PATTERN ANALYSIS THROUGH ACCESS LOGS AND ERROR LOGS USING HIVE WITH HADOOP

[1]Lokesh Narware, **[2]Ajay Phulre**
[1]Student of CSE, [2]Prof. & Head of CSE Dept
[1, 2]Shri Balaji Institute of Tech. & Management, Betul,
Email: [1]lokeshnarware@gmail.com, [2]aphulre@gmail.com

Abstract: *Web usage mining is application of data mining techniques to discover usage patterns from web data, in order to better serve the needs of web based applications. The user access log files present very significant information about a web server. This paper is concerned with the in-depth analysis of Web Log Data of website to find information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining. The obtained results of the study will be used in the further development of the web site in order to increase its effectiveness. Now a days due to increase on internet user logs files are also increase rapidly. Processing this explosive growth of log files using relational database technology has been facing a bottle neck. To analyze such large datasets we need parallel processing system and reliable data storage mechanism.. In this paper we present the hadoop framework for storing and processing large log files and also present the hadoop ecosystems fpr analysis purpose called hive. In this we can takes various kind of apache access log files and error logs which is analyzed by hive.*

Keywords: *Hadoop, web mining, logdata, bigdata, pattern analysis, hive.*

1. INTRODUCTION:

Log files [3][7] provide valuable information about the functioning and performance of applications and devices. These files are used by the developer to monitor, debug, and troubleshoot the errors that may have occurred in the application. Manual processing of log data requires a huge amount of time, and hence it can be a tedious task. The structure of error logs vary from one application to another. Since Volume, Velocity and Variety are being dealt here, Big Data using Hadoop is used. Analytics [2] involves the discovery of meaningful and understandable patterns from the various types of log files. Error Log Analytics deals about the conversion of data from semi-structured to a uniform structured format, such that Analytics can be performed over it. Business Intelligence (BI) functions such as Predictive Analytics is used to predict and forecast the future status of the application based on the current scenario. Proactive measures can be taken rather than reactive measures in order to ensure efficient maintainability of the applications and the devices. Log files are an example of semi-structured data. These files are used by the developer to monitor, debug, and troubleshoot the errors that may have occurred in an application. All the activities of web servers, application servers, database -servers, operating system, firewalls and networking devices are recorded in these log files.[5][6]

There are 2 types of Log files - Access Log and Error Log[4].

- Access Log records all requests that were made of this server including the client IP address, URL, response code, response size, etc.
- Error Log records all the details such as Timestamp, Severity, Application name, Error message ID, Error message details.

1.1. HADOOP:

The Apache Hadoop project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. 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 clusters nodes.

Apache Hive

Facebook created Hive for analyzing large datasets. It is a most widely adopted data warehousing application which can provide the Relational model and SQL interface. Hive infrastructure runs on the top of Hadoop. It mainly helps in providing summary of the data, query and analysis of the unstructured data. Since its incubation in 2008, Apache Hive is considered as standard for Batch and Interactive SQL workloads on data in Hadoop. The Hive tables are similar to relational databases, but the tables in Hive are made up of partitions.

1.2. APACHE PIG

Yahoo started Pig as a research project to focus on analysis of large datasets. It was designed in the style of SQL and also MapReduce. Pig is used with Hadoop in general. Pig Latin is a procedural language used by Apache Pig. The programmers use Pig script and execute the command in the grunt shell. It runs MapReduce programs when running the pig script in grunt shell. The major components of Apache pig are, Parser: Checks the syntax of the script. Optimizer: Carries out the plan of the script as push down. Compiler: Compiles the plan into MapReduce job. Execution engine: Execute the MapReduce jobs and finally Hadoop produce the results.

