Big Data Mining for Web Usage Mining (WUM)

Prof. Dhruva Mistry[#], Prof. Mihir H. Rajyaguru[#], Prof. Prasann Barot[#]

[#] *Assistant Professor, IT department, SPCE, Bakrol, Gujarat, India*

Abstract— **Big Data is a term used to identify the datasets, whose size is beyond the ability of typical database software tools to store, manage and analyse. Web mining techniques primarily focuses on deciphering and scrutinizing the navigational behaviour of user form various aspects and ascertaining the hidden knowledge from these web logs. Log files incorporate traits of user behaviour, therefore it is essential to analyse log data and acquire knowledge from it. In this paper we explain different big data mining techniques.**

Keywords – **Web Usage Mining (WUM), Hadoop, MapReduce, Cloud computing, HDFS, Big Data, Visual Web Mining (VWM), Web Bot.**

I. INTRODUCTION

Big data is used to identify the datasets that whose size is beyond the ability of typical database software tools to store, manage and analyse. Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web documents, hyperlinks between documents, usage logs of web sites, etc. This technique enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. [1]. The goal of Web Usage Mining (WUM) is to understand the behaviour of web site users through the process of data mining of web access data. Knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing [2].

Web Data Mining is not only focused to gain business information but is also used by various organizational departments to make the right predictions and decisions for things like business development, work flow, production processes and more by going through the business models derived from the web data mining as lots is available on the World Wide Web. It can find characteristics and rules of the users' visiting behaviour to improve the service quality to users.

A user profile is often used to classify a given user into predefined user segments (e.g., by demographics or tastes) or to capture the online behaviour of the user including the user's private interests and preferences. In scientific way, we can call cloud computing is an alternative word for distributed computing over a wired or wireless network; it relies on the shared resource to accomplish coherence and scale of economies as similar to utilize the resource over a network, and means the ability to run program or any application on cluster of many computers simultaneously.

As we observed, the Internet has now changed from computing to cloud computing. MapReduce is a great programming model in cloud computing that was introduced at Google. MapReduce is a programming model and an associated implementation for processing and generating large data sets [4].

The Visual Web Mining (VWM) is as the application of Information Visualization techniques on results of Web Mining in order to further amplify the. Perception of extracted patterns, rules and regularities, or to visually explore new ones in web domain [5]. Big data is a heterogeneous collection of both structured and unstructured data [6]. In this paper we are introducing the Web data mining technique and its implementation for handling the big amount web data with VWM and Apache Hadoop MapReduce framework and also cloud computing to handle big data.

II. RELATED WORK

In [1] they analyse the web data of today's advanced world. Internet becomes an important part of many of all organizations, businesspersons and daily individuals. As a web logs are available in different formats,So it is important to mine meaningful data from that. In this paper we have studied two effective technique to mine big data one with Apache Hadoop MapReduce and second with visualization based technique called as Visual Web Mining (VWM).

In [2] it observed that the knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing.

They also provide review of pattern discovery algorithms which utilize association rules, sequencial patterns and classification.

In [3] they had a find new approach through comparing the two tests, the experimental result shows that the new approach can improve the execution efficiency and reduce the execution time. The new algorithm can work well and there is no association rule lost. It can be well used in business. Also noticed that they can aim to find a more accurate and faster approach for Web data mining that is based on cloud computing.

In [4] they observed first, the model is easy to use, even for programmers without experience with parallel and distributed systems, since it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing and secondly a large variety of problems are easily expressible as MapReduce computations.

In [5] they analysed portrays for the users browsing pattern and summarizes the outcome into a graphical report which depicts most visited web pages, browsing session and trending keywords. Hadoop MapReduce framework provides parallel distributed processing and reliable data storage for large volumes of log files. In order to manage such log files, Hadoop MapReduce plays a key role by proficient management of data and decreases the response time. The proposed system with the help of Hadoop MapReduce analyses the log files and segregates the fields of the log files using regular expression mechanism. Regex not only reduces the code length but also reduces the overhead of usage of string functions. The segregated and structured fields are stored in the database in accordance with Hadoop thereby enabling ease of data retrieval.

In [6] they had reviewed that the amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. They discuss issues and challenges related to big data mining and analyse that data with tools like Map Reduce over Hadoop.HDFS which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data.

In [7] Web analysis involves the transformation and interpretation of the Web log records to find out the hidden information or predictive pattern by the data mining and knowledge discovery technique. In this paper they tried to give a clear understanding of the data preparation process and pattern discovery process. Web usage patterns and data mining can be the basis for a great deal of future research. More research needs to be done in ECommerce, Bioinformatics, Computer Security, Web Intelligence, Intelligent Learning, Database Systems, Finance, Marketing, Healthcare and Telecommunications by using Web usage mining.

