQL. Apache hive provides the SQL-like language called HiveQL, which transparently convert queries to MapReduce.

Apache Pig is a tool or a platform for evaluating massive sets of data representing them as data flows. It uses simple SQL scripting language, which is known as PIG LATIN, which is easy to code, manage and optimize the execution.

In this paper, basic concepts, architecture and tools for implementing map reduce programs using Apache pig and Apache hive are introduced. The comparative study of pig and hive coding techniques and an example of map reduce job using hive and pig will be described and also discuss which of these coding techniques are used in different business development types.

From:

Intellipaat: https://intellipaat.com/tutorial/big-data-and-hadoop-tutorial/introduction-to-pig-sqoop-and-hive/

Educbu: https://www.educba.com/apache-pig-vs-apache-hive/

1. INTRODUCTION

The normal database systems are not able to manage large and unstructured data which are in the form of video formats. YouTube has over millions of users and generate billions of views. As its data is getting created in huge amount with great speed it has become a challenge to store, process and study this large amount of data to make it usable. Hadoop is a framework that was created as an alternative to conventional database systems to process large number of datasets, analyze, store and retrieve information by the help of simple programming model. Hadoop Ecosystem comprises of MapReduce Framework, Yarn, Hive, Pig and many more components. MapReduce Framework of Hadoop uses Java for implementing Map and Reduce procedure which is used to process the data in cluster.

The story of Apache pig begins in the year 2006 in Yahoo Research when the researcher was struggling with Map Reduce Java code. It is a platform for managing large sets of data which consists of high-level programming to analyze the data. The story of Apache Hive begins in the year 2007 by the researchers working at Facebook, when non-Java programmers have to struggle while using Hadoop MapReduce. It is a data warehouse software that lets you read, write and manage huge volumes of datasets that is stored in a distributed environment using SQL. By using Apache Pig and Hive developers need not write thousand lines of code to implement Map and Reduce procedure.

2. LITERATURE SURVEY

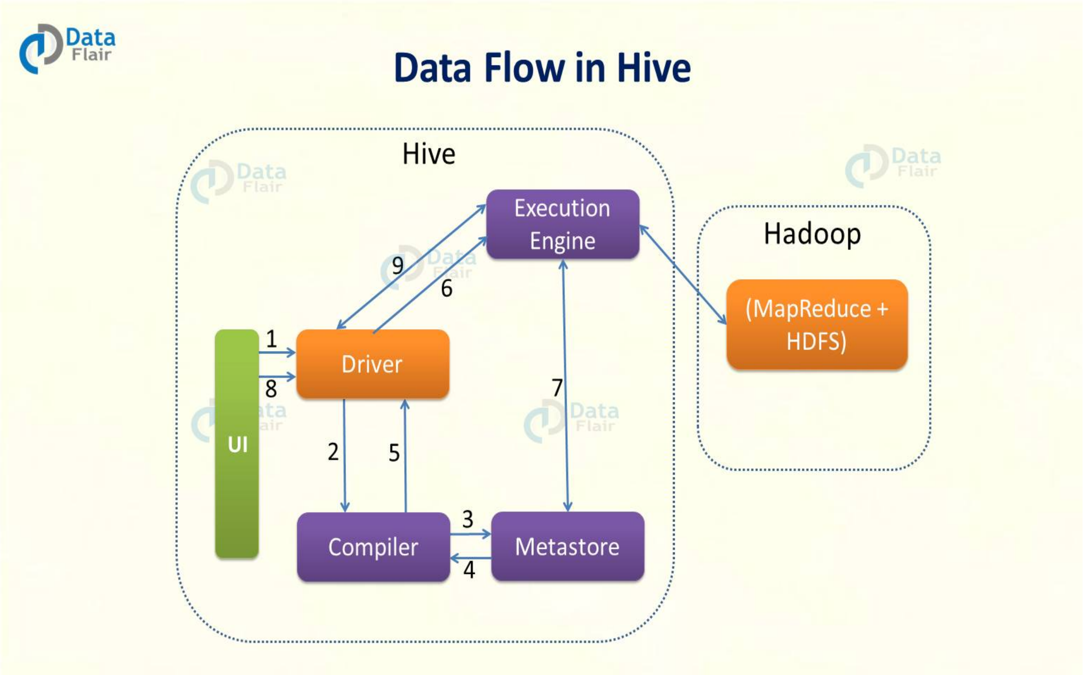
Apache PIG is invented primarily for analyzing large data sets that has high-level programming language. It consists of key properties such as simple to program to achieve parallel programming execution and complex data tasks, Optimization techniques are used to encode the system to optimize the execution automatically to allow user o focus on semantics instead of efficiency, Extensibility for users to create their own functions and for processing. Apache HIVE also supports analysis of huge datasets stored in Hadoop's HDFS and file system compatible. It also provides SQL-like query language known as HIVEQL with schema on read and transparently converts queries to MapReduce. It is also used to accelerate queries, provides indexes and bitmap indexes.