2. LITERATURE REVIEW:

According to [1], Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web document, hyperlink between documents, usage logs of web sites etc. Web usage mining is the process of applying data mining techniques to discover usage pattern from the web data. It is one of the techniques to find personalization of web pages[8]. The collection of web usage data gathered from different levels such as server level, client level and proxy level and also from different resources through the web browser and web server interaction using the HTTP protocol. But in the current scenario the number of online customer's increases day by day and each click from a web page creates on the order hundred bytes data in typical website log file. When a web user submits request to web server at the same time user activities are recorded in server side. These types of web access logs are called log file. Request information sent by the user via protocol to the web server which is recorded in log file. The logfiles are contains some entries like ip address of which computer making the request, the visitor data, line of hit, the request method, location and name of the requested file, the HTTP status code, the size of the requested file.

Log files can be classified into categories depending on the location of their storage that is web server logs and application server logs. A web server maintains two types of log files: Access log and Error log. The access log records all requests that were made of this server. The error log records all request that failed and the reason for the failure as recorded by the application. A log files have lot of parameters which are very useful to recognizing user browsing patterns. In [2], info technology provides utmost importance to process of knowledge. Some petabytes of knowledge isn't spare for storing great amount of knowledge. massive volume of unstructured and structured information that gets created from numerous sources like Emails, web logs, social media like Twitter, Facebook etc. the key obstacles with process huge information embody capturing, storing, searching, sharing and analysis. Hadoop permits to explore advanced information. it's Associate in Nursing open supply framework written in Java that supports parallel and distributed processing and is employed for reliable storage of knowledge. With the assistance of massive information analytics, several enterprises square measure ready to improve client retention, facilitate with development and gain competitive advantage, speed and scale back quality. E-commerce corporations study traffic on websites or navigation patterns to see probable views, interests and dislikes of an individual or a gaggle as a full looking on the previous purchases. during this paper, they compare some usually used information analytic tools.

2.1. 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 amounts 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 so by using using traditional tool and techniques are not feasible, and the tradition tool cannot handle the large amount of dataset or an unstructured data. For this reason, data mining needs pre-processing and analytic method for finding the value. Indeed, data mining is closely related with artificial intelligence and machine learning and so on. Scale of data management in data mining and big data is significantly different in size. Big data came out after solving the requirements and challenges of data mining.[12]

3. PROPOSED WORK:

For analyzing these large and complex data required a power tool, we are using hadoop which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

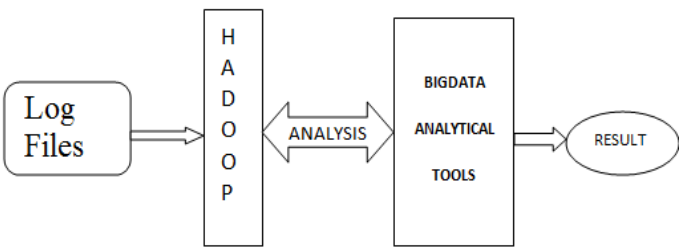

Figure1. Workflow Diagram

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on social media data sets. For analysing first we used standard platform as hadoop on single node ubuntu machine [9] to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling ,After that we can use bigdata analytical tools to refine such unstructured data and analyse the social data using bigdata analytical tools.

4. EXPERIMENTAL & RESULT ANALYSIS:

All the experiments were performed using an i3-2410M CPU @ 2.30 GHz processor and 3 GB of RAM running ubuntu 14 .And than we configure hadoop-1.1.2 on ubuntu and along with hadoop[11] we also integrate bigdata analytical tools hive and pig on top of the hadoop, So to achieve this we are going to follow the following methods:

- Loading Data into HDFS[10].
- Analyzing using Apache Hive.
- Compare with apache pig.

4.1. LOADING DATA INTO HDFS:

First we can loading different access and error log files in to HDFS, in our dissertation we can analyse two access log which are common access log and combined access log, And also we can analyze the error logs such as hadoop logs. Figure 2 shows the common access log which is loaded into HDFS, figure 3 shows the combined access logs and figure 4 shows the hadoop logs file. And in this figures we can clearly seen that there is not any structure between the data of these logs file. After loading these different logs file into HDFS.

