Big Data Analytics Using Hadoop Tools- Apache Hive vs Apache Pig

S. Dixit[1], D. Jaykar[2], S. Gaikwad[3]

CSE Department, SKN Sinhgad College of Engineering, Korti, Pandharpur.

dixitsanika97@gmail.com

jaykardhanashree1996@gmail.com

SayaliGaikwad1001@gmail.com

Abstract— Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. Big data usually includes datasets with sizes beyond the ability of commonly used software tools to capture, manage and process data within a tolerable elapsed time. Big data analytics is the process of examining large and varied datasets. Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn leads to smarter business moves, more efficient operations, higher profits and happier customers. Today's advances in analysing big data allow researchers to decode human DNA in minutes. Predict where terrorists plan to attack, etc. Apache Hadoop is an open source, java-based programming framework that supports the processing and storage of extremely large datasets in distributed computing environment. The most well known technology used for Big data is Hadoop. It is actually large scale batch data processing system. The apache Hadoop framework has Hadoop Distributed File System(HDFS) and Hadoop MapReduce at its core. There are many Big data tools for handling big data which are built around Hadoop. Two popular tools are Apache Pig and Apache Hive. Apache pig is a high-level platform for creating programs that run on apache hadoop. The language for this platform is called Pig Latin. Hive is an open-source data warehouse system for querying and analysing large datasets. This paper aims to analyse some of the different analytics methods and tools which can be applied to big data. It focuses on Hadoop's components and different analytical tools i.e. Pig and Hive

Keywords— Big Data, Map Reduce, Hadoop, Apache Pig, Apache Hive, HDFS

I. INTRODUCTION

Big Data :

Big data is a phrase used to mean a massive volume of both structured and unstructured data i.e. so large it is difficult to process using traditional database and software techniques. 3Vs (Volume, Variety and Velocity) are three defining properties of big data. Volume refers to the amount of data, Variety refers to the number of types of data and Velocity refers to the speed of data processing. Big data helps the organizations to create new growth opportunities and entirely new categories of companies that can combine and analyze industry data.

Hadoop :

Apache Hadoop is an open-source software framework for storing data. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called Map Reduce. Originally hadoop is designed for computer clusters built from commodity hardware. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

Apache Hive :

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. The traditional SQL queries must be implemented in the Map Reduce. Java API to execute SQL applications and queries over a Distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries into underlying java without the need to implement queries in the low-level java API. The language for this platform is called HiveQL. Internally, a compiler translates HiveQL statements into a directed acyclic graph of MapReduce which are submitted to hadoop for execution.

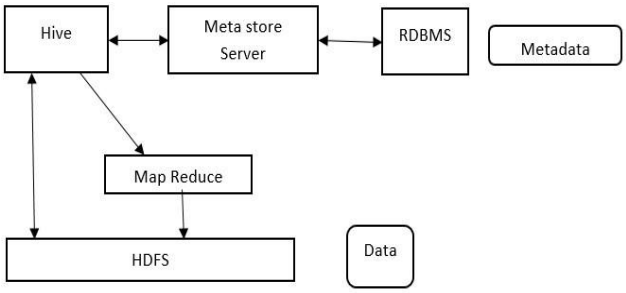


Fig.: Architecture of Hive

Apache Pig :

Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in Map Reduce. Decrease in deployment time. Apache pig provides data operations like ordering, filters and joins. While SQL is designed to query the data, Pig Latin allows you to write a data flow that describes how your data will be transformed (such as aggregate, join and sort).

