International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195250

An Overview of Apache Pig and Apache Hive

Saiyam Arora*[1], Abinesh Verma[2], Richa Vasuja[3]
[1,2]Department of Computer Science, Chandigarh University, Mohali, Punjab, India
[3]Assistant Professor Department of Computer Science, Chandigarh University, Mohali, Punjab, India

ABSTRACT

Ever since the enhancement of technology has taken place, the data is growing at an alarming rate. The most prominent factor of data growth is the "Social Media", leads to the origination of a tremendous amount of data called Big Data. Big Data is a term used for data sets that are extremely large in size as well as complicated to store and process using traditional database processing applications. A saviour to deal with Big Data is "Hadoop" and two major components of Hadoop which are HDFS (Distributed Storage) and Map Reduce(Parallel Processing). Apache Pig and Hive is an essential part of the Hadoop Ecosystem. This paper covers an overview of both Apache Pig and Hive with their architecture. As Hadoop, no doubt is doing tremendously great work by storing and processing the huge volume of data but there are more frameworks now a days to increase the efficiency of Hadoop framework which are basically seen as the layers of Hadoop or a part of Apache Hadoop project. And that is why this paper includes the two most important layers namely Apache Pig and Apache Hive.

Keywords: Big Data, Hadoop, Map Reduce, HDFS, Pig, Hive.

I. INTRODUCTION

Nowadays Technology leads to the origination of a tremendous amount of data. To handle this data there are lots of challenges that needs to be taken care which includes capturing, curating, storing, searching, transferring, analysing and visualising of data. To overcome these challenges the biggest support is Hadoop framework, which is an open-source, Java-based framework that supports storage and processing of extremely large data sets in a distributed computing environment. It is reliable, flexible, scalable and economical(works on commodity hardware). The two core components of Hadoop are HDFS and Map Reduce. HDFS is used for storing the data in a distributed manner and Map Reduce is for processing the huge amount of data using parallel approach[5]. Map Reduce is a java based framework, which makes it possible to process a large set of data in a parallel way. But programmers who are not good at Java normally used to struggle to work with Hadoop, especially while performing any MapReduce tasks. Even if the lines of code is compared then also it is much more lengthy. Java codes are usually too long to write. So there is a need to get a platform where this much lines of code should not be there. For this purpose also, Apache Pig and Hive as introduced. Apache Pig is an abstraction over Map Reduce. Basically Apache Pig is a layer on Hadoop framework only. This Pig supports parallelisation mechanism. For implementation purpose, it provides Pig Latin language which is really a boon and 10 lines of Pig Latin is equal to 200 Lines of java code[1][6]. So this one is the foremost thing in Apache Pig that a different language is used so that the developer should not struggle more into it.

The other framework of Hadoop is Apache Hive, which works on a query language known as HQL(HiveQL).

II. OVERVIEW OF APACHE PIG

A. Apache Pig

Apache Pig was introduced by Yahoo in 2006 as a research project, basically to create and executes Map Reduce tasks on large data sets. In 2007, Apache Pig was open sourced via Apache incubator. In 2008, the first release of Apache Pig came out. In 2010, Apache Pig graduated as an Apache top-level project. Pig is a scripting language. It is an open-source high-level data flow system. It is used for creating programs for Hadoop by using a procedural language known as **Pig Latin**, that is compiled into Map Reduce jobs that run on Hadoop clusters. It deals with all type of data and rapid development. It is used for web crawling, click streaming, searching logs and data analyzing. [2]

B. Pig Execution Environment

Apache Pig scripts can be executed in three ways, namely, interactive mode(Grunt shell), batch mode(Script), and embedded mode(UDF). Apache Pig Execution Environment contains two modes i.e, Local Mode and Default MR Mode.

1. Pig Local Mode:

In this mode, execution takes place on localhost and local file system and no need of Hadoop. This mode is basically used for testing purpose and local mode execution in standalone JVM.

Command : $ ./ pig – x local [3]

2. Pig Default MR Mode:

In this mode, we load or process data that exists in the Hadoop Distributed File System or Hadoop cluster using Apache Pig. By default, Pig executes on MR mode.