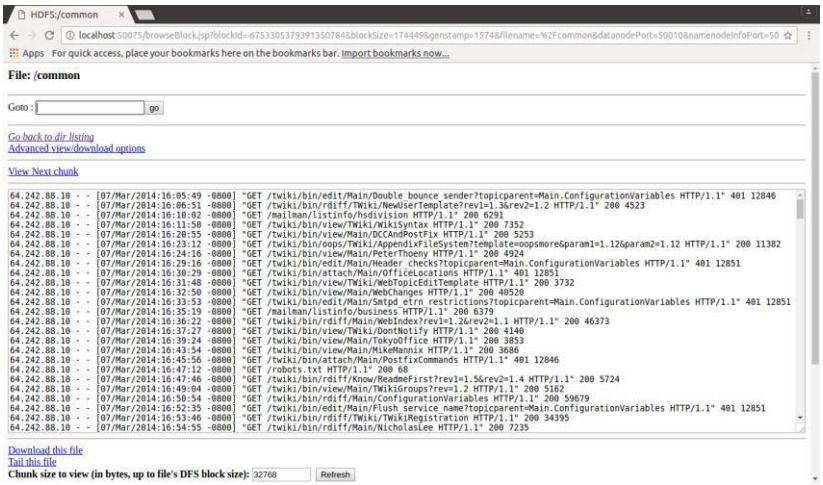

Figure 2 Loading common access logs into HDFS

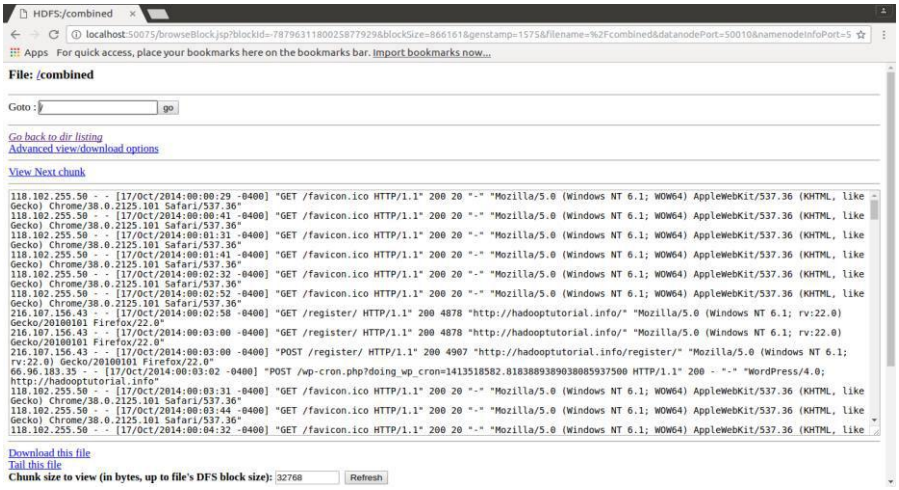

Figure 3 Loading
combined access logs into HDFS

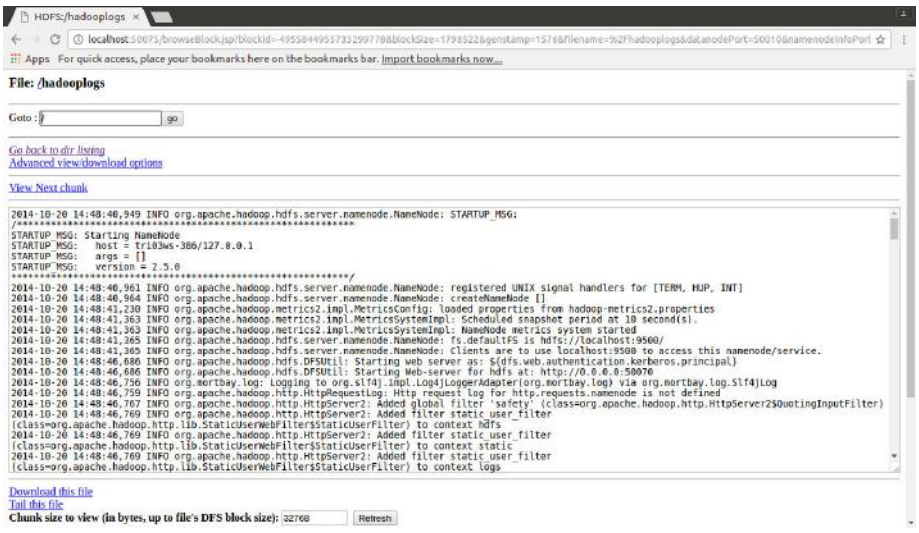

Figure 4. Loading Hadoop logs into HDFS

5. Analyzing using Apache hive:

5.1. Analyzing Common log file

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 can first create a common_log table to store the common_logs 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 unstructure data into structured format. First we can find the hit ratio from different host or ip address for that we can write a hive query which is shown in figure 5.

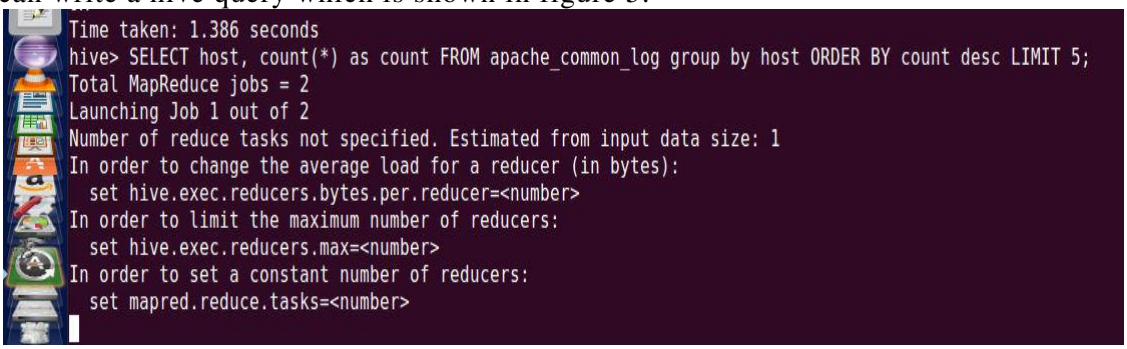

Figure 5. Launching query for Finding maximum hit ratio of ip address

