Web Surfer Tracking using Big Data Technologies

B. B. Jayasingh,
Department of IT, CVR College of Engineering, Ibrahimpatan, Hyderabad, India
Email: bbjayasingh9@rediffmail.com

Abstract- **Log files are semi-structured on the web server-side. Computer generates huge flat text files as log. Therefore, Hadoop file systems are suitable to store as text files. The MapReduce techniques of the Hadoop works well in distributed clusters as it process one line at a time. NASA web server log is considered as input to system in order to perform the session identification task. A Statistical report is produced based on total count of visit per hour, per day, and per date. R language is used in order to find the user sessions and analyzed rigorously. In this paper, the distributed cluster is considered in Hadoop for the session identification in the web log. The analyst loads the log file and does analysis using MapReduce and generating statistical reports. The system developed to analyze the traffic from the log file based on various parameters, such as hours of a day, days of a week, days in a month etc. The system generates the detailed information about the location of a user that includes the country, state, city, zip code and also the co-ordinates of the location (latitude and longitude) by making use of the IP address field of a web server log.**

Index terms – **Web log, Web Mining, Big Data, Hadoop, Map Reduce, R Tool.**

I. Introduction

Web mining is the popular research area as the web generates a huge unstructured data lead to big data. The complexity of the web data requires effective search tools to find relevant information easily and precisely. The amount of information available in the web log is difficult to predict without a tool. How to predict the users' behavior is a matter of concern for the service provider and how to reduce the load on traffic as the personal information are more. The service providers must design the website which must accommodate various independent users.

The importance of user data is well known to companies like Facebook and Twitter though they have the biggest and fastest growing data repository in the world. Their size and diversity is required to be mined for useful information. The algorithms and the data analysis tools for inspecting the data are needed for decision making. The decision makers extract the useful information by considering specific digital footprints. The digital footprints are the representative of collective user behavior. The decision makers analyze a large amount of these traces to find common patterns, make better predictions, build smarter products, extract user models, and gain a better understanding of the dynamics of human behavior.

The data is surely growing in size, but also in complexity as it shows up in different formats and from different sources that are hard to integrate, and in dynamicity as it arrives continuously, changes rapidly and needs to be processed as fast as possible. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly. Another method is putting data in-memory but using a grid computing approach, where many machines are used to solve a problem. Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time. On the other hand, it is a challenge [9]. Current methodologies are often not suitable to handle huge datasets, so new solutions are needed.

The World Wide Web is an interdisciplinary part of human life thereby the volume of data becomes larger and larger time to time. Big data [2] is when the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam. big data is data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time. This means that we can call big an amount of data that forces us to use or create innovative methodologies. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

In the business intelligence, Big data can handle petabytes or terabytes of data in a reasonable amount of time. Big data uses Hadoop framework for data intensive distributed applications. The main principle of hadoop is moving computations on the data rather the moving data for computation. Hadoop is used to breakdown the large number of input data into smaller chunks and each can be processed separately on different machines. To achieve parallel execution, Hadoop implements a MapReduce programming model. MapReduce a java based distributed programming model consists of two phases: a massively parallel "Map" phase, followed by an aggregating "Reduce" phase. MapReduce is a programming model and an associated implementation for processing and generating large data sets [10].

This paper uses the web logs of NASA website accessed by the user which is freely available, to identify the session. It also applies Hadoop framework in turn MapReduce technique to process the web log. The file size of the log is 550MB and the dataset are collected from various time period of the same year. The identified session is analyzed based on hour, day, date and number of times visited using R tool.

