Analyzing Web Access Logs using Spark with Hadoop

Vandita Jain
M.Tech(CSE)
LNCT, Bhopal (M.P.)

Tripti Saxena
Prof & Dept. of CSE,
LNCT, Bhopal (M.P.)

Vineet Richhariya, PhD
Professor & Head, Dept. of CSE,
LNCT, Bhopal (M.P.)

ABSTRACT
Web usage mining is a process for finding a user navigation patterns in web server access logs. These navigation patterns are further analyzed by various data minig techniques. The discovered navigation patterns can be further used for several things like identifying the frequent patterns of the user, predicting the future request of user, etc. and in the recent years there are huge growth in electronic commerce websites like flipkart, amazon, etc. with an huge amount of online shopping websites, it is necessary to notice that how many users are actually reaching to the websites. When user's access any online website, web access logs are generated on the server. Web access logs data helps us to analyze user behavior that contain information like ip address, user name, url, timestamp, bytes transferred. It is very meaningful to analyze the web access logs which helps us in knowing the emergency trends on electronic commerce. These ecommerce websites generates petabytes of log data every day which is not possible by traditional tools and techniques to store and analyze such log data. In these paper we proposed an hadoop framework which is very reliable for storing such huge amount of data in to HDFS and than we can analyze the unstructured logs data using apache spark framework to find user behaviour. And in these paper we can also analyze the log data using mapreduce framework and finally we can compare the performance on spark and mapreduce framework on analyzing the log data.

Keywords
Hadoop, HDFS, Mapreduce, Log analysis, spark, user behaviour.

1. INTRODUCTION
Log files [3] 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. Analytics [7] involves the discovery of meaningful and understandable patterns from the various types of log files.

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.

PURPOSE
A large number of log files [4] are generated by computers nowadays. A Log File is a file that lists actions that have taken place within the application or device. The computer is full of log files that provide evidence of what is going on within the system. Through these log files, a computer user will confirm what internet sites are accessed, United Nations agency accessed and from wherever it had been accessed. conjointly the health of the appliance and device is recorded in these files. Here area unit many places wherever log files is found:

- Operating systems
- Web browsers (in the shape of a cache)
- Web servers (in the shape of Access logs)
- Applications (in the shape of error logs)
- E-mail

Log files area unit Associate in Nursing example of semi-structured knowledge. These files area unit utilized by the developer to watch, debug, Associate in Nursingd troubleshoot the errors that will have occurred in an application. All the activities of internet servers, application servers, info -servers, software, firewalls and networking devices area unit recorded in these log files.

There area unit two forms of Log files - Access Log and Error Log. This paper discusses the Analytics of Error logs.

Access Log files contain the subsequent parameters – scientific discipline Address, User name, visiting path, Path traversed, Time stamp, Page last visited, Success rate, User agent, URL, Request kind.

1. Access Log records all requests that were made from this server together with the consumer scientific discipline address, URL, response code, response size, etc.
2. Error Log records all the main points like Timestamp, Severity, Application name, Error message ID, Error message details.

Error Log may be a file that's created throughout processing to carry knowledge best-known to contain errors and warnings. it's sometimes written when completion of process so the errors is corrected. Error logs contain the parameters such as:

- Timestamp (When the error got generated).
- Severity (Mentions if the message may be a warning, error, emergency, notice or debug).
- Name of application generating the error log.
- Error message ID.
- Error log message description

HADOOP
Hadoop is an open source, distributed computing framework developed and maintained by the Apache Software Foundation written in java.

In hadoop developers can deploy programs written in any other languages or in java for the processing of data parallely across multiple commodity machines despite of the fact that hadoop framework is written in java.

One of the key features of hadoop is that it partitions the computation and data across multiple nodes and then makes the

application computation run in parallel on these nodes. Important features of hadoop are redundancy and reliability which means that if any of nodes fails due to technical fault or other failures, it automatically creates a backup for that node without any intervention of the operator.