Command : $ ./ pig – x mapreduce [3]

C. Architecture of Apache Pig

The architecture includes **Pig Latin Script Interpreter**, used to transform the script into Map-Reduce tasks. After that Syntax checking or analysis is done by **Parser**, which creates logical plan i.e., DAG(Direct Acyclic Graph) contains logical operators (Fig. 3). **Optimizer** optimizes the logical plan and then it forward to **Compiler**, which is used to compile the services of Map Reduce tasks and then **Execution Engine** executes and stores the results on Map Reduce.

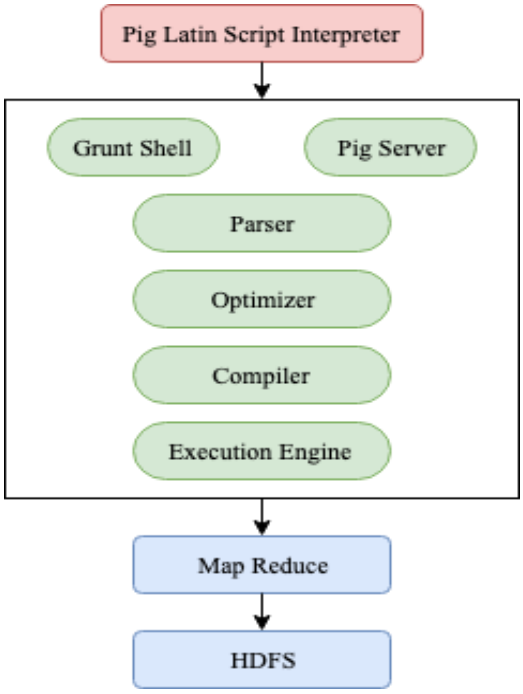

Fig 1. Apache Pig Architecture

D. Plans in Pig

There are mainly two types of plans i.e, Logical Plan and Physical Plans. Logical Plan contains a collection of operators in the script but does not contain the edges between the operators (Fig.3). After the logical plan is generated, the script execution moves to the Physical Plan, that is series of map reduce jobs and how the logical operator converted into backend specific physical operator(Map Reduce Jobs).

Fig 2. Logical Plan

E. Features of Pig

1. Rich set of operators such as Load, Join, Filter etc.
2. Easy to program.
3. Supports UDF(User Defined Functions).
4. Optimization.
5. Ability to handle all type of data.

F. Data Modelling in Pig

Data Modelling contains Fields, Tuples and Bag. Collection of Fields is known as Tuples and collection of tuples is known as Bag. Fig.3.

