Pattern Finding In Log Data Using Hive on Hadoop

Swapna Sahu

M.Tech Scholar
Bansal Institute of Research & Technology
Bhopal,India

***Abstract*: Web log file, in the computing context, is the log file which get routinely generated and maintained by a web server. Analysing web server access logs will give information regarding user's behavior. Log files generate data which contain valuable information from the user which get stored in the web server. Server logs act as a guest sign-in sheet. Log files give information about the pages which had a heavy traffic and least. What sites refer visitors to your site? What pages that your visitors view? Because of the tremendous usage of web, the web log files are growing at faster rate and the size is becoming huge. Processing this explosive growth of log files using relational database technology has been facing a bottle neck. To analyse such large datasets we need parallel processing system and reliable data storage mechanism, Big data uses the Hadoop where massive quantity of information is processed using cluster of commodity hardware. In this paper we present the Hadoop framework for storing and processing large log files and also analysing through hive, Hive is used in pre-processing of voluminous of log files and help us to find out the statics present in website and which help in our learning too.We can also perform optimization on hive query and we also compare the performance of both the analytical tools on analysing log files.**

***Index Terms*: Hadoop, data mining, logfile analysis, behaviour mining, web mining, hive, pig.**

I. INTRODUCTION

Web data research has encountered a lot of challenges in developing an effective web page recommendation system. Because in various surveys of the Web (WWW) it is estimated that roughly one million new web pages are added every day and over 600GB of data changes per month. Data tells around three million webpage can be found online, it also tells one page for every two people on the earth. Such a tremendous data is available, this data is generally unstructured in nature and we need to discover useful knowledge from the domain. We can find relevant information like web page navigation pattern, web page recommendation. Even more information can be taken out such as web browsers, active users. Profile mining is very important because without knowledge of any user profile/web user we will helpless to find or predict next web page for web user or active user with the help of web log and web user session we will find about user profile.

This question is answered in this manner that log comes in various form but as the organization grow ,its application also enhance which results in voluminous amount of data that is useful which we are getting from the web, network devices, firewall logs, and databases server provides a lot of information. This huge datasets can be handled by Apache Hadoop. Log data contain info which we have to extract, then after processing the extract content we get proper information. General use of log file is to record event such as access record or failure record. In Banking transaction log are maintained. How much transaction is happening per second? We are getting meaningful information that consist of extraction (map) and pattern is generated by reducing it from a web log. This is going to help to maintain file access statistics which will support to take proper detection and decision which will ultimately benefits to the web administrator.

As the growth of data increases over years, storage and analysis becomes incredible, this in turn increases the processing time and cost efficiency. Though various techniques and algorithms are used in distributed computing the problem remains still idle. To overcome by this problem Hadoop Map reduce is used, to process massive files in a parallel manner. The use of World Wide Web emits data in larger quantity as users are more interested in performing their day to day activities through online. Interaction between the user in a website is examined through web server log files, a computer generated data which is in semi structured form . This paper we are analysis of weblog files using Hadoop Mapreduce to preprocess the log files and generation of patterns.

HADOOP

Apache is a open source framework which is reliable, scalable and is used for distributed computing. It is generally used for batch and offline processing. Apache Hadoop have a rich library that allows distributed processing of huge datasets(exabyte,petabyte) sets beyond clusters of computers using thousands of computational independent computers. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is a good option for analysis voluminous datasets such as twitter analysis, pattern analysis. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different cluster nodes. Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data of commodity hardware in a reliable and fault-tolerant manner.