Depending on the process complexity the time of execution may vary from minutes to hours. Hadoop has emerged out to be a potential solution for number of applications in web log analysis, visitor behaviour, search indexes, indexing and analysis of text content, applications in biology, genomics and physics, machine learning researches and natural language processing researches and in all sort of data mining.

Spark

Apache Spark may be a lightning-fast cluster computing technology, designed for quick computation. it's supported Hadoop MapReduce and it extends the MapReduce model to expeditiously use it for additional varieties of computations, which incorporates interactive queries and stream process. the most feature of Spark is its in-memory cluster computing that will increase the process speed of associate application.

Spark is meant to hide a good vary of workloads like batch applications, unvaried algorithms, interactive queries and streaming. with the exception of supporting of these employment during a various system, it reduces the management burden of maintaining separate tools.

2. LITERATURE REVIEW

According to [1], net mining[13] is that the application knowledge mining techniques to extract helpful knowledge from net data that features net document, link between documents, usage logs of internet sites etc. net usage mining is that the method of applying data processing techniques to find usage pattern from the online knowledge. it's one in all the techniques to seek out personalization of sites. the gathering of net usage knowledge gathered from completely different levels like server level, shopper level and proxy level and additionally from completely different resources through the net browser and web server interaction exploitation the communications protocol protocol [12]. however within the current situation the quantity of on-line customer's will increase day by day and every click from an internet page creates on the order hundred bytes knowledge in typical web site log file. once a net user submits request to web server at identical time user activities square measure recorded in server facet. These kinds of net access logs square measure known as log file. Request info sent by the user via protocol to the online server that is recorded in log file. The logfiles [13]are contains some entries like science address of which pc creating the request, the visitant knowledge, line of hit, the request technique, location and name of the requested file, the communications protocol standing code, the dimensions of the requested file.

Log files may be classified into classes reckoning on the situation of their storage that's net server logs and application server logs. an internet server [11] maintains 2 kinds of log files: Access log and Error log. The access log records all requests that were product of this server. The error log records all request that unsuccessful and also the reason for the failure as recorded by the applying. A log files have ton of parameters that square measure terribly helpful to recognizing user browsing patterns [11].

Mining the web log file can useful to server and E-commerce for predicting the behavior of their online client. each day increasing on-line customers likewise as increasing the dimensions of net access log [10]..In giant websites handling variant synchronic guests will generate hundred of peta bytes of logs per day. the present data processing techniques store blog files in ancient software and analyze. RDBMS system cannot store and manage the peta bytes of heterogeneous dataset. So, to research such massive blog file with efficiency and effectively we'd like to develop quicker, economical and effective parallel and climbable data processing algorithmic program. additionally want a cluster of storage devices to store peta bytes of blog knowledge and parallel computing model for analyzing Brobdingnagian quantity of knowledge. Hadoop framework provide reliable clusters of storage facility to stay our giant blog file knowledge in an exceedingly distributed manner and parallel methoding options to process an oversized blog file knowledge with efficiency and effectively. The preprocessed net logs by HadoopMapReduce atmosphere is any processed for prediction of user next request while not worrisome them to extend the interest and to scale back the time interval with ecommerce system.

This paper shows the way to method log file exploitation MapReduce and the way Hadoop framework is employed for parallel computation of log files. knowledge collected from varied resources square measure loaded into HDFS for facilitating MapReduce and Hadoop framework.We well-tried that process massive knowledge with the assistance of Hadoop atmosphere ends up in minimum computation and time interval and additionally our HM_PP algorithmic program ends up in sensible accuracy in prediction of user most well-liked pages. So, one will simply access the ecommerce system with the assistance of massive knowledge analytics tools with less time interval and sensible prediction accuracy. In future log analysis may be done by correlation engines like RSA envision and HA cloud atmosphere. The higher than work may also be extended with linguistics analysis for higher prediction.