In this report, we cover introduction to the MapReduce and how it helps in Apache PIG and HIVE [1]. Then we look into the tools used and Architecture of Apache Pig and Hive where we look into the components of HIVE architecture and PIG architecture. We then move on to the requirements necessary for installation of both PIG and HIVE [3]. We later on look into the sample examples of MapReduce programming using PIG in Latin Script and Sample example of MapReduce using Apache Hive in QL script. Next, we learn list of some companies that widely use Apache PIG and HIVE in their daily work and researches done on it [4]. Then we see the comparative analysis of both Apache PIG and Apache HIVE, their differences and similarities they have in common [6]. Finally, we look into the final conclusion or gist of the total report on Apache PIG and HIVE [7].

3. TOOLS AND ARCHITECTURE OF APACHE PIG AND HIVE

Apache Pig is convenient tools developed by Yahoo which is mainly used for the analysis of Huge datasets whereas Hive is developed by Facebook and is a data warehouse which provides a simple language known as HiveQL similar to SQL querying and analyzing.

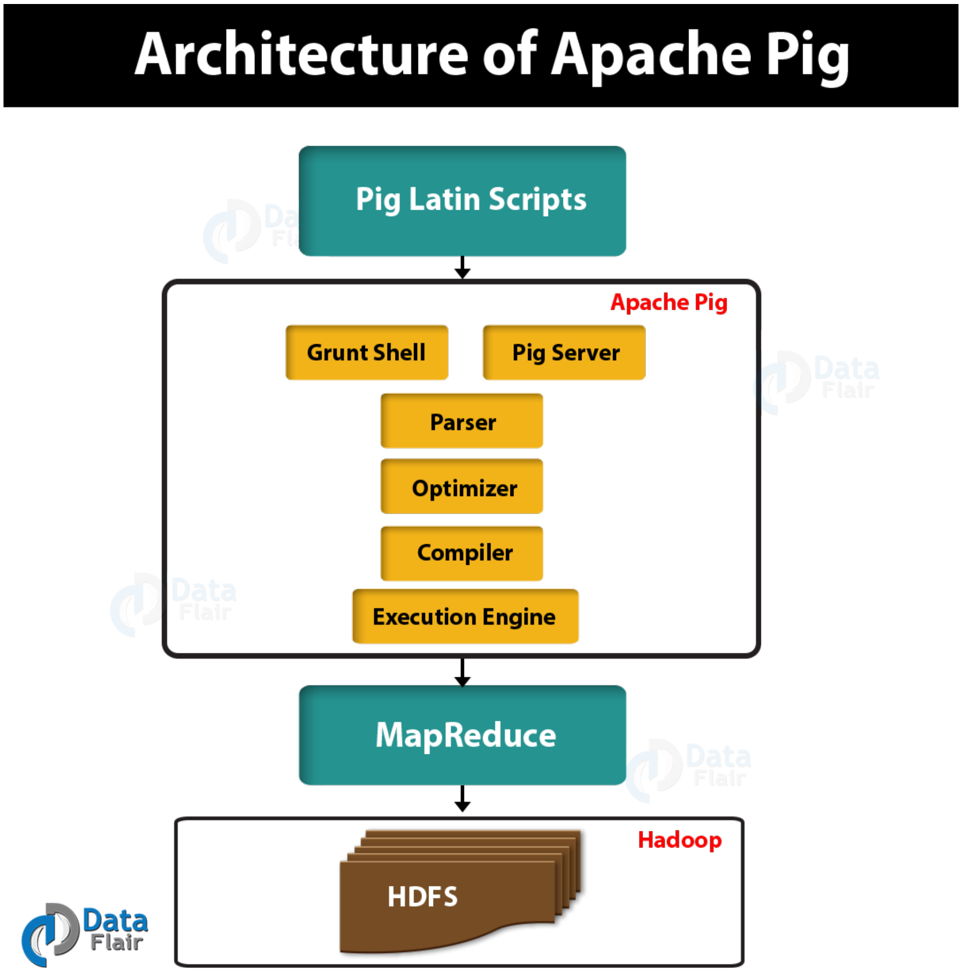
3.1 Hive Architecture Components



- **User Interface:** The interface that Hive supports are Hive Web UI, Hive command line, and Hive HD Insight. The Interface for clients where queries and different operations are submitted to the system.
- **Driver:** The Driver part will receive the queries next.
- **Compiler:** It performs analysis on the different query blocks, query expressions and query parsing generates an execution plan with the help of table and the metadata is partitioned which is looked up from the metastore.

- **MetaStore:** Data is stored related to structure of different partitions and tables in the distribution center with the data related to columns and type of columns. Hive chooses repetitive database servers to store the schema or MetaData of tables, databases, columns in a table and their datatypes.
- **Execution Engine:** Execution Engine process the query and generates the results. The result is same as MapReduce results.
- **HDFS:** Hadoop Distributed File System are the data storage techniques to store data in file system.

3.2 Pig Architecture Components



- **Parser:** All the Pig Latin Scripts are handled by Parser. Parser basically checks the syntax of the script, does type checking and after miscellaneous checks. Directed acyclic graph (DAG) is the yield of parser, that represents pig Latin Statements and logical operators.