Apache Hive

Apache Hive was created by Facebook. The main purpose of the creation of Apache Hive to analyzing voluminous datasets. Apache Hive data warehouse software program allows analyzing, writing, and handling huge datasets. Data is residing in distributed storage and we are handling through SQL. It is one of the most commonly implemented application used for analysis structured data through

Relational model and through SQl interface. Hive is an analytical tool that is built on the top of Hadoop. Data encapsulation, data summarization and analysis of large dataset are the major job done by Apache Hive. Hive helps us to improve the latency and reduce the decreasing efficiency. Hive is believed it is standard for interactive and batch jobs. SQl semantics works on the petabytes with proper time. Hive allow the user to reduce the response time and access the data Hive tables is similar to relational databases. Hive have the partition option in tables.It allow appending and overwriting into the table. Hive table is having the property of serialization, by using serialization in any particular dataset, the tables get serialized and has a particular directory in hdfs. Hive support the data formats like smallint, timestamp, binary, char, boolean, double, bigint, decimal, int, string, float etc. For making complex data type such as map and array we can combine this with primitive data types.

II. LITERATURE REVIEW

According to [1], Web mining[13] is the integration of information that is extracted from web data that is taken from the document, serverlogs, web content ,hyperlink.and usage logs of web sites. Web use mining is the way of applying information mining systems to find utilization design from the web information. It is one of the mechanism to locate personalization of web pages. It is applied in various levels such as server level, client level and proxy levels. It also tells that it comes from various resources through they interact with the browser, HTTP protocol and the status code [3]. However in the existing situation the online user is increasing day by day and use of data per user is increasing randomly. There are various kind of form we are using in web, if user click on the submit button all its activities are saved in server side. All this activities are recorded and maintained by the server which is know as web log or log file. The files are rich in information ,it contain entries[4] like IP address of the system making the request, hit or miss, server location and name of the requested file, the HTTP status code, the file size

We can find various log like Event logs, instant logs, message logs, transaction logs, xml logs and error logs. There are two types of log files. Log files are also categorized on the basis of location that is web server logs and application server logs. A web server [5] maintains two types of log files: Access log and Error log. An access log contains all the request which had been by the customer. A log files have parameters which are very useful to recognizing user browsing patterns [6, 7, 8].

Weblogs are mined and the web mining will be helpful to the server and E-commerce industry wish to predict the behavior of customer and it may generate the pattern of their usage. Each day new users are added and size of web access log is increasing [10]. In websites taking care of a huge number of concurrent users can create hundreds of petabytes of logs every day. Traditional techniques used RDBMS for storage and analyze. RDBMS cannot store the voluminous amount of data and the data should be in the table format. In this way, to examine such enormous web log document productively and successfully we have to grow speedier, proficient and viable parallel and adaptable data mining algorithm. To get faster result we have to run the application on various nodes to store petabytes of web log data and parallel computing model for analysis. Hadoop framework gives trustworthy nodes for data storage to keep large log file data in a distributed manner and parallel processing features to process web log file data efficiently and effectively[11,12]. The preprocessed web logs by Hadoop MapReduce environment is further processed for prediction of user's next request without disturbing them to increase the interest and to reduce the response time with ecommerce system.

This paper shows how log file is analysed through Hadoop framework and how the computation process concurrently and parallel processing. We had various resources through which data are collected then its loaded into HDFS for facilitating MapReduce and Hadoop framework. Hadoop process the huge amount that leads to lesser response time and minimum computation. Ecommerce industries have to process the voluminous amount with the help of big data analytics tools with less response time and accuracy. This task can also be extended with semantic analysis for better prediction.

Big data analytics have fascinated interest from academia, entertainment, ecommerce and various other industries. This analytics help us to extract knowledge and information. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Some of the top most companies like Amazon, Sunglassesindia, Mithaimate, Orosilber.,IndianAirlines, Irctc,,Violetbag.,Snapdeal,Myntra,Bigdata etc. Those companies collect voluminous data and analysis is done with help of tools and have simple interface.

III PROBLEM DEFINITION

Companies like Flipkart, Snapdeal and Amazon routinely produces a huge amount of logs on a daily basis. They continually improve their operations and services by analyzing the data. Analyzing these huge amount of data in a very short period of time is a crucial task for any business analyst. The problem of log files analysis is complicated because of not only its volume but also its disparate structure. The log files are semi-structure or unstructured type. The use of traditional tool and techniques are not feasible, and the traditional tool cannot handle the large amount of dataset or an unstructured data.

