International Journal of Scientific Research & Engineering Trends
Volume 2, Issue 5, Sept.-2016, ISSN (Online): 2395-566X

A Survey: Analytics of Web Log File through Map Reduce and Hadoop

Chetan Sharma, Arun Jhapate
Email: chetansearchjob@gmail.com

***Abstract* – The web is vast, diverse and dynamic and increases scalability, temporal data and multimedia issues respectively. The expansion of the Internet has given rise to a wealth of data as big data that is now available for user access. Different types of data must be managed and organized so that can be accessed by different users effectively and efficiently. The log analysis is an important issue for the web application. This paper is a review of the basics of log analysis as big data in the web environment.**

***Keywords* – Web Application, Log File, Data Mining, Big Data, Cloud.**

I. INTRODUCTION

Big Data is a term refers to Structured, unstructured and Semi structured data that is this data is having variety. Big Data is also referred a term as a data is a huge data set having really huge magnitude [1] i.e. volume (really a huge volume). Big data is that term which arrives before you and your organization has had to deal with before i.e. big data have velocity [4][7]. Big Data may be generated either by human or by machine. Human generates data as documents, emails, images, videos, posts on facebook or tweeter etc. Data comes into machine generated category are sensor data and logs data i.e. web logs, click logs, email logs. Machine generated data are of larger size than human generated data.

This flood of data is generated by connected devices from PCs and smart phones to sensors such as RFID readers and traffic cams, In health care, for instance, clinical data can now come in the form of images (e.g. from X-rays, CT-scan, and ultrasound) and videos [4]. Imaging data collected from one patient alone can easily consume several Gigabytes of storage space.

The Web is heterogeneous: data lives in various formats that are best modeled in different ways. Effectively extracting information requires careful design of algorithms for specific categories of data.

A Web application is an application that uses a Web browser as a client. The application can be as simple as a message board or guest book entry on a website, or as complex as a word processor or spreadsheet.

A web application relieves the developer of the responsibility of building a client for a specific type of computer or operating system specific. Since the client operates in a web browser, the user may be using a IBM compatible or Mac. You can run Windows XP or Windows Vista. They can even be using Internet Explorer or Firefox, though some applications require a specific web browser.

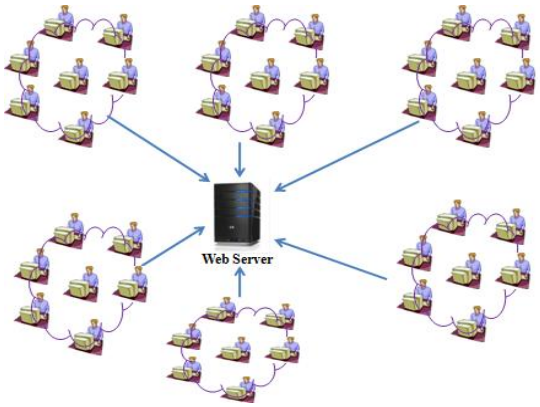

Fig.1. Client Server Architecture

Web applications typically use a combination of server scripts (ASP, PHP, etc.) and client-side script (HTML, Javascript, etc.) to develop the application. The script on the client side deals with the presentation of information, while the server-side script handles all the hard stuff like storing and retrieving information.

A Client Server architecture in which each computer or process on the network is either a client or a server. Servers are powerful computers or processes dedicated to managing disk drives (file servers), printers (print servers), or network traffic (network servers). Clients are PCs or workstations on which users run applications. Clients rely on servers for resources, such as files, devices, and even processing power.

Another type of network architecture is known as a peer-to-peer architecture because each node has equivalent responsibilities. Both client/server and peer are widely used, and each has unique advantages and disadvantages. Client-server architectures are sometimes called two-tier architectures.

II. WEB MINING

Web is the complex system of interconnected elements and mining is the process to extract data or information, Means web mining is the process of extract data from this

interconnected system. Mining of web is one amongst all the application of mining of data. Mining of data essentially deals with the organized form of data, while mining of web deals with the unorganized and partially organized form of data. Mining of web is divided into three categories i.e. web content mining, web structure mining and web usage mining [7].There are 3 types of web mining.

III. Web Content Mining

Web content mining is the important field of Web mining. Unlike search engines that simply extracted keywords to index and locate data related websites web documents (keywords) of Web-based applications, Web content mining is an automatic process goes beyond the keyword extraction. The extraction of Web contents directly into the internal content of web pages to find interesting information and knowledge. Basically, the web data consists of text, images, audio, video, metadata and hyperlinks. However, many of the data tape are unstructured text data. Research on the application of exploration techniques of unstructured data in text called knowledge discovery in texts (KDT), or extraction of text data or text mining. According to the data sources used for mining, mining, we can divide Web content in two categories: the web page of mining and web content Result of mining research. Site content directly Mining mines the content of Web pages. Web mining research results aims to improve the search results of certain research tools such as search engines.