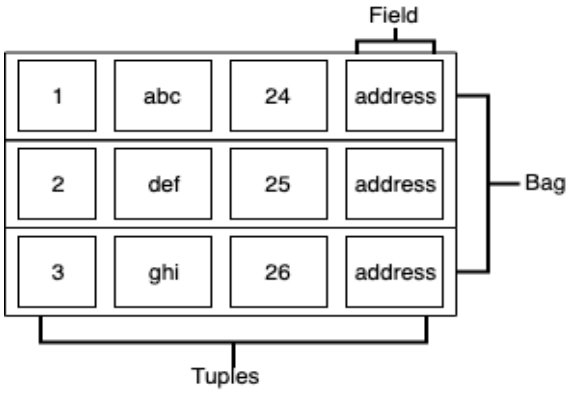

Fig 3. *Apache Pig Data Modelling*

III.OVERVIEW OF APACHE HIVE

A. Apache Hive

Initially, when Facebook started in (2004) it dealt with smaller data sets and as the time passed Facebook become popular day by day and now in (2019) Facebook is one of the biggest social media and social networking service company in the world. Initially, when Facebook started in (2004) it dealt with smaller data sets and as the time passed Facebook become popular day by day and now in (2019) Facebook is one of the biggest social media and social networking service company in the world. Facebook Initially was using a traditional Oracle database to Capture and to analyze the user behavior. So in this particular situation, they started moving towards to have a big data kind of solution and that's when Facebook become an early adopter of the Hadoop platform. when they started using this particular Hadoop to store and process the data, they faced a couple of challenges while doing that Like they had a huge number of user base so they receive terabytes of data every day and to Process that data SQL(Structured query language) standard database language which Facebook wanted to run on the top of the user database was only capable to store the structured data and process Smaller data sets only . But the challenge that Hadoop bring was that the data received was in Hadoop infrastructure and Hadoop used Java-based application to process the data on the other hand Facebook was having people's who was expert in SQL they have a very good understanding of SQL. How to run SQL queries but the kind of data they received was in Hadoop framework and if the wanted to process anything inside Hadoop it had to happen in MapReduce code. Facebook analyze the particular situation and they felt that the kind of gap between the expertise the had and the kind of tools they had to use to do the programming and that where they come up with the project called Apache Hive.

Apache Hive was introduced by Facebook in 2010, written in Java and available in SQL. it is built on the top of Apache Hadoop for providing data query and analysis. Hive give a SQL like interface, to query data and store it in various databases. The file system that integrates with Hadoop initially, was developed by Facebook later then Apache foundation took it up and developed it further as an open source under the name **Apache Hive**.

Apache Hive is an open-source interface. It is a data warehousing package built on top of Hadoop. Generally, used for Data Analysis. Hive language is similar to SQL known as HQL(Hive Query Language). It's an ETL tool.

B. Hive Execution Environment

Apache hive basically supports two execution engines i.e., MapReduce and Spark. To configure an execution engine user should perform one of the following steps: Beeline-(can be set per query) Run the set Hive.execution.engine which is the execution command of hive engine where the engine represents either MapReduce or spark engine but by default engine in MapReduce.

1. MapReduce (Execution engine of Hive):

In past, some years we require a single machine to process larger data set and the processing of data on bigger machines is called scaling up. But scaling has some problems regarding financial and technical issues. And to solve this complication the approach of a cluster of the machine were introduced and this concept is known as scaling out. [4].The idea should be very much feasible for distributed processing, for this there is need of a new program. It provides a mechanism for writing a program which helps to process the data across miscellaneous machines parallelly. MapReduce is divided into two tasks Map and Reduce. Map phase is followed by the Reduce phase. Reduce phase is always not necessary. MapReduce program is written in various programming or scripting languages.[6]

2. Spark (Execution engine of Hive):

Apache Spark is an open-source, distributed computing engine generally used for processing and analysing a large amount of data. It also works with the system to distribute data across the cluster and process it in parallel. Spark uses the library of machine learning (ML) and graph algorithm. It also supports real-time streaming and SQL apps, via Spark Streaming and Shark, respectively.

C. Architecture of Hive

The major components of the Hive are Hive Client, Services, Hadoop. Shown in Fig.4.

1. Hive Client: Hive is cross-language service development platform(means multiple language supportive platforms) like C++, Python, Java etc. using JDBC, ODBC, Thrift Drivers.

2. Hive Services: It consists of various types of Interfaces like CLI(Command Line Interface), Web Interface etc. It provides services like Hive Server, Driver, and MetaStore. It supports five backend databases i.e., Derby, MySQL, MS SQL Server, Oracle and Postgres.

3. Hadoop: It internally uses Hadoop to perform operations or to execute the queries. Hive uses MapReduce for execution and HDFS for Storage purpose.

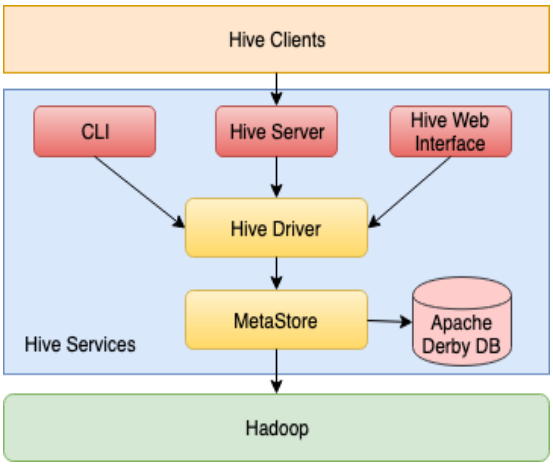

Fig.4. Hive Architecture.