In [2], the author describes that massive knowledge analytics has attracted intense interest from all academe and trade recently for its arrange to extract data, info and knowledge type massive knowledge. massive knowledge and cloud computing, 2 of the foremost necessary trends that square measure process the new rising analytical tools. massive knowledge analytical capabilities exploitation cloud delivery models might ease adoption for several trade, and most significant thinking to value saving, it might alter helpful insights that might providing them with completely different styles of competitive advantage. several firms to produce on-line massive knowledge analytical tools a number of the highest most firms like Amazon massive knowledge Analytics Platform ,HIVE net based mostly Interface, SAP massive knowledge Analytics, IBM InfoSphere BigInsights, TERADATA massive knowledge Analytics, 1010data massive knowledge Platform, Cloudera massive knowledge answer etc. Those firms analyze Brobdingnagian quantity knowledge with facilitate of various sort of tools and additionally give straightforward or straightforward program for analyzing data.

3. PROBLEM DEFINITION

E-commerce websites like flipkart, snapdeal generates petabytes of log data every day. 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. Scale of data management in data mining and big data is significantly different in size. However,

the basic method to extract the value is very similar. Big data came out after solving the requirements and challenges of data mining[13].

4. PROPOSED WORK

For storing and analyzing these large and complex data we need a powerful tool [10], we introduces apache hadoop which is a open source framework for storing and processing large datasets.

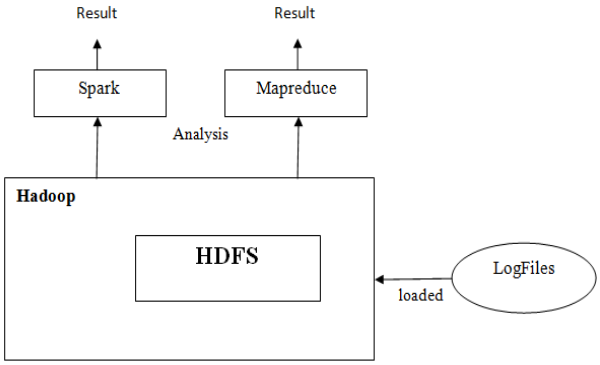

Figure1. Workflow Diagram

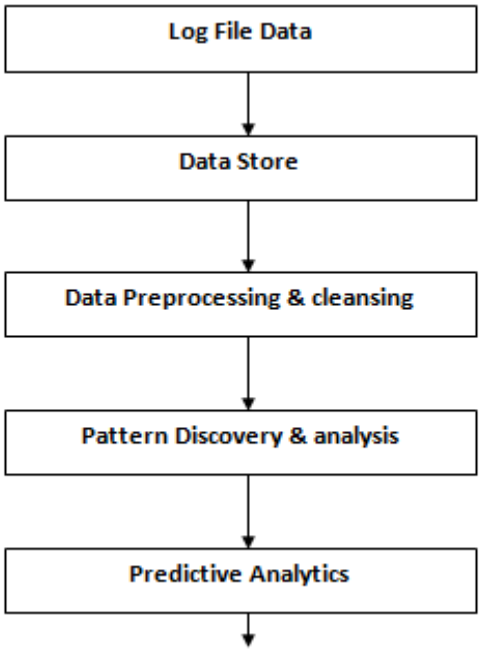

Figure 2. Analysis Steps

Our Steps or Algorithm Steps will follow:

Step 1: First collect the different types of log file for analysis, in our dissertation we collect common log file, combine log file and custom log file.

Step 2: All these three types of log file are unstructured in nature , first we can store these files into HDFS for analysis.

Step 3: these unstructured log files are first preprocess by which the unwanted data are removed, all the pre-processing is done by the spark framework which is running top of the hadoop.

Step 4: We can also analyse these data using mapreduce by writing hive query on top of the hadoop which launches a mapreduce job for hive query and find the page rank, maximum hit rate by the ip address , etc.