IV. Web Structure Mining

Web structure mining studies the topology of hyperlinks with or without the description of links to discover the model or knowledge underlying the Web. The discovered model can be used to categorize the similarity and relationship between different Web sites. Web structure mining could be used to discover authority Web pages for the subjects (authorities) and overview pages for the subjects that point to many authorities (hubs). Some Web structure mining tasks try to infer Web communities according to the Web topology.

Web page cleaning is a crucial preprocessing of Web pages for most Web structure mining tasks since the linkages in noisy parts of the Web pages are usually harmful to Web connectivity analysis.

V. Web Usage Mining

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. Usage mining also allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. Web usage mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

VI. Log File

Log files on different Web servers include various types of information. The basic information found in the log file are as show in figure 2.

```
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 16, 523, 278, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/ja-transmenuh.css, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 514, 278, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/blue.css, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 520, 277, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/ja-transmenu.js, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 519, 278, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/logo-blue.jpg, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 435, 4203, 404, 3, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/img/tableft.gif, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 436, 4203, 404, 3, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/img/tabright.gif, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 516, 278, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/MANIT1.jpg, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 357, 4203, 404, 3, GET,
/images/blue/containerwrap-bg.gif, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 526, 277, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/triangle_animated.gif, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 512, 278, 304, 0, GET,
/University+Institute+of+Technology,+Bhopal+(MP)+-+Institute_files/gl.jpg, -,
192.168.1.8, -, 7/1/2016, 15:05:35, W3SVC1, RAHUL, 192.168.1.2, 0, 347, 4203, 404, 3, GET,
/images/blue/box-bl.gif, -,
```

Fig.2: Web Log File

• **User name:** This identifies who had visited the site. The user ID is mostly the IP address assigned by the Internet service provider (ISP). This can be a temporary address is assigned. It is the user's unique ID is lagging behind both here. On some sites, the identification of the user is done by getting the user's profile and allows it to access the site using a username and password. In this type of access the user is identified so that the visit of the new user can be identified.
• **Visiting Path:** The road traveled by the user when visiting the site. This can be directly using the URL or clicking a link or through a search engine.
• **Path Traversed:** This identifies the path taken by the user with the website using the different links.
• **Time stamp:** The time spent by the user in each Web page while browsing the site. This is identified as the session.
• **Success rate:** The rate of success of the site can be determined by the number of downloads made and the user activity was low copy number. If things software or purchase this also the success rate is added.
• **User Agent:** This is just the user's browser sends the request to the Web server. It's just a string describing the type and version of browser software you use.
• **URL:** The resource accessible by the user. It can be an HTML page, a CGI program or script.

VII. LOG ANALYSIS

Analysis of Web log is an innovative and unique domain constantly formed and modified by the convergence of several emerging web technologies. Because of its interdisciplinary nature, the diversity of issues addressed, the variety and the number of Web applications, is subject to many different and various research methodologies.
The expansion of the World Wide Web has led to a large amount of data that is now generally freely available to access different types of data users must be managed and organized so it can accessed by different users effectively. Therefore, the application of data mining techniques on the web is now the focus area to the point of a growing number of researchers. Several extraction methods were used to discover hidden information on the Web. Extracting Web became an autonomous research area.

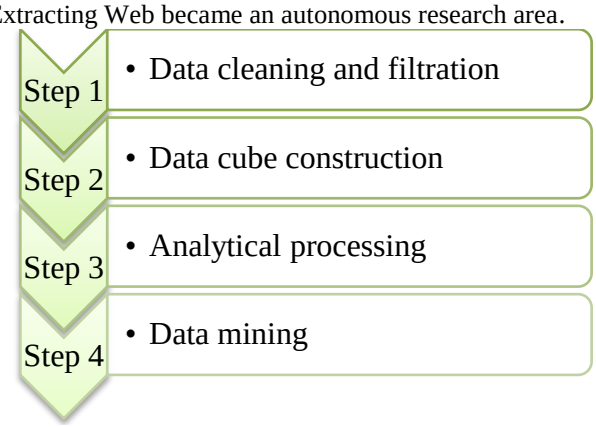

Fig.3: Log Analysis

• **Data cleaning and filtration.** In the first step, the data is filtered to remove all the relevant information. The remaining significant data are converted into a relational database. The database is used as an effective benchmark for the extraction of information for the aggregation of simple and summary data based on the simple attributes.
• **Data cube construction.** In the second step, a data cube is created using all available dimensions.
• **Analytical processing**. Building a data cube allows the application of analytical processing techniques in this stage. It can be done online which is known as online analytical processing or OLAP
• **Data mining.** In the final phase, the data-mining techniques are put to use with the cube interested in finding the information you want data. Common data exploration objectives are: analyzing characterization data, the comparison of the class series, association, prediction, classification and time

