International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

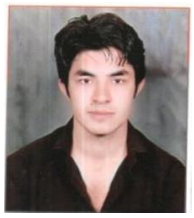
Vol.3., Issue.5., 2015
(Sept.-Oct.)

RESEARCH ARTICLE

ISSN: 2321-7758

WEB LOG MINING USING MAPREDUCE AND APACHE SPARK

AVNEESH TIWARI, RISHABH SONI, Dr. SANJAY AGRAWAL
NITTTR BHOPAL



AVNEESH TIWARI

RISHABH SONI

Dr. SANJAY AGRAWAL

ABSTRACT

In today's Internet word every one connects with internet directly or indirectly. Modern software based systems collect information about their activity and history in logs. These logs information can be used to analyze. The logs having timestamps, IP address, month, date, etc. In today's internet scenario log file analysis become necessary task for analyzing the customer's behaviour and for improve the web applications and banking systems, etc. The rapid changes in technology made it possible to capture the user's interactions with web in the form of web log file. Log file is in the form of text file in systems. Very large amount of data in the web log cannot be directly used in the process of web log mining. Log files are generated very fast in machines and these datasets are very huge. So to analyse these huge datasets we need a parallel processing mechanism. Hadoop and apache spark are well known parallel processing system. HDFS and MapReduce are parallel processing systems. In this paper, authors propose a log analysis system which is run over hadoop MapReduce and Apache Spark environment. Both the framework proposes log data in parallel system using all the mechanism in the hadoop and spark cluster and computes result efficiently. Output in the form of some parameters of log such as IP addresses, month, date, time etc so that it can be useful for real time projects, companies, banks etc.

Keywords: Web Log Mining, Hadoop, Map Reduce, Apache Spark.

INTRODUCTION

Web mining is major and important fields of data mining. Web log mining techniques are applied [1] on structures, contents and log files of web sites to achieve better performance, schema modifications and web personalization and of web sites. Web log mining has three categories [2] such as:

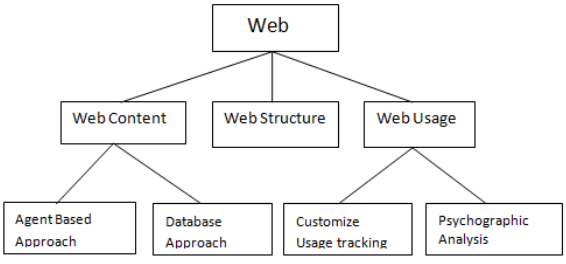


Figure1 Types of Web Mining [8]

I) Web Usage Mining
II) Web Structure Mining and
III) Content Mining

In web usage mining (WUM) or web log mining, users' behaviour or interests are revealed by applying data mining techniques on web log file. In web structure mining, we mine the structure of website on the basis of hyperlinks and intra-links inside and outside the web pages. In web content mining [2] we discover useful information from the contents of web site which may include text, hyperlinks, metadata, images, videos, and audios.

Hadoop

Hadoop is a frame work it is used to process the data very fast. Hadoop has become known as an excellent platform in the area of Big Data for data processing. It produces an authentic storage and high performance in the area of Big Data [3]. Hadoop analysis system builds from commodity hardware.

Map Reduce

Map reduce is a programming model that is used in the Hadoop to process the data [3]. Map-Reduce basically uses the java programming with the Hadoop jar file support it execute the java program in the HDFS environment. This is very good and efficient programming techniques to process the data using any programming language. Even we can use C, python to process the data in place of java bit java is easy and generally popular language now a days so most companies prefer Java.

Map stage:

Input text is one record each line, we use the Long Writable and Text in MapReduce package as the initial input types of key and value respectively, where the value of key is the offset of each line, and the value of value is the content of the corresponding line.

Reduce stage:

Firstly, statistics on the keywords of the same user ID. If one same keyword appears many times, plus one every time. At last, put final result into HDFS. Because keys have been sorted before Reduce, all the values of same key are put together, namely have been encapsulated as Iterator <Value Type>. These all can be processed at the same time, and the final output is value which encapsulates the same user's all key words and weights. The middle key is taken as the final output without processing.

Apache Spark: Apache Spark is a parallel cluster computing framework. Compared with Hadoop two-stage disk-based Map Reduce system, Spark provide efficient result and gives very less processing time in certain applications. [5][6] Spark is very efficient for data analytics and machine learning. In Spark we can write applications in Scala, Java, R, and Python. With compare to Map and Reduce operations, Apache Spark supports streaming data, SQL queries, machine learning and graph data processing.