Step 5: Finally we can compare execution performance of these two framework spark and mapreduce.

5. EXPERIMENTAL & RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14 . As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

- Loading Data into HDFS.
- Analyzing using spark.
- Analyzing using Mapreduce.
- Compare the performance.

Loading Data into HDFS

First we can loading different access logs in to HDFS, Figure 3 shows the loading a log file into HDFS. 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 we can analyze using spark and mapreduce framework.

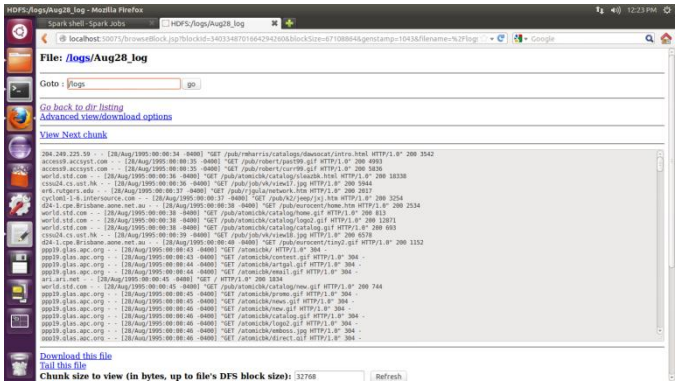

Figure 3. Web access logs stored into the HDFS

Analyzing Using Spark

After loading such huge amount of unstructured data into HDFS, we can analyse the data using spark which read the data from HDFS and preprocessing and analysis is done by spark. By these we can write the rdd script to find top ip address who has maximum hit ratio. The script are shown in figure 4.

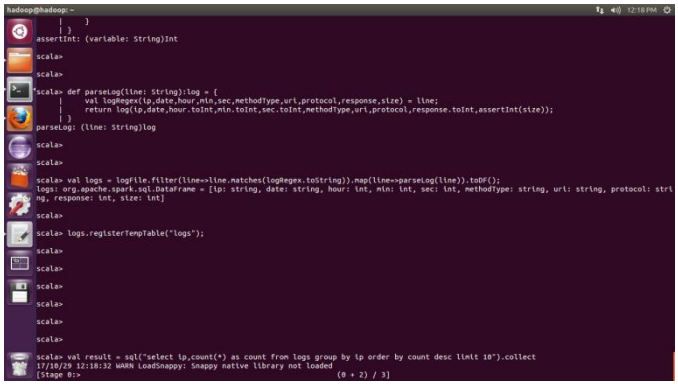

Figure 4. Processing log data using spark

After executing these , we can get the output , for analyzing we are using SQL in spark, because SQL is very easy to use for

analyzing the log records. And the result of these scripts are shown in figure 5.

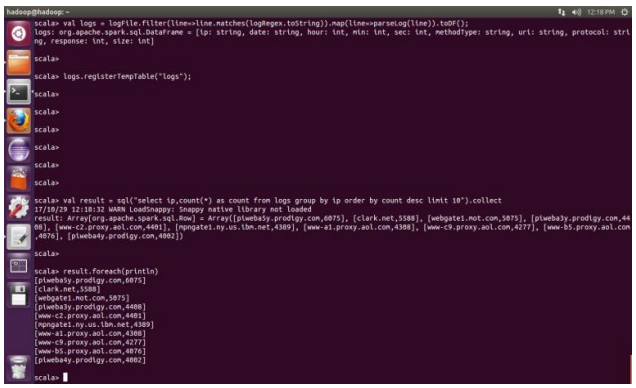

Figure 5. IP address along with its frequency

Similarly with the help of sql context in spark we can find the response statis code along with its frequency and the most frequent access page, and these result are shown in figure 6 and figure 7.