Data mining techniques first do the pre-processing and then applying analytics methods after using analytics methods we get the value. Data mining is related with many research technologies like artificial intelligence and machine learning and so on. Data management and big data in data mining is significantly different in size. Usually the primitive method is to find out the similar value. Data mining follows the mechanism of extracting knowledge needs data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation etc. Big data came out after solving the requirements and challenges of data mining [13].

IV PROPOSED WORK

For analyzing complex datasets, we are using Hadoop framework of Apache Software foundation,. We are using an analytical tool Hive having which is a powerful tool designed for deep analysis.

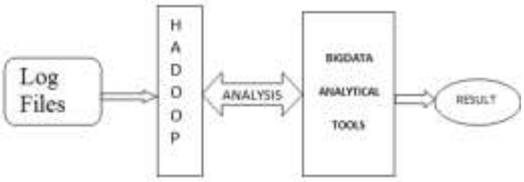


Figure1. Workflow Diagram

In this paper we are using analytical tool Hive that can handle the problems which are raised by massive datasets and to find out the dynamic characteristics of data and to carry out task on social media data sets. Hadoop is a standard platform we are using on single node ubuntu machine which provides the solution of big data, Hadoop uses MapReduce framework where our data gets mapped to frequent datasets and we reduced it to smaller size which is manageable [9]. Finally Bigdata analytical tools can be used for refinement of unstructured data and analyse them.

V.EXPERIMENT RESULT ANALYSIS

The following system configuration we require 2 GB of RAM running ubuntu 14, i3-2410M CPU @ 2.30 processor and. Then configure the Hadoop, we are using hadoop-1.1.2 on ubuntu and along with Hadoop. We also integrate bigdata analytical tools hive and pig on top of the Hadoop, to achieve the results we are pursuing the following methods:

- Loading Data into HDFS.
- Analyzing using Apache Hive.

Loading Data into HDFS

First we are loading different access log files in to HDFS, in our dissertation we can analyse NASA web access log which are common access log. Figure 2 shows the loading a log file into HDFS. And in this figures we can clearly see that there is no structure between the data of these logs file. After loading these different log file into HDFS we can analyze using bigdata analytical tool such as apache hive. In the next section we will analyze these complex log files.

File: /logfile

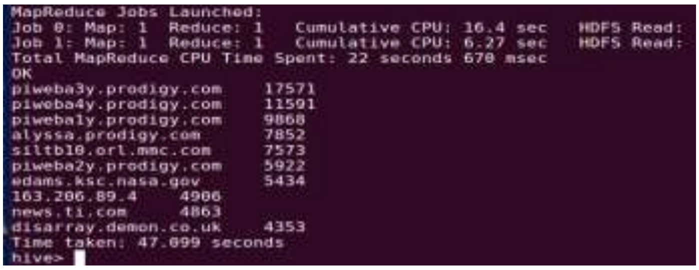
Goto :

```
[illegible]
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size)

Analyzing using Apache Hive

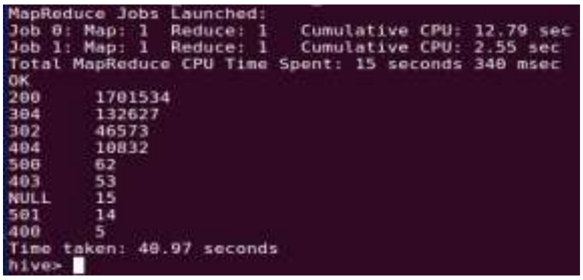
After storing the log raw data into HDFS, now we can start analyzing these complex log files using apache hive. For analyzing common log file we have created a table named as nasa_log table to store the access log data efficiently in structured manner. For converting the unstructured and complex log file into structure tabular format, we can use Regex SerDe properties into hive which can transform the unstructured data into structured format. For creating table and applying regex serde properties into table. For these hive queries, hive engine launches a mapreduce job for pre-processing the log files, the mapreduce job is launched by running a hive query on terminal. After finishing an execution of mapreduce job we can get the output of that query. In figure 3 we are getting the host or ip address which has maximum frequency or hit counts and the time taken by hive query is also shown in figure 3 that is takes 47.099 seconds to finish the execution.