- **Optimizer:** DAG is passed to the logical optimizer. It carries out the optimization future such and push down and projection.
- **Compiler:** The optimized logical plan is complied by the compiler into a series of MapReduce jobs.
- **Execution Engine:** All the jobs are submitted to Hadoop n a sorted order. The required result is provided ones these MapReduce jobs are executed on Hadoop.

3.3 Requirements for installation of Pig

It is important that we have Hadoop and Java installed on our local machine before we go for Apache Pig, which means before installing Apache Pig, install Hadoop and Java. Java 1.6, Hadoop 0.2 is required for Pig versions 0.5 through 0.9 to connect to Hadoop Cluster.

Pig can also be downloaded from Apache's Maven repository. It has JAR files for Pig, source code of Pig and for Javadocs.

Pig can be downloaded from Cloudera where all the tools are wrapped and tested together and it can also provide Professional support, if needed.

3.4 Requirements for installation of Hive

Similar to pig, Java and Hadoop must be installed on our system before installing Hive.

Java 1.7

Hadoop 2.x

Any binary of Hive from the source available on the internet can be downloaded and installed.

4. MAPREDUCE EXAMPLE

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs filtering and sorting, and a reduce method, which performs a summary operation.

4.1 MapReduce Program Example Using Pig

Example:

```
A = LOAD ‘data’ USING PigStorage() AS ( f1:int, f2:int, f3:int );

B = GROUP A BY f1;

C = FOR EACH B GENERATE COUNT ($0);

DUMP C;
```

Keywords:

LOAD= Loads data from the file system.

GROUP= Groups the data in one or more relations.

FOREACH= Generates data transformations based on columns of data.

4.2 MapReduce Program Example Using Hive

Example:

```
CREATE TABLE page_view(viewTime INT, userid BIGINT,

            page_url STRING, referrer_url STRING,
        ip STRING COMMENT 'IP Address of the User')
COMMENT 'This is the page view table'
PARTITIONED BY(dt STRING, country STRING)
STORED AS SEQUENCEFILE;
```

Keywords:

CREATE TABLE: Create is to create a table.

COMMENT: Comment can be attached both at column level and table level.

PARTITIONED BY: partitioned by clause defines the partitioning columns which are different from the data columns and are actually not stored with the data.

STORED: How the file is stored.