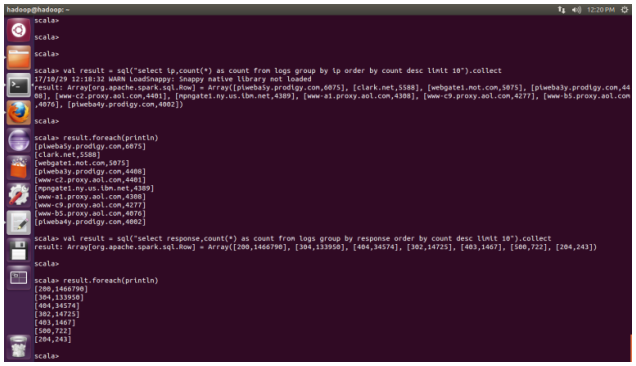

Figure 6. Response status code with frequency

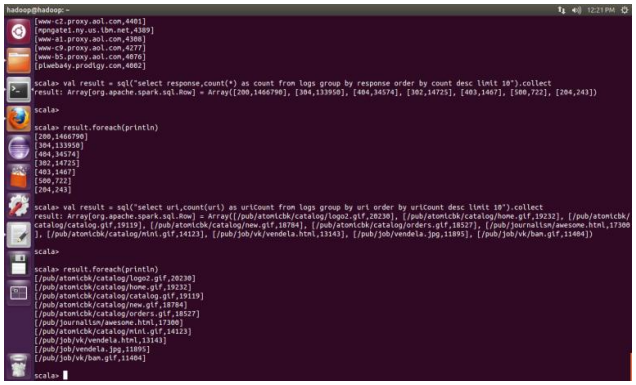

Figure 7. Most frequent access pages

Figure 8 shows the time taken by these three sql query to analyze the log data on spark framework.

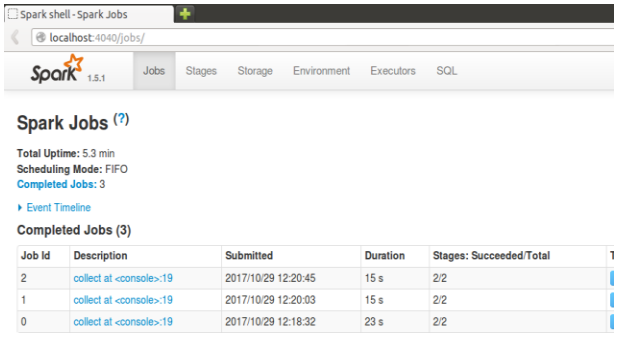

Figure 8. time taken by spark framework

Analyzing using Mapreduce

We can analyze the complex log data using mapreduce framework. And for sql like interface we can integrate apache hive on top of the mapreduce, we can write sql query on hive and hive compiler validate the sql query and convert sql query into mapreduce program which is executed by mapreduce framework and we can get the output. Figure 9 shows the hive query and the mapreduce job which is launched by hive.

Figure 9. Mapreduce job is launched by hive query

After the execution is done by mapreduce framework we can get the output through mapreduce ans figure 10 shows the result of mapreduce job.

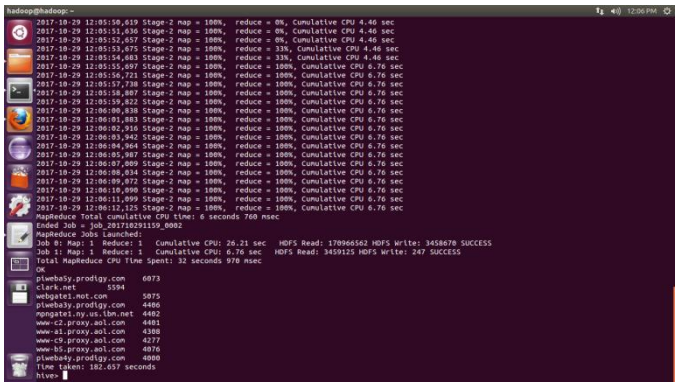

Figure 10. output is generated from mapreduce

Similarly we can find the response status code and the most visited web pages using mapreduce framework and we can get the same result as spark result, both the framework are accurate in processing log data. The time taken by mapreduce framework are shown in figure 11.