```
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 16.4 sec   HDFS Read:
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 6.27 sec   HDFS Read:
Total MapReduce CPU Time Spent: 22 seconds 670 msec
OK
piweba3y.prodigy.com    17571
piweba4y.prodigy.com    11591
piweba1y.prodigy.com    9868
alyssa.prodigy.com      7852
siltb10.orl.mmc.com     7573
piweba2y.prodigy.com    5922
edams.ksc.nasa.gov      5434
163.206.89.4    4906
news.ti.com     4863
disarray.demon.co.uk    4353
Time taken: 47.099 seconds
hive>
```

Figure 3. Maximum Hits From IP Addresses

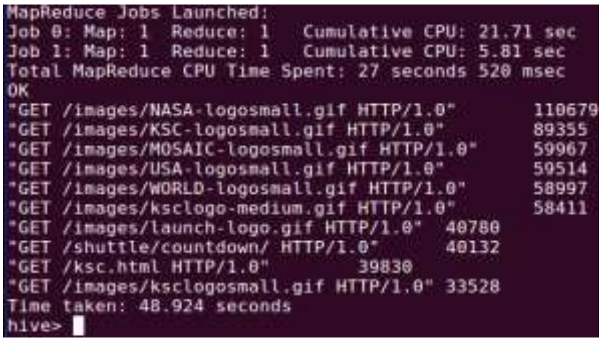
Similarly we can also find the various status code which we can get along with its frequency and the time taken by hive, figure 4 shows the various status code which we are along with its frequency and the time taken by hive query.



```
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 12.79 sec
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.55 sec
Total MapReduce CPU Time Spent: 15 seconds 340 msec
OK
200     1701534
304     132627
302     46573
404     10832
500     62
403     53
NULL    15
501     14
400     5
Time taken: 40.97 seconds
hive>
```

Figure 4. Various Status Code Along With Frequency

Similarly we can find the maximum hitting pages along with frequency which user can access and the time taken by hive are shown in figure-5.