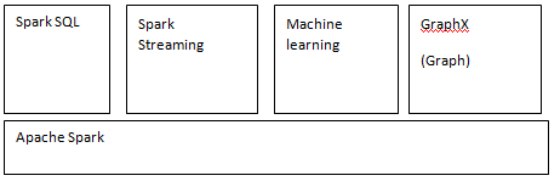


Figure 2 Operations of Apache Spark

Spark Architecture:

Apache Spark uses hadoop file system for data storage. And it works with HDFS, HBase, Cassandra or any other hadoop compatible data source.Scala, Java, and Python programming languages use API in Apache Spark. Apache Spark can be utilized as a Stand-alone server and also it can be on a distributed computing framework.

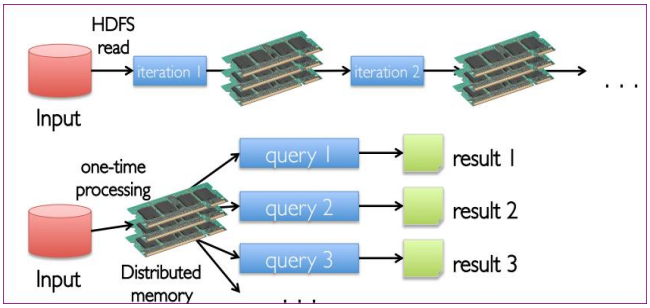


Figure3 In memory data sharing [4]

Related work: Many researchers have proposed their techniques for web log mining. Some of them we are highlighted hear. Yan Liu [13] States that System anomaly detection is much focused area for performance refinement development, maintenance in large scale of distributed systems. This is important to generalize problem diagnosis and troubleshooting by analyzing web logs. Although due to the growing scale and complexity of distributed systems, the category of logs must be very large. It is not efficient for common methods to analyze system logs on

single node. According to author, [6] Web usage mining is the kind of Web mining activity that demand the automatic find of what user access patterns from Web servers. Author describe in this paper analyze the web log using distinct algorithms like Hash tree, Apriori, and Fuzzy and then author proposed enhanced Apriori algorithm to give thesolution with higher optimized efficiency for Crisp Boundary problem. The proposed algorithm is based on the Hash tree Algorithm steps of frequent item sets and rule generation phases. Pabarskaite (2002) [14] states that pre-processing of web log file plays an important role in WUM and takes 80% of total time of web mining.

Experimental Setup: For this research we have taken log files of NASA server. These log files contain fields like IP address, URL, date, month, year, time, images. We have installed Hadoop and Apache Spark. Hadoop is built on a single node cluster have Intel(R) Core(TM) W3565 CPU @ 3.20GHz Processors and System type 64 bit Operating system, 4 GB memory. The operating system is GNU/ CentOS-6.5-x86_64, VMware Workstation 10.0.1 Build 1379776, hadoop-1.0.4**,** we implement our method using the Map Reduce programming paradigm to and Java version is 1.7.0_79.Eclipse-standard-kepler-SR2-linux-gtk-x 86 _ 64, the number of replicas is set to 1 and HDFS block size is 64 MB during the tests. For running Apache Spark, Virtual Box and Vagrant VM are used. Oracle Virtual Box is a hypervisor for x86 computers from Oracle Corporation. For this research, Virtual Box is installed on the existing host operating system, which is Windows 8. Vagrant is a tool for building complete development environments. It can be seen as a higher-level wrapper around virtualization software such as Virtual Box and VMware. Log Files are stored on Hadoop and Apache Spark. Log files are distributed evenly on this node. MapReduce job and Apache Spark (pySpark) are then run on these files and get analysed result in graphical formats. Analyzed result shows total hits of each URL and gives total number of count in every month. We have also taken result of AWK programming uses the same log file for comparison with both MapReduce and Apache Spark. Figure shows the execution time of Mapreduce, Apache Spark and AWK.

