International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

Web Visitors Data Analytics using Hadoop Ecosystem

Urvashi Sharma, Dr. Sunita Varma
Shri G.S.I.T.S Indore, 452003,
INDIA

Abstract. **Hadoop is an open source framework used for storing and processing huge amount of data i.e big data. In this paper by using these tools the data is stored, processed and analysed. Data is stored in HDFS by using flume, filtered and analysed through pig and hive. The comparsion is done to find out time required by pig and hive for processing the same query. It is concluded that the time required to process the same query to hive is less as to find through pig.**

1 INTRODUCTION :

Big data means a massive data set. It cannot be study by traditional computing techniques. Immense volume of the data cannot be processed or stored by applying traditional method. Normal relational database system cannot accumulate this data that's why we require some separate type of data structure and database over which we can able to analyze data which outgrowth proper output. Hadoop is an open source framework for storing data. Hadoop can handle various form of structured and unstructured data.

2 TECHNOLOGIES USED :

Hadoop 1.0 consist of two components, HDFS and map-reduce programming tool. Hadoop 2.0 is also called Hadoop Ecosystem which consist of following components-

Apache Pig : Pig is high level procedural language platform used for programming on Hadoop and Map reduce. Pig [7] is an Apache open source project.

Aapache hive : Hive is the tool to process structured data in HDFS. It remains on the top of the HDFS to help, summarize and analyse data.

HDFS: Hadoop distributed file system- It stores data files as similar to the original form as possible.

HBase: It is Hadoop's database and compares well with an RDBMS. It reinforce structured data storage for large tables.

Zookeeper: It is combination service for distributed application.

Sqoop : It is used to move bulk data between Hadoop and structured data stored such as relational data.

Flume : Flume is system used for moving large quantities of streaming data into hdfs .

3 PROCESS OF ANALYSIS THROUGH HADOOP ECOSYSTEM

Hadoop Ecosystem The process of Hadoop services typically involves technique to analyze data. (i) The data is stored in HDFS by using flume. (ii) This data is filtered and analyzed through pig and hive. (iii) The data is stored to SQL.

Data Ingestion

Data ingestion is done through flume utility. This process involves storage of the data from local system to HDFS. For this we have source, sink and channel. Each of them will assign a variable name, then configure the source and describe its type give the path where we have to move the file. Now describe the sink, define it, then the path given to the project directory will be created and flume data directory will be created, then assign the packet size, time interval and roll count. Now describe the channel, define memory its capacity and connect source and sink to channel.

By flume utility which is one of the main component of Hadoop ecosystem, data is stored from local storage to HDFS. The channel is used, which buffers events in the memory. Finally connect the source and the sink to the channel.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

Data Filtration and Analysis

In the process of filtration only required columns get filtered in which, we need to select columns according to given queries and finally analysis will be done through pig and hive. Analysis through pig can be done by creating an individual script for different queries. Pig is an open source technology that enables cheap storage and processing of large datasets without requiring any specific formats. Pig is used for analyzing and querying on large data set which is stored in HDFS. Required query will be run by pig script name.pig command and the result will get stored in part file and displayed by HDFS dfs –cat /project directory /script name /part file.

Analysis through hive will be done by firing query (HQL). Hive is SQL inspired language. Hence filtration and analysis is important part of the proposed approach in order to analyze

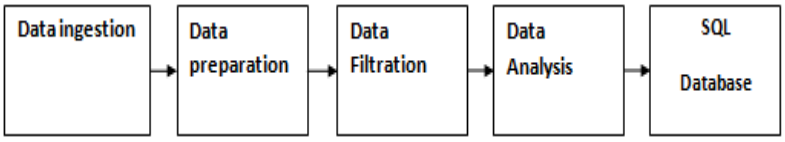

Fig. 1. Architecture Diagram of Process of analysis through Hadoop ecosystem

Large data set is being analyzed through Apache pig that consist of simple format and easy to program. It is similar to SQL. Pig was initially developed by yahoo. With pig language, you have facility to write commands in java etc. In both the cases, pig and hive are use to analyze data For many traditional data operation pig latin includes operators. For reading, processing and analyzing data, Pig latin developing their own functions hence it is extensible.

DATA TRANSFER AND STORAGE

Unstructured and structured data are two type of data. Analysis through pig and hive is done on structured or semi structured data where as relational databases deals with structured data in which tables are created. In order to transfer the results from unstructured to structured, we require sqoop which is one of the main component of Hadoop ecosystem. Sqoop is used transfer data between HDFS and relational database system. It is used to transfer data from files system to relational database and import data from database such as Mysql to HDFS.

4 EXPERIMENTS AND RESULTS

This segment shows the experiments and results of the system explained in the Section 3. To conduct experimentation, analysis through both pig and hive is done.

4.1 Performance Analysis of result

Analysis and comparison of pig and hive is performed on these datasets by using the languages of pig and hive. Following are the required queries which are used to extract the information from the dataset:

_ Which is the most viewed page?
_ Which is the most viewed product?
_ which is most frequently used web browser?
_ Report top 3 product, ip address , product corresponds to those ip address?
_ Report the count of product and ip address?