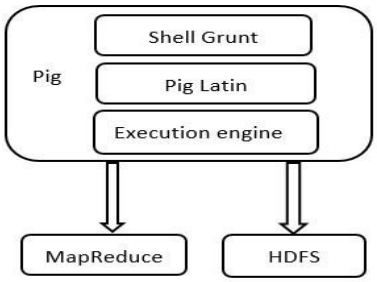


Fig.: Architecture of Pig

TABLE I

Apache Pig	Apache Hive
Apache pig was created at Yahoo	Apache Hive created at Facebook
Pig uses language called Pig Latin	Hive uses language called HiveQL
Pig loads the data effectively and quickly	Hive execute quickly, but not load quickly
Apache Pig can handle Structured, Unstructured, Semi-structured data	Hive is mostly for Structured data
Pig has a procedural data flow language	Hive has a declarative SQLish language

II. LITERATURE REVIEW

- Crime data analysis using pig with hadoop
  This paper suggested that, the best place to look up to find room for improvement is the voluminous raw data that is generated on a regular basis from various sources by applying Big Data Analysis which helps to analyze certain trends that must be discovered, so that law and order can be maintained properly and there is a sense of safety and well-being among the citizens of the country.
- Big data analytics using hadoop tools- Apache hive vs apache pig
  Big data technologies continue to gain popularity as large volumes of data are generated around us every minute and the demand to understand the value of big data grows. The Apache Hadoop framework has Hadoop Distributed File System (HDFS) and Hadoop MapReduce at its core. There are a number of big data tools built around Hadoop which together form the 'Hadoop Ecosystem.' Two popular big data analytical platforms built around Hadoop framework are Apache Pig and Apache Hive. The purpose of this paper is to explore big data analytics using Hadoop. It focuses on Hadoop's core components and supporting analytical tools Pig and Hive.

III. PROBLEM STATEMENT

Big data analysis using Hadoop tools - Apache Hive vs Apache Pig.

IV. OBJECTIVE

- To analyze Big Data
- To handle structured, unstructured and semi-structured data
- To processing, transforming and analyzing data

V. EXECUTION

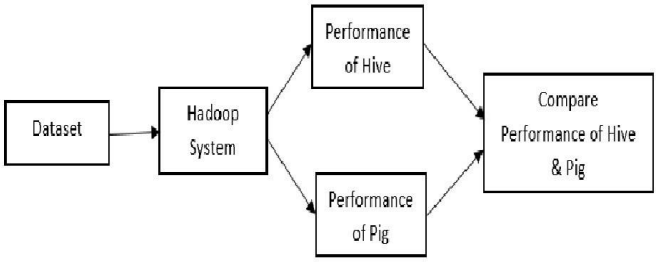


Fig.: Flow of execution

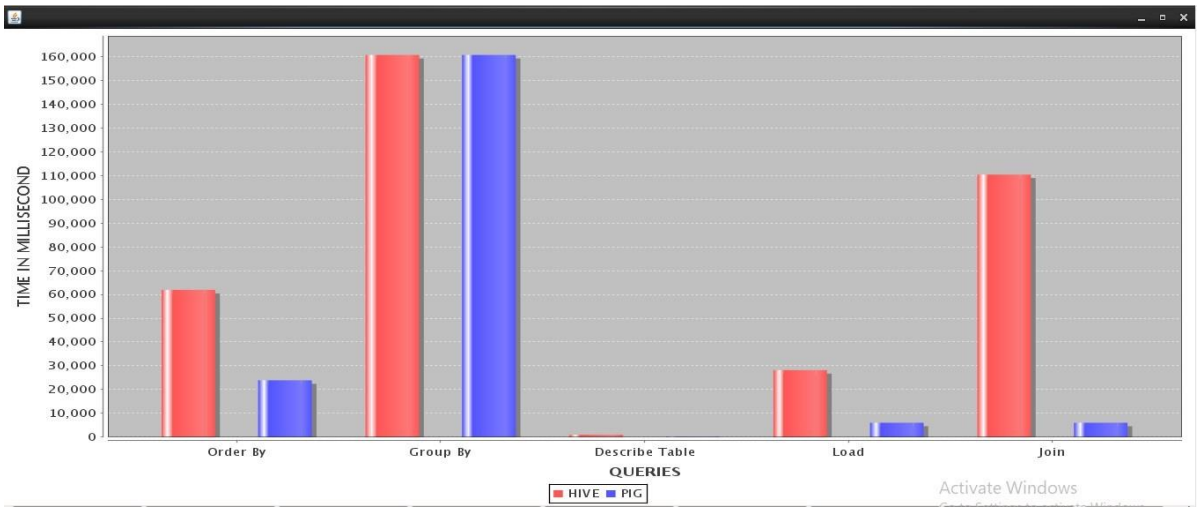
First import the dataset into the hadoop distributed file system (HDFS) for process on it using different hadoop tools. Data will process using Apache Hive and calculate the require processing time. At the same time data will process using Apache Pig and calculate the require processing time. Display the comparative result by using graph.