In [8] they analyze that Web usage mining is a task of data mining of web data that extract knowledge about web users and how these users interacting with the web site. Our motive of web usage mining is to provide web intelligence by discovering users' access patterns of web pages, such as frequently visited hyperlinks, frequently accessed web pages, popular access sequence of web pages, and users grouping etc. automatically and quickly from the huge sever access log records.

In this paper we reviwed few techniques for mining large volume of web data. Hadoop MapReduce architecture is one of these techniques to process distributedly on large datasets.The other technique Cloud computing works with the Internet,Hadware and software to store big data.At the same time one visual application of information is also there to find meaningful patterns from web log data and explore it to new domain.

III. TECHNIQUES FOR BIG DATA MINING

A. Cloud Computing

Cloud computing is for long-held dream of computing as a utility which has recently emerged as a commercial reality [3]. Cloud computing refers to both the application delivered as services over the Internet and the hardware and system software in the data centres that provide these services. Cloud computing is at the pick in the computer world these days, as the usage show it is too big of a buzz, cloud computing is stand for different things to different people, it is so much vast and versatile, not only for branch of IT [9]. Research firm IDC thinks that, by 2020, cloud spending will reach $203.4 billion worldwide [11]. Cloud computing started with a risk-free concept: Let someone else take the ownership of setting up of IT infrastructure and let end-users tap into it, paying only for what is been used. A service offering computation resources is frequently referred to as Infrastructure as a Service (IaaS) and the applications as Software as a Service (SaaS). An environment used for construction, deployment, and management of applications is called PaaS (Platform as a Service) [10].

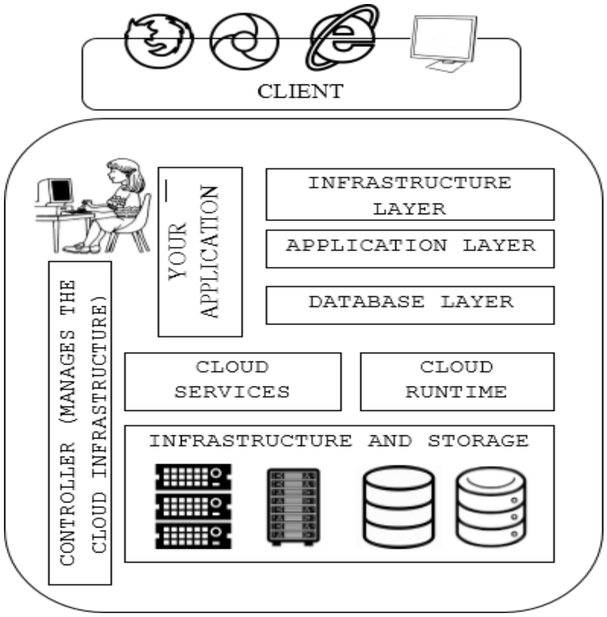

Fig. 1 Cloud Computing Architecture.

B. Hadoop MapReduce

Hadoop is an open source distributed computing framework which is used for distributed processing of large data sets and designed to satisfy clusters scaled from a single server to thousands of servers. Hadoop is the most widely used cloud computing platform in recent years and has been adopted by major Internet companies and research institutions. A Hadoop cluster is composed of two parts: first is the Hadoop Distributed File System (HDFS) and second is the MapReduce. Hadoop is the optimal choice to realize our approach. Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The Hadoop and MapReduce communities have developed a powerful framework for performing predictive analytics against complex distributed information sources [5].

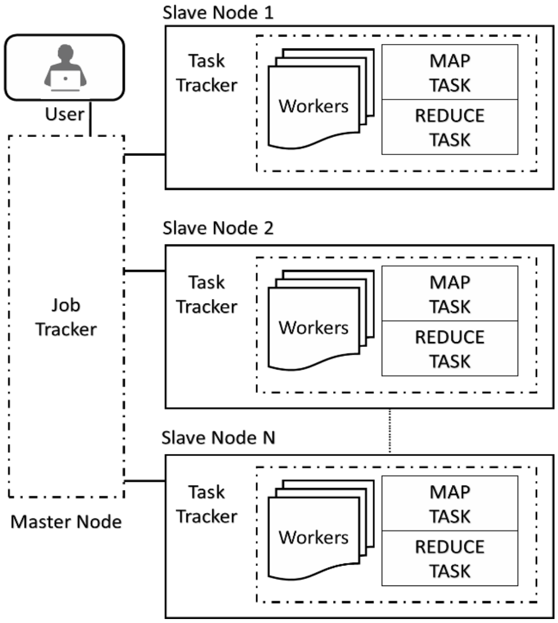

Fig. 2 Hadoop MapReduce Programming Model.