VIII. LITERATURE SURVEY

The Log file is the main source of the state of the system[1], the analysis of user behavior, etc. log analysis system should not only the amount of data processing capacity and stable, but also adapt to a variety of settings the need for efficiency, which can be realized independently of analytical tools and even under single cloud computing. It presents a platform in the unified data analysis batch file with the combination of Hadoop and Spark clouds. Hadoop offers a distributed file system and the offline batch part of the computer while the Spark calculation model is based on the distributed memory. The articulation of Hadoop, Spark and storage and analysis tools of the hive and shark data helps provide a unified platform with a batch analysis capacity and memory cloud to connect in a double process High sense available, stable and efficient.

There are two main difficulties in analysis website [2] user demand for information by the traditional method of analysis of the recording. First, it is difficult to associate the request with a specific user recover with precision, so it can not take into account the relationship between user characteristics and behavior of the user. Second, it is difficult to extract complete interaction accurately processing information. This paper presents an improved method of analysis of log depth, whose central idea is to rebuild the log data using partnership. Specific measures include the association of behaviors partnership glossary and IP association. Experiments were conducted to demonstrate that the registration of reconstructed data is the most effective request for information from the user mining. The method can solve the shortcomings of the traditional method of analysis register, and get the good result of the experiment.

In ubiquitous computing high levels of connectivity are required [3]. As concerns the security aspects are essential. A strategy that can be applied to improve the analysis of security log. These strategies can be used to promote

understanding of systems, in particular, detection of intrusion attempts. The functioning of modern computer systems, as used in the ubiquitous computing, tend to generate a large number of files that require the use of automated tools to facilitate analysis. The tools we use data mining techniques to save the analysis were used to detect attempts to attack computer systems, assistance in security management. Therefore, this article proposes an approach to carry logs with Intuit analyzes to avoid a strike. The proposed solution explores two fronts: (i) the registration dossiers of applications, and (ii) the log records of network and transportation layers. To evaluate the proposed approach using a prototype modules for data collection and standardization of data was designed. The normalization module further add contextual information to aid in the analysis of emergencies for safety. To preserve the autonomy of the system, records network and transport layers are collected and evaluated current connections. The tests were developed in the proposed solution, showing good result for typical attack categories.

Analysis of the Security log [4] is extremely useful for discovering anomalies and intrusions. However, the large amount of log data, new frameworks and techniques of science and of computer security. A framework for analysis of distributed security log and parallel light that enables organizations to analyze a massive number of system, network and transaction logs efficient and scalable manner is presented. Unlike frameworks distributed, for example, MapReduce, our framework is designed specifically for the analysis of security log. It has a minimal set of necessary properties such as dynamic programming tasks for streaming recordings. For prototyping, we implement our environments executives Amazon Cloud (EC2 and S3) with a basic analysis application. Our evaluation shows the effectiveness of our design and shows the potential of our scenarios distributed cloud-based framework in large log analysis.

Log is the main source [5]of the state of the system, the behavior of the user, the actions of the system, etc. Analysis system connection should not only the amount of data processing capacity and stable, but also to accommodate a variety of scenarios in accordance with the requirement of efficiency and performance, which can 't be reached tools Independent analysis available computer chassis or even individually. Therefore, a log analyzer with the combination of Hadoop and Map-Reduce paradigm is proposed. The articulation of Hadoop MapReduce programming tools and allows to provide batch analysis in the most the ability to promptly response and computer memory to process connect a high fashion available, efficient and stable.

IX. TECHNOLOGICAL ASPECT

Complex data analysis and identifying patterns on web based environment is explained with Figure 3. Basically here we have many sources of the data generation. This data is basically massive and progressive. Every data centers have to supply this data in distributed environment and here load balancing is crucial issue. These scalable database management systems have to supply massive data to number of application. Load balancing and security is major issue in big data. From fig we can simply analysed that these data generation from different sites in massive amount is progressive and has to distribute over cloud environment. Mapping this data to primary sources to their required destination required Hadoop like framework and Map Reduce like technology.