Fire Query :

In this page, user can fire the queries and that queries executed using Hadoop tools – Apache hive and apache pig. After selecting the query and clicked on OK button, it will display result. There are different queries which are Describe table, load data, join tables, group by and order by. To applying these queries we use the dataset named as retail_db. In this dataset there are multiple tables which are having large number of records.

VI. EXPERIMENTAL RESULT

Time Analysis :



The above graph shown that time analysis between different queries using Hadoop tools. Each query executed using Hadoop tools i.e. Apache hive and Apache pig. According to processing time it will display the result.

TABLE II

Sr. No.	Query	Time in Minute	
		Hive	Pig
1	Describe Table	0.02	0.01
2	Order By	1.03	0.39
3	Load	0.46	0.10
4	Join	1.84	0.12
5	Group By	2.67	2.48

The above table shown that time (in minutes) required for execution of queries using Apache Hive and Apache Pig.

VII. CONCLUSIONS

Big data contains huge data in size, analysis and structuring is big challenge. Implementation of hadoop on big data gives solution for big data i.e. how it is manageable by reducing our time. After executing some queries on apache hive and apache pig. Both component will help you achieve the same goal. According to final result, apache pig is more faster comparing apache hive. Apache pig requires less time for executing query than apache hive.

REFERENCES

[1] Prof R. Angelin Preethi and Prof J. Elavarasi, "Big data analytics using hadoop tools-apache hive vs apache pig", International journal of emerging technology in computer science & electronics(IJETCSE) ISSN:0976-1353 Volume 24 Issue 3-february 2017.

[2] Ramensh R, Divya G, Divya D, Merin K Kurian, "Big Data Sentiment Analysis using Hadoop", (IJIRST) International journal for Innovative Research in Science & Technology, Volume 1, Issue 11, April 2015 ISSN: 2349-6010.

[3] S. Dhawan and S. Rathee, "Big Data analytics using hadoop components like pig and hive", American International journal of research in science, technology, engineering and mathematics, vol. 2, (2013), pp. 88-93.

[4] Ms. Sarika Rathi, "A brief Study of Big Data Analytics using Apache Pig and Hadoop Distributed File System", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) ISSN: 2278-1323 vol. 6, Issue 1, January 2017.

[5] K. Ramesh, K Raghavendra Rao, "Enhancing the Processing Time By Managing MySQL Cluster, Apache Pig and Apache Hive Methods", International Journal of Advanced Technology and Innovative Research, ISSN: 2348-2370, vol. 08, October-2016.

[6] Dr. E. Laxmi Lydia, Dr. M. Ben Swarup, "Analysis of Big Data through Hadoop Ecosystem Component like Flume, MapReduce, Pig and Hive", International Journal Of Computer Science Engineering(IJCSE).

[7] Arushi Jain, Vishal Bhatnagar, "Crime Data Analysis using Pig with Hadoop", International Conference on Information Security & Privacy(ICISP2015), 11-12 December 2015, Nagpur, INDIA

[8] Pooja Jain, prof Jay Prakash Maurya, "Comparative Analysis using Hive and Pig on Consumers Data", International Journal of Computer Science and Information Technologies(IJCSIT), vol. 8(2), 2017, 285-291.