C. *Visual Web Mining Architecture*

The Architecture of implementing the Visual Web Mining (VWM) is shown below in Figure 3. Visual Web Mining (VWM) as application of Information Visualization techniques on results of Web Mining in order to further amplify the perception of extracted patterns and visually explore new ones in web domain[12].

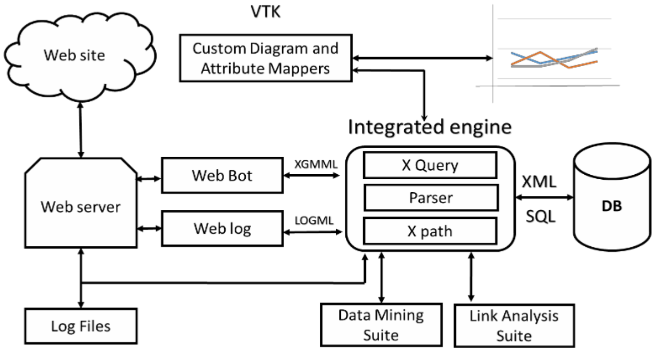

Fig. 3 Visual Web Mining Architecture.

Analysis of web site usage data involves two significant challenges: firstly the volume of data, arising from the growth of the web, and secondly, the structural complexity of web sites. We target one or a group of websites for analysis. Input of the system consists of web pages and web server log files. Access to web log is done by the local file system, or by downloading it from a remote web server. A web robot (Web Bot [12]) is used to retrieve the pages of the website. In parallel, Web Server Log files are downloaded and processed through a sessionizer and a Log Markup Language (LOGML) file is generated.

IV. CONCLUSIONS

Now in digital world amount of web data is increasing much faster day by day, so it is necessary to justify that which data are useful and which are not. To overcome this situation we have done analysis of different solutions. Big data analysis tools like MapReduce over Hadoop and HDFS which helps organizations to better understand their customers and the marketplace and to take better.

The main objective of cloud computing is to make better use of distributed resources and to solve large scale computation problems. We noticed that this technique is also faster and accurate for web data mining. The drawbacks of existing system is can be solved by Hadoop and MapReduce framework. Cloud computing uses the same framework that plays a key role by proficient management of data and decreases the response time.

REFERENCES

[1]. Pranit B. Mohata, Prof. Sheetal Dhande"Web Data Mining Techniques and Implementation for Handling Big Data", IJCSMC, Vol. 4, Issue. 4, ISSN 2320–088X April 2015.

[2]. Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah. "Web usage mining: A review of recent works", The 5th International Conference on ICT4M, Electronic ISBN: 978-1-4799-6242-6, Publisher: IEEE, 17-18 Nov. 2014.

[3]. Wenzheng Zhu* and Changhoon Lee"A New Approach to Web Data Mining Based on Cloud Computing", Journal of Computing Science and Engineering,Vol. 8, No. 4, December 2014.

[4]. Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters", OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[5]. Jalpa Mehta,Amir Ansari,Aseem Girkar, Ayesha Khanna,Ankit Nagda "Trend Analysis based on Access Pattern over Web Logs using Hadoop" International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 8, April 2015.

[6]. Jaseena K.U and Julie M. David"Issues,challenges and solutions: Big data mining" CSIT, GRAPH-HOC, SPTM – 2014,CS & IT-CSCP 2014.

[7]. S. P. Nina, M. Rahman, K. I. Bhuiyan and K. E. U. Ahmed, "Pattern Discovery of Web Usage Mining," *2009 International Conference on Computer Technology and Development*, Kota Kinabalu, 2009, pp. 499-503. doi: 10.1109/ICCTD.2009.199, Publisher: IEEE.

[8]. M. Dhandi and R. K. Chakrawarti, "A comprehensive study of web usage mining," Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-5. doi: 10.1109/CDAN.2016.7570889, Publisher: IEEE.

[9]. S. Patidar, D. Rane and P. Jain, "A Survey Paper on Cloud

Computing," 2012 Second International Conference on Advanced Computing & Communication Technologies, Rohtak, Haryana, 2012, pp.394-398. doi: 10.1109/ACCT.2012.15, Publisher: IEEE.

[10]. Rajkumar Buyya, "Introduction to the IEEE Transactions on Cloud Computing", IEEE Transactions on Cloud Computing, (TCC), Vol. 1, No. 1, January-June 2013

[11]. FRAMINGHAM, Mass, IDC-Analyze the future, Article, February 20, 2017.

[12]. Amir H. Youssefi, David J. Duke, J. Zaki Rensselaer, "Visual Web Mining", New York, New York, USA, ISBN: 1-58113-912-8, ACM 1- 58113-912-8/04/0005, WWW2004, May 17–22, 2004.