Real-Time Web Log Analysis and Hadoop for Data Analytics on Large Web Logs

Mayur Mahajan[1], Omkar Akolkar[2], Nikhil Nagmule[3] , Sandeep Revanwar[4],

[1234] BE IT, Department of Information Technology, RMD Sinhgad School of Engineering, Pune, Maharashtra, India.

---***---

Abstract - ***Organizations from the public and private sector are making a strategic decision to use the web log generated to gain competitive advantage. The main hurdle is to process this huge data efficiently for analytics purpose. It can be achieved by mining of this data. Web Log analyser is a tool used for finding the statics of web sites. Our project aims at implementing the web log analyzer for handling exception and errors. The web log data will be of unstructured form having XML data. Through Web Log analyzer the web log files are uploaded into the Hadoop Distributed Framework where parallel procession on log files is carried in the form of master and slave structure. Various association rule mining algorithms available will be applied onto this web log. The results of each implementation of algorithms would be noted. Finally, the noted results will be analyzed to find out the most suitable algorithm in terms of time and computational efficiency. The web log will be split up using Hadoop and then performed upon. Hadoop is the core platform for structuring big data and also used for analytics purpose. It will be a deep study of these algorithms along with the use of Hadoop during this process. This paper discuss about these log files, their formats, access procedures, their uses, the additional parameters that can be used in the log files which in turn gives way to an effective mining and the tools used to process the log files. It also provides the idea of creating an extended log file and learning the user behaviour. Also this paper presents the details of the working of by web log analyzer. In addition to this, Recommendations are given to administrator for the improvement of website.***

Key Words: web log files, Hadoop, real-time analysis, map reduce, goal conversion, recommendations.

1. INTRODUCTION

In today's world internet usage has been increased day-by-day. Data is captured from different sources such as sensor data, web logs of search engine, social media sites, digital pictures and videos, purchase transaction records and cell phones GPS signals. To handle such type of data is a tedious job. This unstructured data is Big Data. Hadoop is core platform for structuring Big Data and it solves the problem of making it useful for analytics purposes. Hadoop's parallel processing of data is major advantage for analytics purpose . Web log analyser is a tool which is used for analytics of web logs which are generated at server side and for finding statistics of a particular web site. Web log analyser is a fast and a powerful tool for log analysis [5]. It gives information about site visitors and their user activities, accessed files, paths in the sites, user navigation between pages, various browser statistics, operating system statistics[6]. It generates easy to read reports which includes pie charts, bar graphs, text information. Reports are in the form of HTML ,PDF and CSV formats. The web server supporting dynamic HTML reports is also included[5] . The Browser makes the requests for data to Web server, which then uses HTTP to deliver the data back to the browser that had requested the web page [4]. Then the browser converts and/or formats, the files into a page which can be viewed by the user [7].

The much needed robustness and scalability option to a distributed system is provided by Hadoop, which also provides inexpensive as well as reliable storage. Tools for analysing structured and unstructured data are also provided by Hadoop. So, Map Reduce and HDFS of Hadoop use simple & robust techniques for delivering very high availability of data and for analysis enormous amounts of information quickly. It may not be possible to convert all the sequential algorithms into the parallel form, which in turn could be converted to the map/reduce format. There could emerge circumstance that the calculations may not be viably actualized in the guide/lessen organization, or usage of calculations could require more prominent overheads and not adjust for the points of interest Hadoop gives, along these lines making execution on Hadoop a terrible decision. Affiliation guideline mining is a sort of information mining process. Affiliation standard mining is done to separate fascinating connections, designs, relationship among things in the exchange database or other information archives. For instance an affiliation guideline natural product => milk created from the exchange database of a supermarket can help in planning promoting system around the standard. Affiliation standards are generally utilized as a part of different ranges, for example, telecom systems, promoting and hazard administration, and stock control and so on. Numerous organizations and firms keep huge amounts of their everyday exchange information. These information could be dissected to take in the buying pattern of the client. Such important understanding can be utilized to bolster assortment of business-related applications, for example, advertising and advancement of the items, stock administration and so on.

2. RELATED WORK DONE

Contents of Log Files:

The Log records in various web servers keep up various sorts of data [8]. The essential data present in the log record is as per the following.

•IP Address/User name: This recognizes who had gone by the site. The recognizable proof of the client is for the most part by utilizing the IP address. This might be an impermanent location that had designated. In this way the novel distinguishing proof of the client is not accomplished. In some sites the client recognizable proof is made by getting the client profile and permits them to get to the site by utilizing a client name and secret word. In this sort of access the client is being distinguished extraordinarily so that the return to of the client can likewise be recognized.