For these hive query, hive engine launches a mapreduce job for pre-processing the log files. After finishing a execution of mapreduce job we can get the output of that query which is shown in figure 6.

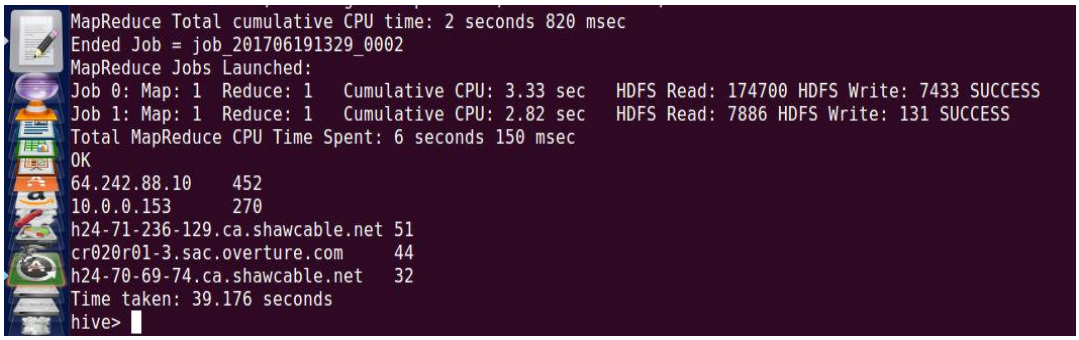

Figure 6. Maximum hits from ip addresses

After finding maximum and minimum hit count from the ip address , the next important field in the common log file is the status code which we can get from websites when accessing a web page, so for finding various status codes along with its frequency we can launch hive query. After finishing execution of mapreduce job we can get the various status code from common logs along with their count frequency which is shown in figure 7.