```
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 21.71 sec
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 5.81 sec
Total MapReduce CPU Time Spent: 27 seconds 520 msec
OK
"GET /images/NASA-logosmall.gif HTTP/1.0"       110679
"GET /images/KSC-logosmall.gif HTTP/1.0"        89355
"GET /images/MOSAIC-logosmall.gif HTTP/1.0"     59967
"GET /images/USA-logosmall.gif HTTP/1.0"        59514
"GET /images/WORLD-logosmall.gif HTTP/1.0"      58997
"GET /images/ksclogo-medium.gif HTTP/1.0"       58411
"GET /images/launch-logo.gif HTTP/1.0"  40780
"GET /shuttle/countdown/ HTTP/1.0"      40132
"GET /ksc.html HTTP/1.0"        39830
"GET /images/ksclogosmall.gif HTTP/1.0" 33528
Time taken: 48.924 seconds
hive>
```

Figure 5. Maximum Hitting Pages

Optimizing Query Performance

In this we can also optimize the hive query performance, we can perform serialization process at the starting table and store the resultant table into new table and then apply all the query on these new resultant table by which we can get the result faster as compared to perform same operation on deserialize table. For this we can execute the different query on two hive tables first table in which the optimization is not present and second in which we perform optimization process for which we can get output of queries with different execution time and the time taken by query in both the tables. For this we can create another table call lognew and the schema difference between both table are shown in figure -6.

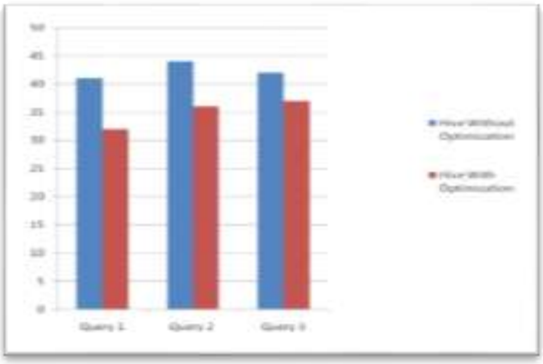


Figure 6. Execution time taken by query on hive tables

Comparison

After analyzing the access logs from analytical tool hive and pig we can see the result. The results of both the tools are same means both are accurate in terms of accuracy but both are taking different execution time to generate the result. Table 1 shows the time taken by hive and pig to generate the result.

Analysis of Data	Hive	Pig
Table Creation	3.139 seconds	2 seconds
Loading/Storing data	4.246 seconds	1 min 27 seconds
Querying dataset	1.626 seconds	14 seconds

Table 1. Time taken by hive and pig

From the above table, we observed that time taken by Hive is lesser than time taken by Pig in all aspects.

VII CONCLUSION:

World Wide Web has given necessitated the customers to make use of automated tools to find desired information resources and to pursue and to find out relevant usage pattern. We have applied best programming model ie Hadoop MapReduce programming model for analyzing web application log files. In the current system, we are doing storage using HDFS and log files analysis part is dobe by MapReduce model and we get results minimum response time. We get classified results of analysis through hive query which is written over Mapreduce and we can also compare the performance on hive and pig and the hive perform better in processing access logs over pig in terms of execution.

REFERENCES

[1] Dr.S.Suguna, M.Vithya, J.I.Christy Eunaicy, “Big Data
Analysis in E-Commerce System Using HadoopMapReduce” in 2016 IEEE.

[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3] M.Santhanakumar and C.Christopher Columbus, “Web Usage Analysis of Web pages UsingRapidminer”, WSEAS Transactions on computers, EISSN: 2224-2872, vol.3, May 2015.

[4] Shaily G.Langhnoja ,MehulP.Barot and DarshakB.Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery “,International Journal of Data Mining Techniques and Applications, vol.2 ,Issue.1, June 2013.

[5] Web server logs ://http. Sever side log.org.

[6] Nanhay Singh, Achin Jain, Ram and Shringar Raw, “Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques”, International Journal of Data Mining & Knowledge Process(IJDKP) vol.3, Issue.4, July 2013.

[7] J.Srivastava et al, “Web usage Mining: Discoveryand Applications of usage patterns from Web Data“, ACM SIGKDD Explorations, vol.1, Issue. 2, pp.12-23, 2000.

[8] S.Saravanan and B.UmaMaheswari, “Analyzing Large Web Log Files in A HadoopDistributedCluster Environment”, International Journal of Computer Technology & Applications, vol.5, pp. 1677-1681.

[9] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[10] K.V.Shvachko, “ TheHadoop Distributed File System Requirements”, MSST ’10 Proceeding of the 2010 IEEE 26th Symposium on Mass Storage System and Technologies(MSST).

[11] Apache Hadoop ://http://hadoop.apache.org.

[12] A white paper by OrzotaInc, “Beyond Web Application Log Analysis using Apache Hadoop”.

[13] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More--MattheA. Russell.

[14] R Chaure, SK Shandilya, Firewall Anamolies Detection and Removal Techniques – A Survey, International Journal on Emerging Technologies, 2010

[15] AK Dubey, SK Shandilya, A Comprehensive Survey of Grid Computing Mechanism in J2ME for Effective Mobile Computing Techniques, Industrial and Information Systems (ICIIS), 2010

[16] AK Dubey, SK Shandilya, Exploiting need of data mining services in mobile computing environments, International Conference on Networks (CICN),2010