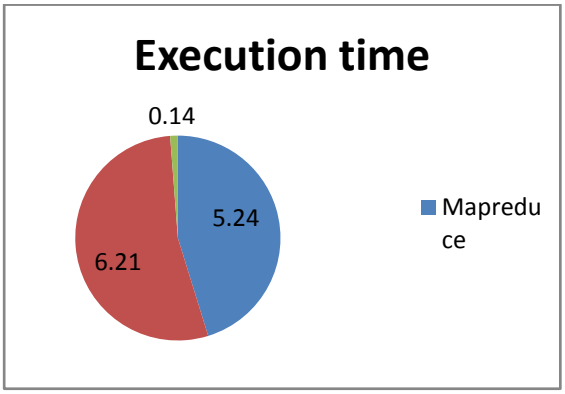


Figure 4. Shows the comparison between the Mapreduce, AWK and Apache Spark

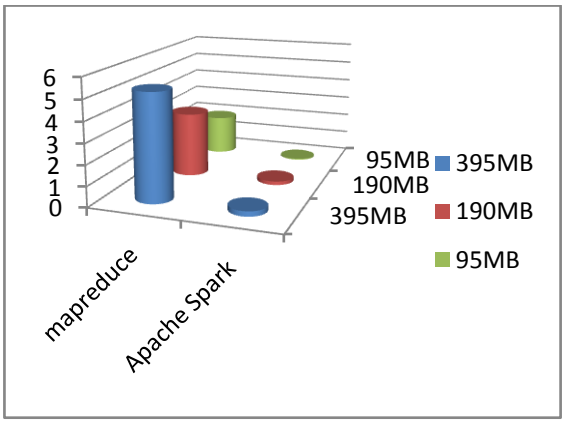


Figure5. Comparison of Execution Time with different data sets

Conclusion:

Web Log Mining is an important area in every fields of computer technology. This plays a vital role in many areas such as application design, anomaly detection, debugging, and security threats. Web Log Mining also has challenges in its mining techniques, large size of log data, and structure of data. Large size is important factor which affects the efficiency to analyze log data. This paper works focus on efficiently web log mining technique over large data sets. Different architecture has been used for Web log mining. In this paperwork we have applies web log mining on the Hadoop MapReduce and Apache Spark frameworks. Hadoop and Apache Spark both are able to work on the every type of data, structures unstructured and semi-structured. We have tested unstructured data sets in both frameworks. We have taken files of different size that contains a huge number of log data. This paper is based on the concept of Web log mining. Earlier days web log mining done by some tools and technique. But

because of structure and huge size of log data, these traditional tools and technique are not well suited for Log file Analysis. By implementing both frameworks on Web log mining, performance of Mapreduce and Apache Spark has been compared by applying different data sets.Apache Spark gives better execution time as compared to MapReduce.

REFERENCES

[1]. S.Veeramalai,N.Jaisankar and A.Kannan.Efficient Web Log Mining Using Enhanced

[2]. Apriori Algorithm with Hash Tree and Fuzzy. (IJCSIT) Vol.2, No.4, August 2010

[3]. Qiang FU, Jian-Guang LOU, Yi WANG, Jiang.Execution Anomaly Detection in Distributed Systems through Unstructured Log Analysis. 2009 Ninth IEEE International Conference on Data Mining.

[4]. Suneetha K. R. and D. R. Krishnamoorthi (2009). Identifying User Behaviour by Analyzing Web Server Access Log File. IJCSNS VOL.9 No.4, April 2009.

[5]. https://www.virtualbox.org/

[6]. https://www.vagrantup.com

[7]. S.Veeramalai ,N.Jaisankar and A.Kannan. Efficient Web Log Mining Using Enhanced

[8]. Apriori Algorithm with Hash Tree and Fuzzy. (IJCSIT) Vol.2, No.4, August 2010

[9]. Ang FU, Jian-Guang Lou, Yi Wang, Jiang Li.Execution Anomaly Detection in Distributed Systems through Unstructured Log Analysis. 2009 Ninth IEEE International Conference on Data Mining

[10]. R.Shanthi, Dr.S.P.Rajagopalan.An Efficient Web Mining Algorithm To Mine Web Log Information. IJIRCE2013

[11]. MeiyappanNagappan, Mladen A. Vouk. Abstracting Log Lines to Log Event Types for Mining Software System Logs.

[12]. SayaleeNarkhedeand TriptiBaraskar. HMR Log Analyzer: Analyzer Web Application Logs over Hadoop Mapreduce. International Journal of UbiComp (IJU), Vol.4, No.3, July 2011

[13]. Mirghani. A. Eltahi,andAnour F.A. Dafa-Alla. Extracting Knowledge from Web Server Logs Using Web Usage Mining. 2013 (ICCEEE)

[14]. S Saravanan, B Uma Maheswari. Analysing Large Web Log Files in a Hadoop Distributed Cluster Environment.Int.J.Computer Technology &Applications,Vol 5 ,1677-1681

[15]. Yan Liu, Ning Cao, Wei Pan. System Anomaly Detection in Distributed Systems through MapReduce- Based Log Analysis. 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)

[16]. Pabarskaite Z. (2002).Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.