4.3 Companies researches using Pig and Hive

There are wide range of companies that are using PIG and Hive. For example, AOL company uses PIG for analytics and batch data processing for various applications. PIG is also used in Mendeley to help their business analytics, feature feedback, user experience evaluation and more. Salesforce also uses PIG to develop easy custom UDFs. They developed their own library containing UDFs and loaders and are actively contributing back to the community. Twitter uses PIG extensively to process usage logs, mine tweet data, and more. Stanford University also uses PIG to power an emerging interface to these archives for social scientists. Pig and Hadoop are used for the underlying processing and indexing. The goal is to support these scientists in analyzing the archives for their research. PIG is widely used by PayPal to analyze transaction data in order to prevent fraud. They are the main contributors to the Pig-Eclipse project.

APACHE HIVE is also very popular in usage of many companies. HIVE is used for Trending topics, Hot Wikipedia Topics, Served Fresh Daily. Powered by Cloudera Hadoop Distribution & Hive on EC2. They use Hive for log data normalization and building sample datasets for trend detection R&D. In NexR, HIVE is used for replacing Oracle DW, big data analysis and integrating R. They develop the enterprise Hive. SaaSPulse uses HIVE for analytics, machine learning and customer interaction analysis of web applications. TaoBao uses HIVE for data mining, internal log analysis and ad-hoc queries. We also do some extensively developing work on Hive.

5. COMPARATIVE ANALYSIS OF PIG AND HIVE

- **Data Processing:** Apache PIG is high level data flow language. Apache HIVE is used for batch processing that is Online Analytical Processing(OLAP).
- **Processing Speed:** Apache PIG has higher latency because of executing MapReduce job in background. Apache HIVE also has higher latency because of executing MapReduce job in background.
- **Compatibility with Hadoop:** Apache PIG runs on top of MapReduce. Apache HIVE also runs on top of MapReduce.
- **Definition:** Apache PIG is open source, high level data flow system that renders you a simple language platform properly known as PIG Latin that can be used for manipulating data and queries. Apache HIVE is open source and similar to SQL used for Analytical Queries.
- **Language Used:** Apache PIG uses procedural data flow language called PIG Latin. Apache HIVE uses a declarative language called HiveQL.
- **Schema:** Apache PIG don’t have concept of schema. You can store data in allas. Apache HIVE supports schema for inserting data in tables.
- **Web Interface:** Apache PIG does not support web interface. Apache HIVE supports web interface.

- **Operations:** Apache PIG is used for structured and semi-structured data. Apache Hive is used for structured data.
- **User Specification:** Apache PIG is used by researchers and programmers. Apache HIVE is used by Data Analyst.
- **Operates On:** Apache PIG operates on client side of cluster. Apache Hive operates on server side of cluster.
- **Partition Methods:** There is no concept of partition in Apache PIG. Apache HIVE supports sharding features.
- **File Format:** Apache PIG supports Avro file format. Apache HIVE directly does not support Avro format but can support using.

6. CONCLUSION

Hadoop MapReduce is at present a well-known idea for expansive scale information examination performance. PIG and HIVE utilization in Bigdata examination reveals view into critical issues looked by purchasers and enables the establishments or partnerships to redress these issues, change in administrations, to give appropriate fulfillment to the buyers, to keep beware of issues and to develop cooperative attitude in the market. Then also, it gives buyers to recognize legitimately among the organizations and make the specialist co-op determination overwhelmingly. Based on the parameters like execution time, number of MapReduce jobs, lines of code it has been examined that hive holds better and efficient than pig. In light of the parameters like execution time, number of MapReduce occupations, lines of code it has been inspected that hive holds preferable and effective over pig.

BIBLIOGRAPHY

[1] Comparative Analysis Using Hive and Pig on Consumers Data: Pooja Jain et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (2), 2017, 285-291.

[2] Analysis of Pig and Hive: International Journal of Research in Advent Technology, Vol.6, No.5, May 2018 E-ISSN: 2321-9637

[3] https://cwiki.apache.org/confluence/display/Hive/Tutorial

[4] https://pig.apache.org/docs/latest/basic.html

[5] https://cwiki.apache.org/confluence/display/Hive/PoweredBy

[6] https://cwiki.apache.org/confluence/display/PIG/PoweredBy

[7] https://www.educba.com/apache-pig-vs-apache-hive/

[8] https://en.wikipedia.org/wiki/Apache_Hive

[9] https://pig.apache.org

[10] https://www.dezyre.com/article/hadoop-ecosystem-components-and-its-architecture/114

[11] https://data-flair.training/blogs/pig-architecture/

[12] https://www.tutorialspoint.com/hive/hive_introduction.htm