X. HADOOP PLATFORM

Hadoop is open source framework and has two components which are HDFS and Map Reduce. HDFS is distributed file system for storing and retrieving data for Map Reduce and help to executes jobs for users. Hadoopp Form the cluster of data nodes and store data on space utilization of data nodes on cluster [11]. Hadoop runs on heterogeneous environment and may leads to workload problem while data distribution and access. Above the HDFS layer there is Map Reduce engine which consist on operating part for the server. One problem with Hadoop is workload balancing will be handled by transfer of data between racks. Modifying data transfer rate between the racks is another technique.

It is a platform which provides solution for big data in a dispersed environment [3]. It is a software framework which was developed by Google's MapReduce, where an application dissociates into various parts. The two main components of Hadoop which are HDFS and MapReduce are explained as follows:

XI. HDFS

It is a failure resistant storage architecture included in Hadoop called as HDFS [3-4]. HDFS has the ability to extend incrementally and resist fault from important portions of infrastructure without data loss and store huge amounts of data. Clusters of machines are created by Hadoop and work is coordinated among them. If any machine of the cluster fails still, the operation of Hadoop is by transferring task to other system in cluster without losing data or affecting our work. Cluster storage in HDFS is managed by dividing the files which enters known as "chunks," and then these chunks are stored repeatedly across servers. HDFS stores 3 complete replicas of these file copying every part to same number of servers.

HDFS Cluster comprise of two types of nodes (i) name node (the master) which manages the file system namespace, the metadata of all files and file system tree. (ii) Number of data nodes (workers) whose work is to store and retrieve block as per instruction of name node. Without the name node accessing the file is a difficult task. The stored chunks of data which are retrieved is informed back to the name node. Therefore it is essential for name node to prevent from failure.

XII. MAP REDUCE

Map Reduce framework is the major component of the Hadoop system [4]. This platform permits the order of a process that can be used for big data set and break data along with problem and execute it simultaneously. This can be applied on multiple dimensions from analyst point of view.

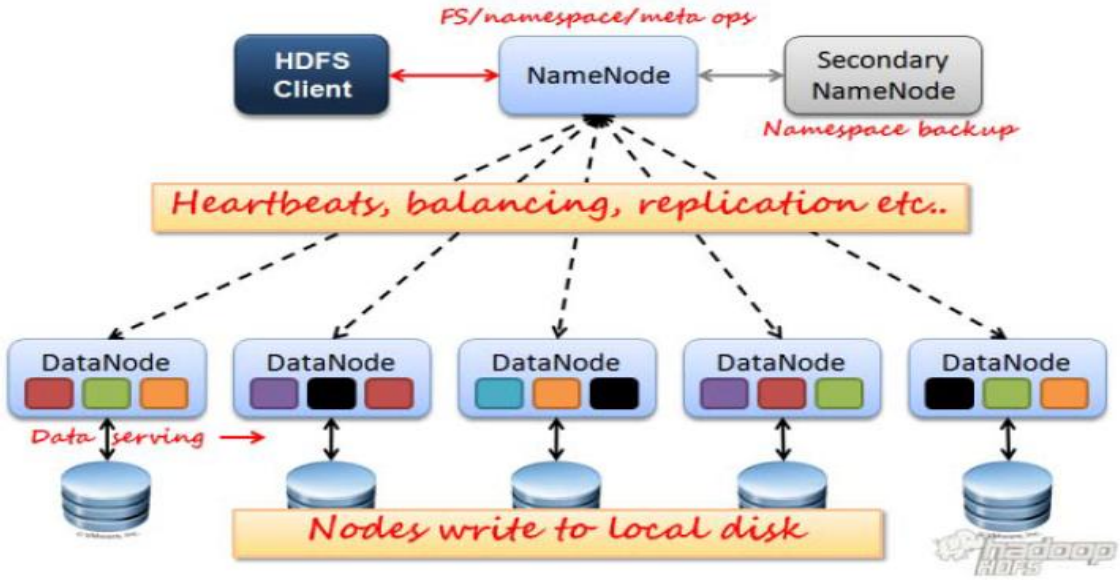

Fig.4:. Hadoop Architecture

Map Reduce works with HDFS with functions namely Map and Reduce. Whenever job is assign by the users to Hadoop, immediately input spited in multiple pieces and Map function will be applied to data for generating intermediate result. Intermediate result will be monitor and shuffled for generating final result. Whenever a job tracker gets job from client it will first execute Map task first and then finds processes for every split data [5][7].

The two functions of MapReduce are map and reduce. Key/value pairs are taken as input by the function and intermediate sets of key/values pairs are generated is known as map. The values which are Intermediate associated with the same key are merged through this function known as reduce[3].