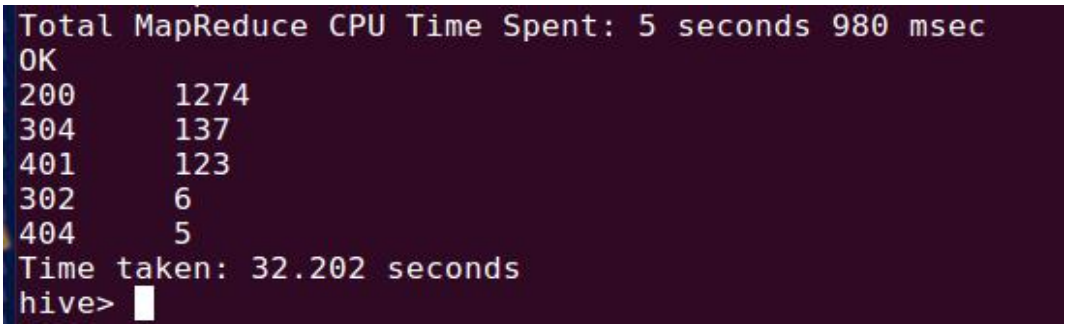

Figure 7. Status code along with its frequency

Analyzing Combined (Extended) log file

For analyzing combined log file we can first create a combined_log table to store the combined logs 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 unstructure data into structured format. Now we can find the ip address hit count and various status code as same as we find in common log file. The combined log files have a extra details like referrer page and the browser agent details so we can also perform some analysis on that part by launching a hive query and the result are shown in figure 8.

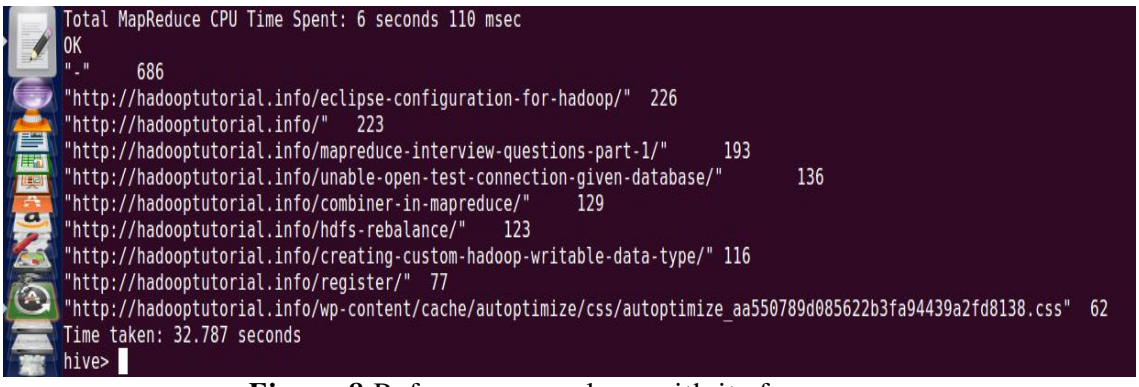

Figure 8 Referrer pages along with its frequency

5.2. Analyzing Custom logs (hadoop logs) files:

For analyzing combined log file we can first create a hadoop_log table to store the hadoop logs data efficiently in structured manner. After loading these file we can start analyzing by launching different hive queries by which we can find various error, warning and info message because by analyzing error message we can check the what error are coming in to hadoop application log file , And analyzing warn message we can verify the unauthorized activities or detect some warning message and by analyzing info message we can check the hadoop application is work well according to our configuration. The error message which we can detect are shown in figure 9.