DOI:10.32377/cvrjst0812

II. Literature Survey

There are variety of Web logs in the world and The data has been so large that it becomes difficult to analyze it with the help of our traditional mining methods. Therefore the Big data term has been introduced that exceeds the processing capability [11]. Now, Data mining techniques are replaced with Big Data Mining to discover the usage patterns from the logged data. To understand customer behavior, evaluate the effectiveness of a particular website and the user, it is now big data mining techniques that plays an important role. It has three main key characteristics (1) volume (2) velocity and (3) variety. Therefore, Volume is the size of data which is now larger than terabytes and petabytes. So it is very difficult to analyse using conventional methods due to large scale. The Velocity is the pre-defined period of time to mine large amount of data using big data technology. The traditional methods of mining are not an appropriate solution as it may take huge time to mine such a volume of data. The Variety is the various heterogeneous sources the data comes from and forms a Bigdata. But the Bigdata is designed to handle structured, semi-structured as well as unstructured data which is not designed in the traditional methods that handle only structured data and not such large volume.

Hadoop [4] proposed the smart miner framework that extracts the user behaviors from web log. Web log contains data not only structured traditional relational data, but also semi-structured and unstructured data come from a variety of sources and in a variety of types. There are more focused research towards web log mining using big data technologies. Such a framework [5] is used the smart session construction to trace the frequent user access paths.

A traditional data warehouse cannot accommodate the data generated by machine such as click stream logs, email logs. Though the data are of unstructured that are larger in volume in comparison with human generated data. So Big data is the only alternative for the recent trend to store and analyze. The work embodied in this paper [6] is a generic log analyzer framework for different kinds of log file. The other work proposes the train model where the data is stored in HDFS and the test model categories the text document is the Parallelization of Genetic

Algorithm (PGA) is suggested [7] that uses OlexGA package for classifying the document.

There are two kinds of node in a cluster where one node acts as both master and slave and the other node acts as a slave and the data is transferred in 100Mb/s speed. There are five daemons such as namenode, datanode, jobtracker, tasktracker and Secondary namenode that contains a master node. A slave node contains tasktracker and datanode daemons and the master node contains the IP address of the slave. The slave node is identified by the master using its ip address. The master node read the log files that are of block size of 64MB, so it is default to save in hard drive. A framework for unstructured data analysis was proposed [8] by using big data of public tweets from twitter. The tweets are stored in Hbase using Hadoop cluster through Rest Calls and text mining algorithms that are processed for data analysis.

III. Hadoop Framework

Hadoop framework [11], In pseudo distributed mode contains all the five daemons run on local machine simulating a cluster. It process mostly unstructured text files, so the text files are generated and stored in the HDFS after applying data cleaning step. The cleaned web log data is used to analyze the session identification, unique user and unique URLs. HDFS stores large files across multiple machines typically in the range of gigabytes to terabytes. It achieves reliability by replicating the data across multiple hosts. Hadoop implements a computational paradigm known as MapReduce.

MapReduce [10] is a computational paradigm designed to process very large sets of data in a distributed fashion. The model is based on the concept of breaking the data processing task into two smaller tasks of mapping and reduction. During the map process, a key-value pair in one domain is mapped to a key-value pair in another pair, where the ‘value’ can be a singleora list of multiple values. The keys from the mapping process are then aggregated and the values for the same key combined together. This aggregated data is then fed to the reducer (one call per key) and the reducer then processes this data to produce a final value. The list of all final values for all the keys is the result set.

The key issue in breaking a problem into the MapReduce model is that the map and reduce operations can be performed in parallel on different keys, without the results of one operation affecting the other. This independence of results allows the map/reduce tasks to be distributed in parallel to multiple nodes, which can then perform the respective operations independent of each other. The final results are then aggregated together to produce the final result list.

IV. Web Log Analysis

User logs are collected by the web server and typically include IP address, page reference and access time. Mining web data provides a lot of information, which can be better understood with visualization tools. This makes concepts clearer than is possible with pure textual representation. Hence, there is a need to develop tools that provide a graphical interface that aids in visualizing results of web mining. Some of the most prominent technologies are [3] NoSQL database.