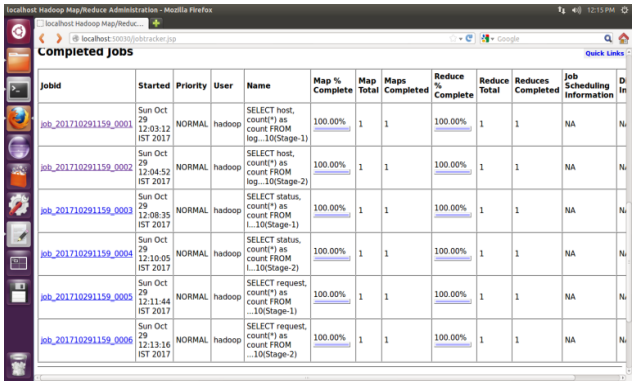

Figure 11. Time taken by mapreduce

Comparison between Spark and Mapreduce

After analyzing the complex log data using spark and mapreduce framework we can say that both the framework are very accurate in processing log data but differ in execution time, table 1 shows the execution time taken by mapreduce and spark framework.

Table-1 Execution time taken by spark & mapreduce

Execution time (in seconds)	Mapreduce	Spark
Query-1	32.9	23
Query-2	31.5	15
Query-3	33.4	15

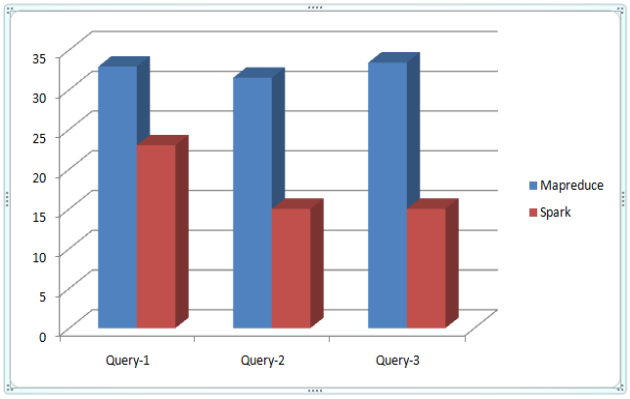

Figure 12. Execution time taken by spark & mapreduce

6. CONCLUSION

World Wide Web has necessitated the users to make use of automated tools to find desired information resources and to follow and access their usage pattern. We have presented best fit Spark and MapReduce programming model for analyzing web application log files . In this paper, data storage is provided using HDFS and Spark and MapReduce model applied over log files gives analyzed results in minimal response time. To get categorized results of analysis hive query is written over MapReduce result. Finally we can compare the performance comparison between spark and mapreduce framework for analyzing log data and we can say that spark holds better efficiency as compared to mapreduce.

7. 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] G.S.Katkar, A.D.Kasliwal, "Use of Log Data for Predictive Analytics through Data Mining", Current Trends in Technology and Science, ISSN: 2279-0535. Volume: 3, Issue: 3(Apr-May 2014).

[4] Savitha K, Vijaya MS, "Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies", IJACSA, Vol. 5, 2014.

[5] McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011, http: //www.mckinsey.com/.

[6] White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel corporation.

[7] Qureshi, S. R., & Gupta, A, "Towards efficient Big Data and data analytics: A review", IEEE International Conference on IT in Business, Industry and Government (CSIBIG),March 2014 pp-1-6.

[8] http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS.

[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] Chuck Lam, "Hadoop in Action", Manning Publications.

[11] Harish Kumar B T, Dr. Vibha L, Dr. Venugopal K R, "Web Page Access Prediction Using Hierarchical Clustering Based on Modified Levenshtein Distance and Higher Order Markov Model" in 2016 IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia

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

[13] 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