•Time stamp: The time spent by the client in every page while surfing through the site. This is distinguished as the session.

•Page went to in conclusion: The page that was gone to by the client before he or she leaves the site [8].

•Success rate: The achievement rate of the site can be dictated by the quantity of downloads made and the number duplicating action under passed by the client [9].

•User Agent: This is only the program from where the client sends the solicitation to the web server. It's only a string portraying the sort and form of program programming being utilized.

•URL: The asset got to by the client. It might be a HTML page or a script.

•Request type: The method used for information transfer is noted. The methods like GET, POST.

The paper [3] includes brief description of how log files are generated (location), contents in a log file (username, visiting path, path traversed, timestamp, page last visited, etc.). Experimental Results gives us detailed description of types of web server logs i.e., from where the log are generated, status codes sent by the servers to the clients. The overall process of web mining in this paper is useful for extraction of information from World Wide Web. This includes three major steps: 1.preprocessing, 2.pattern discovery, 3.pattern analysis. It does not provide the idea about concept of learning the user's area of interest.

The paper [2] discusses about the approach E-Web Miner and is the proposed web mining algorithms that removes the flaws of Improved Apriori-All algorithm and improves upon the time complexity of the earlier Apriori-All algorithm. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines correct result of candidate set whereas the Improved Apriori-All algorithm fails to deliver the correct result. An efficient web mining algorithm for web log analysis which can be traced for its valid results and can be verified by computational comparative performance analysis. The most criticized ethical issue involving web usage mining is the invasion of privacy.

The paper [1] introduces the Web Page Collection Algorithm that uses cluster mining in order to find a group of connected pages at a web site. The proposed algorithm takes web server access log as input and maps it into form clustering. Afterwards the cluster mining is applied to the output data. Finally the output is obtained by mining the web user's logs. This paper presents strategy for automatic web log information mining that has been proved to be more effective. There needs to be information integration of content knowledge and knowledge extraction from the various web sites.

3. PROPOSED METHOD

In this paper, we enhance our system for analysis. Specifically, we present an advanced scheme to support faster analysis of real time web logs generated by various web sites with parallel processing using Apache Hadoop. In this way, the users can be provided with browser wise statistics, region wise statistics, client wise statistics, etc. Web log analysis demonstrates the analysis in terms of bar graphs, pie charts and reports in the form of PDF, etc. as the definitions specified in the proposed analysis model. Recommendations will help the administrator for the improvements of website in various locations on various factors.

1) This project is basically deals with the analysis purpose. Using analysis report developer can fix the bugs in the website.

2) We present an advanced scheme to support stronger analysis by analysing the real time web log generated file with Apache Hadoop.

3) The report also contains errors and exceptions and the no of times they have occurred. This tool saves the important time of developer to analysis the logs.

3.1 SYSTEM ARCHITECTURE

Working of the system with respect to architectural component is explained below:

Third Party Web-site:

This section covers the users of the system, i.e., the admin of the web-sites who want to analyze the performance and the logs of their web-sites. There are multiple users of this third-party web-site which are responsible for the generation of the logs. Admin of a particular web-site have to register on the analyzer web-site, i.e., the system where the processing will be done.

Web Log Analyzer:

This is the main part of the architecture, i.e., the web-site for the analysis of web-logs. First of all, after admin of third party site logs in to the system, the web-site verification will be done so as to provide security to the analysis of logs.