Finding hidden patterns in the large database needs analytical techniques. There are so many software tools available for predictive analysis including big data analytics to find useful information. The recent technologies focus on big data to analyze the logged data to track the user’s behavior. We try to track the user behavior by analyzing the logged datasets that are in semi structured fashion maintained by Hadoop. We took a log

DOI:10.32377/cvrjst0812

record as an example and analyzed all the fields incorporated in the table given herewith.

An example of log record is 125.125.125.125 - uche [20/Jul/2008:12:30:45 +0700] "GET /index.html HTTP/1.1" 200 2345

Field name	Example value	Description
host	125.125.125.125	IP address or host name of the HTTP client that made the request
identd	-	Authentication Server Protocol (RFC 931) identifier for the client; this field is rarely used. If unused it's given as "-".
username	uche	HTTP authenticated user name (via 401 response handshake); this is the login and password dialog you see on some sites, as opposed to a login form embedded in a Web page, where your ID information is stored in a server-side session. If unused (for example, when the request is for an unrestricted resource) it's given as "-".
date/time	20/Jul/2008:12:30:45 +0700	Date then time then timezone, in the format [dd/MMM/yyyy:hh:mm:ss +-hhmm]
request line	3ET /index.html HTTP/1.1"	The leading line of the HTTP request, which includes the method ("GET"), the requested resource, and the HTTP protocol version
Status code	200	Numeric code used in the response to indicating the disposition of the request, for example to indicate success, failure, redirect, or authentication requirement
bytes		Number of bytes transferred in the body of the response

V. Output Visualization

The system can be used to visualize the traffic from a log file based on various parameters, such as hours of a day, days of a week, days in a month etc. The system generates the detailed information about the location of a user, who requests the web site for a URL. The information includes the country, state, city , zip code and also the co-ordinates of the location(latitude and longitude) by making use of the IP address field of a web server log. The system provides the user with the option of visualizing the results of log analysis. The system plots the geolocation of all the ip addresses present in the log file on a world map using R programming. This can be used to analyze the traffic distribution based on the location of the user. The system generates line graphs to visualize the traffic distribution for hours of the day, days of a week and days of a month. These graphs can be used to detect patterns in traffic distribution and those patterns can be applied for business purposes.

As more and more businesses are discovering, data visualization is becoming an increasingly important component of analytics in the age of big data. Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information. So we use a free statistical computing tool and plot the graphs that are through R language.

The following R script in Fig. 1 takes the hours and requests as input and generates a line plot with hours on x-axis and requests on y-axis.

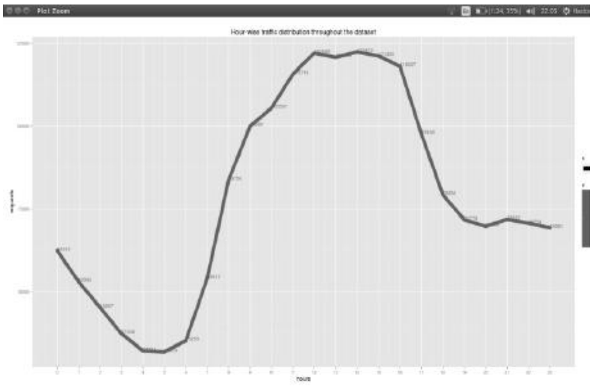

Fig. 1 Traffic distribution for hours of the day

The following R script in Fig. 2 takes the dates and requests as input and generates a line plot with dates on x-axis and requests on y-axis

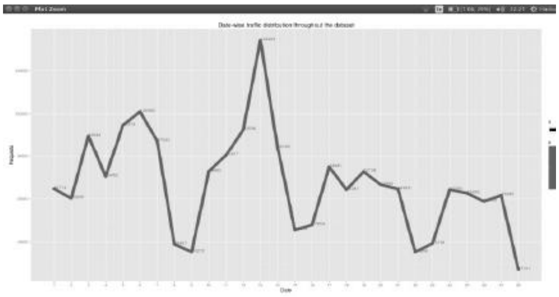

Fig. 2 Traffic distribution for Dates of the month

The following R script in Fig. 3 takes the days and requests as input and generates a line plot with days on x-axis and requests on y-axis