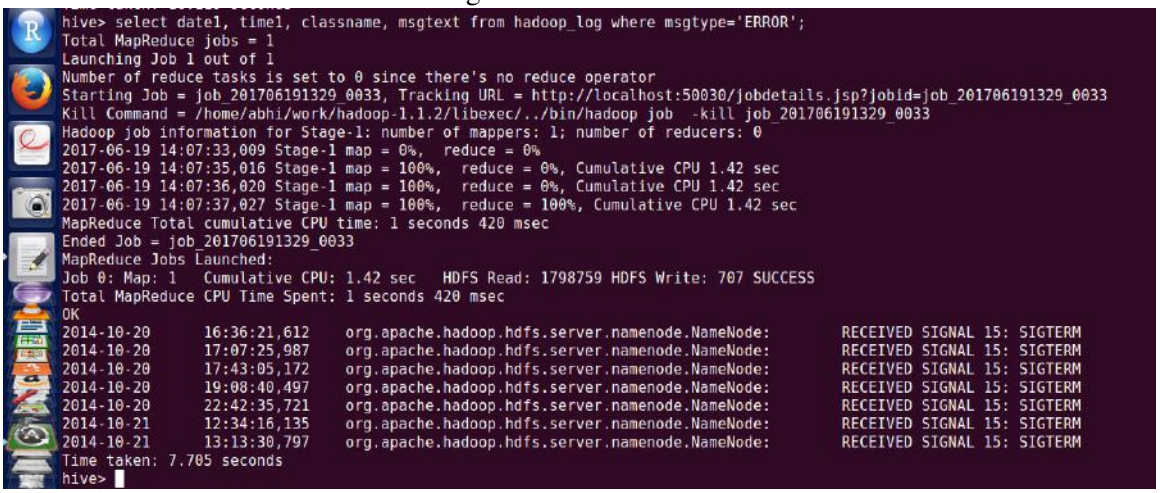

Figure 9. Finding various error comes in hadoop logs

COMPARISON

After analyzing the access logs and error logs using hive we can also analyse the access logs using pig which is an another hadoop ecosystem used for analytics purpose. In thiswe can compare performance of hive and pig by analyzing access log . First we can load the large logs file into the HDFS and than we can analyse using hive and the result and time taken by hive query are shown in figure 10.

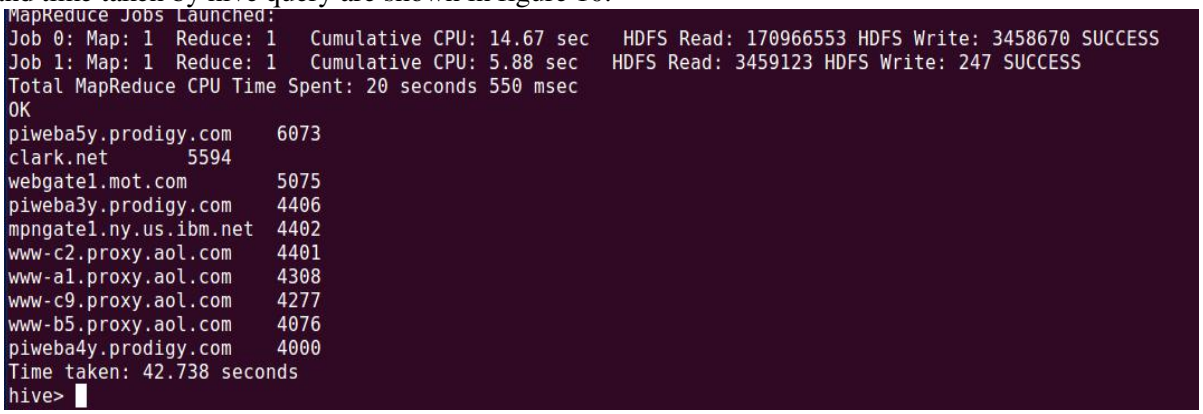

Figure 10. Time taken by hive query

After analyzing using hive we can analyse the logs data using pig by writing an pig script. An pseudo algorithm are shown below:

```
REGISTER /home/Desktop/piggybank-0.11.0.jar;
A = LOAD Data using piggybankloader;
B = GROUP A BY B;
C = FOREACH B GENERATE flatten($0), COUNT($1) as count;
D = order C by $1 DESC;
E = LIMIT D 10;
DUMP E;
```

After executing the pig script into the grunt shell we can get the output which is shown in figure 11.