All Hadoop-related projects lack an user friendly interface and are difficult to learn but are mostly used by companies working with big data[5]. There are excellent data mining tools and commercial solutions where Rapid Miner can be applied. The most popular data mining tool is Rapid Miner which is easy to learn with a clean user interface. Moreover, it is extendable, providing additional functionality to the developer to develop the basic software [6]. It is used for business and engineering applications as well as for rapid prototyping, application development, research, education, training [7] .

XIII. CONCLUSION

Analysis of Web log is an innovative and unique domain constantly formed and modified by the convergence of several emerging web technologies. Because of its interdisciplinary nature, the diversity of issues addressed, the variety and the number of Web applications, is subject to many different methodologies and different research. Log file is a crucial part of web application. In this manner the log analysis is also plays an important role in the various applications and treats as big data.

The present paper elaborates the concept of Hadoop a Big data tool. Our conclusion in the present paper is based on the fact that Hadoop is a tool which is made for handling big data analytics and meets with the expectation of ever growing demands of the data. Any general integration on this work has not yet been known but some work has been done on integrating Log file to hadoop.

REFERENCES

[1] Xiuqin Lin; Peng Wang; Bin Wu, "Log analysis in cloud computing environment with Hadoop and Spark," in *Broadband Network & Multimedia Technology (IC-BNMT), 2013 5th IEEE International Conference on* , vol., no., pp.273-276, 17-19 Nov. 2013

[2] Chaofei Wang; Jing Chen; Xiaopeng Liu; Jinwei Zhao, "An improved deep log analysis method based on data reconstruction," in Cloud Computing and Intelligence Systems (CCIS), 2014 IEEE 3rd International Conference on , vol., no., pp.86-90, 27-29 Nov. 2014

[3] da Silva Machado, Roger; Borges Almeida, Ricardo; Correa Yamin, Adenauer; Marilza Pernas, Ana, "LogA-DM: An Approach of Dynamic Log Analysis," in Latin America Transactions, IEEE (Revista IEEE America Latina) , vol.13, no.9, pp.3096-3102, Sept. 2015

[4] Xiaokui Shu; Smiy, J.; Danfeng Yao; Heshan Lin, "Massive distributed and parallel log analysis for organizational security," in Globecom Workshops (GC

Wkshps), 2013 IEEE , vol., no., pp.194-199, 9-13 Dec. 2013

[5] Hingave, H.; Ingle, R., "An approach for MapReduce based log analysis using Hadoop," in Electronics and Communication Systems (ICECS), 2015 2nd International Conference on , vol., no., pp.1264-1268, 26-27 Feb. 2015

[6] K Savitha and MS Vijaya , "Mining of Web Server Logs in a Distributed Cluster using Big Data Technologies" , International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5 , 2014

[7] T. K. Das , "BIG Data Analytics: A Framework for Unstructured Data Analysis" , International Journal of Engineering and Technology (IJET) , vol. 5 , no. 1 , 2013

[8] Jiang Dawei, K.H. Antony and Gang Chen , "MAP-JOINREDUCE: Toward Scalable and Efficient Data Analysis on Large Clusters" , IEEE Transactions on Knowledge and Data Engineering, pp.1299 -1311 , 2011

[9] Pavlo Andrew, Paulson Erik, Rasin Alexander, J. Daniel, J. David, De Witt, Samuel Madden and Michael Stonebraker , "A Comparison of Approaches to Large-Scale Data Analysis" , ACM SIGMOD International Conference on Management of data , pp.165 -178 , 2009

[10] W. Xu, L. Huang, A. Fox, D. Patterson, M. Jordan. "Online System Problem Detection by Mining Patterns of Console Logs". In the Proceeding of ICDM '09 Proceedings of the 2009 Ninth IEEE International Conference on Data Mining.

[11] AiLing Duan et. al. “Research and Practice of Distributed Parallel Search Algorithm on Hadoop_MapReduce”, 2012 International Conference on Control Engineering and Communication Technology, 2012 IEEE DOI 10.1109/ICCECT.2012.131, pp 105108

[12] D. Agrawal, A. El Abbadi, S. Antony, and S. Das. Data ManagementChallenges in Cloud Computing Infrastructures. In *DNIS*, pages 1–10, 2010.

[13] White paper on “Solution Brief Big Data in the Cloud: Converging Technologies , How to Create Competitive Advantage Using Cloud- Based Big Data Analytics.

[14] Tharam Dillon et. al. “Cloud Computing: Issues and Challenges”, 2010 24th IEEE International Conference on Advanced Information Networking and Applications,

[15] D. Agrawal, S. Das, and A. E. Abbadi. Big data and cloud computing: New wine or just new bottles? *PVLDB*, 3(2):1647– 1648.