D. Hive Data Flow Model:

1. **User interface (UI):** The user interface of hive enable the user to submit queries and the other operation that is to be performed on the System
2. **Driver:** Basically, it is responsible for receiving the queries submitted by Hive Client (Thrift, JDBC, ODBC, CLI, Web UL interface).
3. **Compiler**: queries are passed semantic analysis on different query block and query expression is done by the compiler.
4. **Metastore:** It stores the metadata for Hive relations and tables (Schema and their location). It provides directly the information to the client

using the Metastore service API. Three modes i.e., Embedded, Local and Remote mode.

5. **Execution Engine:** After completion of compilation and optimization, the Execution Engine will execute the tasks in the order of their dependencies using Hadoop.

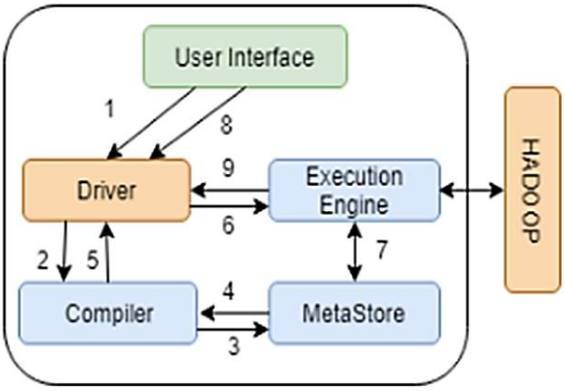

Fig:5 Data Flow Model(Hive)

E. Features of Hive:

1. It provides an easy way to summarise data, analyse and query.
2. HQL doesn't require any additional knowledge of Language. It is similar to SQL.
3. Also runs Ad-hoc queries for data analysis.
4. It supports partitioning of data to improve performance.

F. Limitations of Hive:

1. Not recommended for row-level updates.
2. Latency for hive queries is high.
3. Not designed for OLTP.

V. CONCLUSION

Apache Pig and Hive both are the foremost frame-work which helps Hadoop to work more efficiently. This paper briefly explains the overview of Apache Pig and Hive that how these two framework exe-cutes i.e about execution engine, architecture and data modelling of both the frameworks and last but not the least features of both frameworks in accord-ance to Hadoop framework. This paper basically highlights the mechanism of Pig and Hive that how it deals with the Big Data that the world is having, how it helps to process that data which is not struc-tured. Today the world is using unstructured data and that too in huge volume so these two frame-works are proved to be the boon in the field of Big Data Analytics. This field is very trendy and the data is growing with every second, so these two plays a very significant role in Apache' Hadoop project.

VI.REFERENCES

[1]. Kadhar Bhasha J, Dr. M. Balamurugan, "A Review on Hive and Pig", International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), 2017, ISSN 2456-5717.

[2]. Vaishali Chauhan, Meenakshi Sharma, "Hive, Pig & HBase Performance Evaluation for Data Processing Applications", International Journals of Advanced Research in Science Engineering, 2016, ISSN 2319-8354.

[3]. Ms. Sarika Rathi, "A Brief Study of Big Data Analytics using Apache Pig and Hadoop Distributed File System", International Journals of Advanced Research in Science Engineering, 2017, ISSN 2278-1323.

[4]. Rupinder Singh, Puneet Jai Kaur, "Analyzing performance of Apache Tez and MapReduce with Hadoop multinode cluster on Amazon cloud", 2016, DOI 10.1186/s40537-016-0051-6.

[5]. Richa Vasuja, Ayesha Bhandralia, Kanika Chuchra, "Daemons of Hadoop: An Overview", International Journal of Engineering Research and Technology, 2018, ISSN: 2278-0181.

[6]. Maitrey S, Jha CK. "Handling Big Data efficiently by using MapReduce technique.", IEEE international conference on computational intelligence & communication technology (CICT), 2015, pp 703-8.

Cite this article as : Saiyam Arora, Abinesh Verma, Richa Vasuja, "An Overview of Apache Pig and Apache Hive", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 432-436, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195250
Journal URL : http://ijsrcseit.com/CSEIT195250