The comparison is done to find out time required by pig and hive for giving result of the query. It is concluded that the time required for processing the same query is less in hive as compared to pig. Pig takes more time because it is written in pig latin language in the form of scripts and this scripts run for each query repeatedly. We need to run different script of different query, so it will take more time to run pig scripts where as hive is a data warehouse in which we only need to fire query to get required results so it will take less amount of time and hence the performance of hive better than pig.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

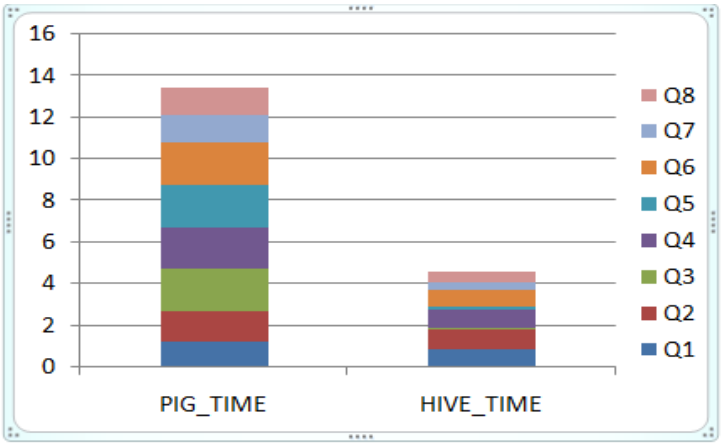

Fig. 2. Graphical representation of analysis through pig and hive

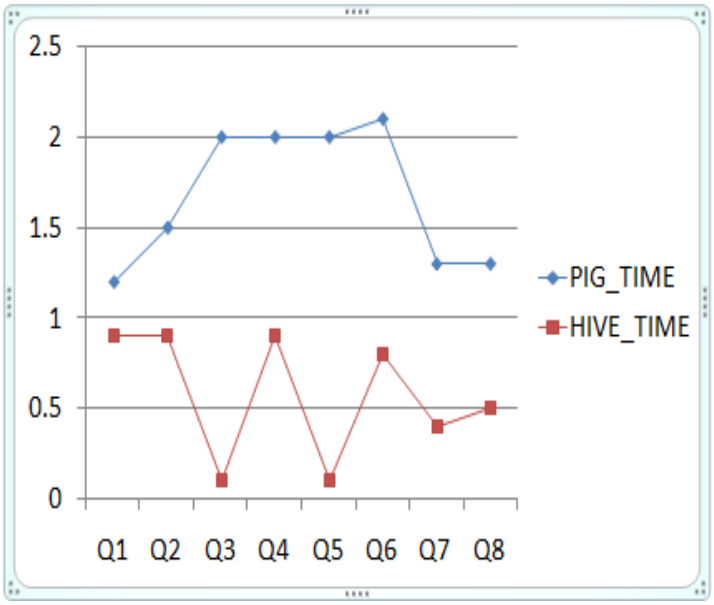

Fig. 3.Visualization of the Result

From the experiments and result appear in Table 1, we notice that as we increase the number of instances, the time to produce the plan also increases. The number of nodes grnerated increase with the increasing number of objects of each instance.

After applying the analysis method of both pig and hive, it has been observed that graph of the hive time is less than one minute and pig time is more than hive time hence performance of hive is better.

It has been observed that the graph represented after applying, analyzing method through pig is above 1 minute than that of hive which takes less than 1 minute every time. Hence it is clear from the graphical representation that takes more time than that of hive so the performance of pig is poor than hive and pig is more suitable for filtration where as hive is for analysis of log data of web visitors.

CONCLUSION

The outcome of the paper is focused on analyze of log data of web visitors to find out queries like most viewed page, product, user ip , report which will further proved profitable to most of the firm to improve sale of their product and earn profit. In the proposed approach various components of Hadoop ecosystem has been used for example flume is used for data ingestion from local system to HDFS, pig is use for filtration of the required columns, analysis using pig script in pig latin and hive is use for analysis, storage of data using hql language by firing queries and sqoop is used to store data in my sql that is relational data base in the form of tables. Various queries used in this project are to find out most viewed page on the web portal, most viewed product, most viewed web browser generate report of product, generate report of user ip and their counts.

Finally comparison of time taken by pig and hive is done and hence we conclude that time taken by pig is more than that of pig that is time taken to run the query of most viewed page through pig is 1 minute more .

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 10, October-2020

REFERENCES

[1] V.Maria Antoniate Martin, Dr. K. David :" Big Data and challenges " In International Journal of scientific Research in computer engineering and Information Technology 2018 IJSRCSEIT volume :3
[2] Keren Quaknine, Michael Carey: " The Pig Mix Benchmark on pig,Map reduce and HPCC System", 2015 IEEE International conference on big data.
[3] Ashish Thusoo, Joydeep Sen Sarma, Namit jain , Facebook Data Infrastructured.
[4] Zahid Ansari, Swarna: "Apache pig – A Data Flow framework based on Hadoop Map reduce". International Journal of Engineering Trends and Technology (IJETT)-volume 50 Number 5 August 2017.
[5] Sooyong Jung, Yongtae Shin, " Study of Big Data Collection Scheme Based Apache Flume for Log Collection " International Journal of computer Theory and Engineering, Vol. 10 No. 3 ,June 2018
[6] Ashwini A. Pandagale, Anil R.Surve " Big data Analysis using Hadoop Framework" IJRAR –International Journal of Research and Analytical Reviews, vol 3 Issue: 1, Jan – March 2016 .
[7] Mr.S.S Arvinth, Ms. A. Haseenah Begam "An efficient Hadoop Framework Sqoop and Ambari for Big Data Processing"International Journal for Innovative Research in Science & Technology,Volume 1,march 2015