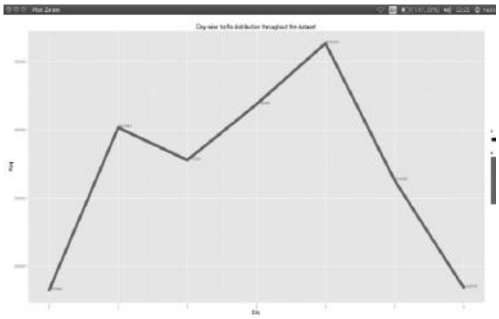

Fig. 3 Traffic distribution for Days of the week

DOI:10.32377/cvrjst0812

The following R script in Fig. 4 takes the co-ordinates of the geolocation of the IP addresses from which the requests have been generated. The co-ordinates are plotted on a world map, visualising the distribution of traffic to the web-site.

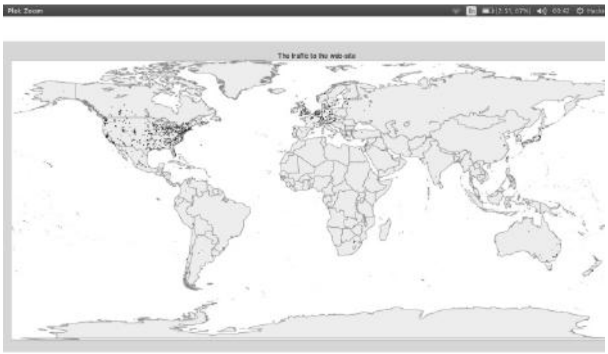

Fig. 4 Traffic distribution for location of the user

VI. Conclusions

The importance of the Big data technology and Hadoop framework is discussed in this paper. The Map Reduce techniques are implemented and tracked for the web surfer information. Our algorithms process the NASA web server logs using MapReduce task to produce a statistical report based on total count of visit per hour, per day, and per date. R programming language tool is used to visualize the details about the web surfer.

References

[1] "G. C Bodyan, T.V Shestakov, "Web Mining in Technology Management, Engineering Universe for Scientific Research and Management, Vol 1 Issue 2, April 2009.

[2] "R. Verma, S. R. Mani, "Use of Big Data Tehnologies in Capital Markets," 2012 Infosys Limited, Bangalore, India

.

[3] "J. Manyika, B. Brown et.al, "Big Data: The next frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, June 2011.

[4] "M. Ali , I. Hakki Toroslu, "Smart Miner: A New Framework for mining Large Scale Web Usage Data," WWW 2009, April 20-24. 2009 Madrid, Spain. ACM 978-1-60558-487-4/09/04.

[5] "S. Narkhede and Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs Over Hadoop MapReduce," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.

[6] "M. Bhandare, Vikas Nagare et al., "Generic Log Analyzer Using Hadoop Mapreduce Framework," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, September 2013.

[7] "K Sharadchandra Rahate et al., "A Novel Technique for Parallelization of Genetic Algorithm using Hadoop," International Journal of Engineering Trends and Technology (IJETT), vol.4, issue 8, August 2013.

[8] "T. K. Das et al., "BIG Data Analytics: A Framework for Unstructured Data Analysis," International Journal of Engineering and Technology (IJET) vol 5, No 1, Feb-Mar 2013.

[9] "J. McKendrick, "Big Data, Big Challeneges, Big Opportunities: 2012 IOUG Big Data strategies survey," September 2012.

[10] "J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Google, Inc

[11] "T. White, "Hadoop: The definitive Guide," Third Edition, ISBN: 978-1-449-31152-0-1327616795.

DOI:10.32377/cvrjst0812