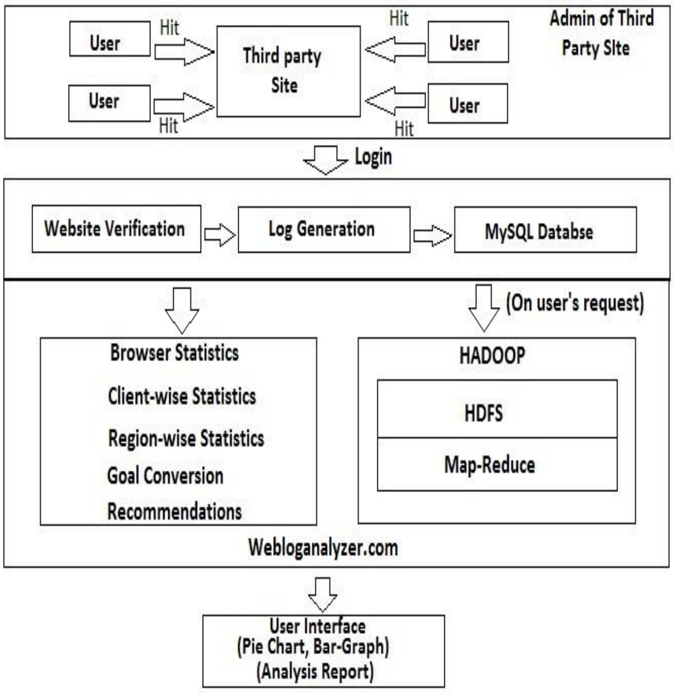


Fig 1. System Architecture

Admin will have to copy paste the scripting code to his web-site which is present on the system and will be unique. After security check, the log generation will start and will be pushed into the MySQL database for the further processing. System then will do the analysis and will generate Browser-wise, Client-wise and Region-wise statistics. Goal conversion rate will also be calculated based on which performance recommendations will be provided. This analysis will be real-time, i.e., the analysis will be done in parallel with logs generation.

If the admin wants statistics of particular period such as a particular month, then on the request of the admin of web-site the system will generate a text file which will be an input to Hadoop framework.

Hadoop Distributed File System (HDFS) will store this file and this file then will be analyzed by Map-Reduce method. The results will again be stored in text file and will be returned to the system.

User-interface:

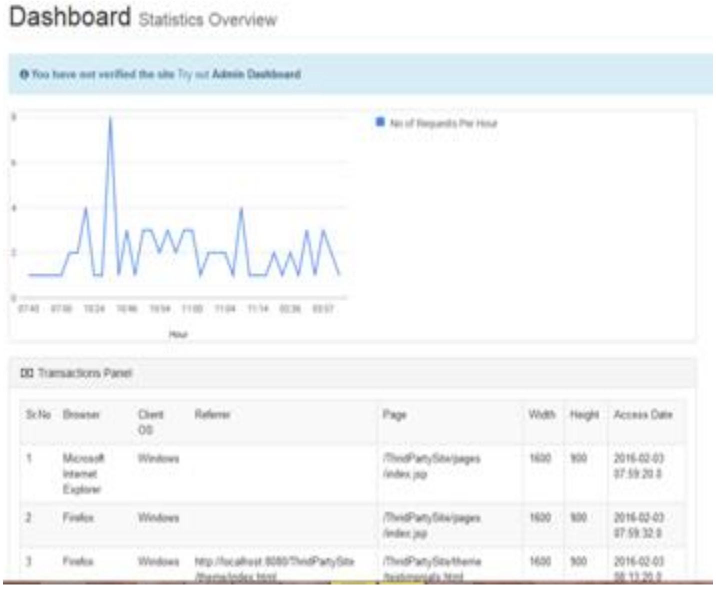
The web-site of analyzer will display the analysis report in the form of Pie-charts, Bar-graphs. The recommendations for the performance of web-site and the goal conversion rate will be displayed to the user via user interface. The report generated on the request of admin of third party site for a certain period (Hadoop output text file) will be displayed in user-understandable form through this interface.

4. EXPERIMENTS AND RESULTS

In this section, we show the study results from different sources that the feasibility, speedup, validity and efficiency of real-time analysis and Hadoop analysis. The data set is processed and statistics about the log data are generated. These statistics are made available to the user interface where they are displayed on a button click. The line graphs, pie-charts and bar graphs are used to display the results of the analysis. Following shows the statistics generated after processing log data sets.

1.Real-time Line Graph:

The real time graph generation based on the analysis of the generated graphs is shown by the system as Line graphs. This line graph represents the statistics of the number of requests coming to the site in particular time intervals. The generated logs have the parameter as Access Date which will show the time as well as the date of the user's request to the site through which the time of user's request can be accessed for the generation of the graph.
On click, the graph generated will show the exact number of requests at that particular time. This line graph will be Time vs. Number of request and will be in user-understandable form.