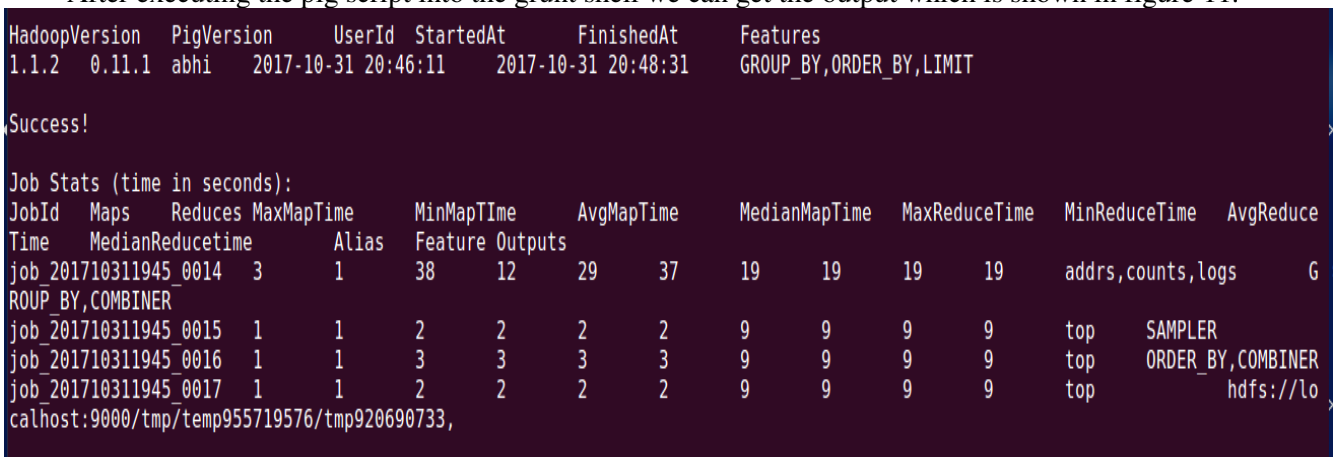

Figure 11. Time taken by pig

We can launched 3 query from hive and same as pig to compare the result on execution time taken from both the tools.

Execution time in seconds	Pig	Hive
Query-1	140	61.54
Query-2	127	47.72
Query-3	131	52.84

Table-1 Execution time taken by pig & hive

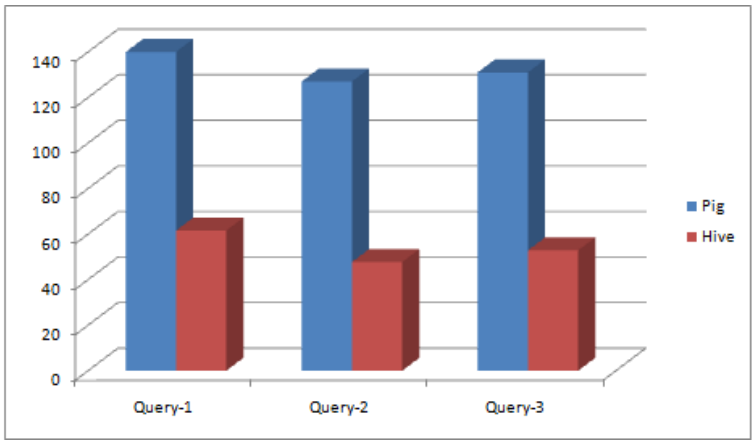

Figure 12. time taken by pig & hive

6. CONCLUSION:

The user access log files present very significant information about a web server. This paper is concerned with the in-depth analysis of Web Log Data of website to find information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining. To get categorized results of analysis hive query is written over MapReduce result. In this we can taken common access log file , combined access log file and also analyse the error logs using hive. And we can also compare the performance with 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. Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, “BIG DATA ANALYTICS: HADOOP AND TOOLS” in 2016 IEEE Bombay Section Symposium (IBSS), IEEE 2016.
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”.