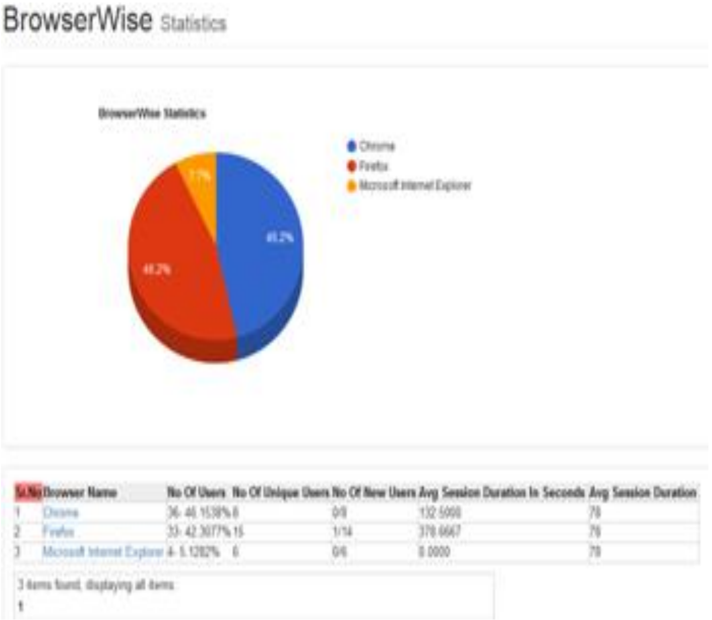


Dashboard Statistics Overview

You have not verified the site Try out Admin Dashboard

Transactions Panel

Sr.No	Browser	Client OS	Referer	Page	Width	Height	Access Date
1	Microsoft Internet Explorer	Windows		/ThirdPartySite/pages /index.jsp	1600	900	2016-02-03 07:59:20.0
2	Firefox	Windows		/ThirdPartySite/pages /index.jsp	1600	900	2016-02-03 07:59:32.0
3	Firefox	Windows	http://localhost:8080/ThirdPartySite /theme/index.html	/ThirdPartySite/theme /testimonials.html	1600	900	2016-02-03 08:13:20.0



BrowserWise Statistics

Sr.No	Browser Name	No Of Users	No Of Unique Users	No Of New Users	Avg Session Duration In Seconds	Avg Session Duration
1	Chrome	36- 46.1538%	8	0/8	132.5000	78
2	Firefox	33- 42.3077%	15	1/14	378.6667	78
3	Microsoft Internet Explorer	4- 5.1282%	6	0/6	0.0000	78

3 items found, displaying all items
1

2. Pie-charts and bar graphs:

Generation of Browser-wise, client-wise and region-wise statistical analysis will be displayed to the user of the system in the form of pie-charts or bar graphs. These graphs display the statistics of different browsers, different clients or new users of the web-site and the locations of the clients from where they are accessing the website.

5. CONCLUSION

This Paper describes a detailed view of Hadoop framework used to process big data . This paper describes how a log file is processed using map reduce technique. Hadoop framework is used as it is beneficial for parallel computation of log files.The system will analyze the log files and present it to the user in more user understandable form such as pie-charts and bar-graphs. And the system will give the recommendations to the user for improvement of website for the popularity of it. In the future work, Real time web log analysis can be done. The fluctuations can be shown in reports online as per the changes of log.

6.ACKNOWLEDGEMENT

We take this opportunity to thank our project guide Ms. Snehal Nargundi and Head of the Department Ms. Sweta Kale for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report.

We are also thankful to all the staff members of the Department of Information Technology of RMD Sinhgad School of Engineering, Warje for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

7.REFERENCES

[1] R.Shanthi , Dr.S.P.Rajagopalan "An Efficient Web Mining Algorithm To Mine Web Log Information", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 1, Issue 7, September 2013

[2] M.P. Yadav, P.Keserwani "An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner",978-1-4577-0697-4/12/$26.00 ©2012 IEEE .

[3] L.K. Joshila Grace, V.Maheswari , Dhinaharan Nagamalai , "ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING" , International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[4] Paul Hern´andez, Irene Garrig´os, and Jose-Norberto Maz´on "Modeling Web logs to enhance the analysis of Web usage data", Lucentia Research Group, Dept. Software and Computing Systems, University of Alicante, Spain, 2013 Workshops on Database and Expert Systems Applications.

[5] M. Zaharia, D. Borthakur, J. S. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Job scheduling for multi-user map reduce clusters,"EECS Department, University of California, Berkeley, Tech. Rep., Apr 2009.

[6] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. Journal of the Royal Statistical Society B, pages 155–176, 1996 J. Dean and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters". Commune. ACM, 51(1):107–113, 2008.

[7] J. Dean, S. Ghemawat, *"MapReduce: Simplified Data Processing on Large Clusters,"* In Proc. of the 6th Symposium on Operating SystemsDesign and Implementation, San Francisco CA, Dec. 2004.

[8] R. J. Williams, D. E. Rumelhart, G. E. Hinton. Learning representation by back-propagating errors. InNature, volume 323, pages 533–536,1986.

[9] Daryl Pregibon. Logistic regression diagnostics. In The Annals of Statistics, volume 9, pages 705–724, 1981.

R. L¨ammel. Google's MapReduce Programming Model – Revisited. Draft; Online since 2 January, 2006; 26 pages, 22 Jan. 2006.