ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Statistical Analysis of Web Server Logs Using Apache Hive in Hadoop Framework

Harish S[1], Kavitha G[2]

PG Student, Dept. of Studies in CSE, UBDT College of Engineering, Davangere, Karnataka, India[1]

Assistant Professor, Dept. of Studies in CSE, UBDT College of Engineering, Davangere, Karnataka, India[2]

ABSTRACT: Web log file is log file automatically created and maintained by a web server.Analyzing web server access logs files will offer valuable insight into website usage. Because of the tremendous usage of web, the web log files are growing at faster rate and the size is becoming huge. Processing this explosive growth of log files using relational database technology has been facing a bottle neck. To analyze such large datasets we need parallel processing system and reliable data storage mechanism. Hadoop rides the big data where massive quantity of information is processed using cluster of commodity hardware. In this paper based on the architecture of Hadoop Distributed File System and HadoopMapReduce framework and HiveQL query language, we present the methodology used in pre-processing of huge volume of web log files and finding the statics of website and learning the user behavior.

KEYWORDS: big data; hadoop; mapreduce; web server logs; log analysis; hive

I. INTRODUCTION

In today's world, everything is going online. Insuch a competitive environment, service providers are eager to know about, are they providing the best service in the market, whether people are purchasing their product, are they findingapplication interesting and friendly to use, or in the field of banking they need to know about howmany customers are looking forward to their bank scheme. Service providers also need to know, how to make websites or webapplication interesting, which products people are not purchasing and in that case how to improveadvertising strategies to attract customer, what will be the future marketing plans [1]. To answer these questions, log files are helpful. Log files contain list of actions that have beenoccurred whenever someone accesses the website or web application. These log files reside in web servers. Every "hit" tothe Website, including each view of a document, image or other object, is logged in a log file. The raw web log file format is one line of text for each hit to the website. This contains information about who was visiting the site, where they came from, and what they were doing on the website.These log files have tons of useful information for service providers,analyzing these log files can give lots of insights that help understand website traffic patterns,user activity, there interest etc.[2][3]. Thus, through the log file analysis we can get theinformation about the people interaction with websitesand applications.

II. RELATED WORK

A data center generates thousands of terabytes or petabytes of log files in a day. It is challenging to store and analyze these huge volumes of log files. The problem of analyzing log files is difficult not only because of its volume butalso because of the structure of the log file. Traditional database techniques are notsuitable for analyzing such log files because they are not capable of handling such a large volume of logs efficiently. Andrew Pavlo and Erik Paulson in 2009 [4] compared the SQL DBMS andHadoopMapReduce and suggested that HadoopMapReduce loads data faster than RDBMS. Also traditional RDBMS cannot handle large datasets. This iswhere big data technologies come to the rescue [5]. Hadoop-MapReduce [6] [5] is applicable in many areas of Big Data analysis. As log files is one of the type of big data so Hadoop is the bestsuitable platform for storing log files and parallel implementation of MapReduce [7] program foranalyzing them. Apache Hadoop is a new way for enterprises to store and analyze data. Hadoop isan open-source project created by Doug Cutting [8], administered by the Apache SoftwareFoundation. It enables applications to work with thousands of nodes and petabytes of data. Whileit can be used on a single machine, its true power lies in its ability to scale to hundreds orthousands of computers. As described by Tom White [6] Hadoop isspecially designed to work on large volume of information by using commodity hardware in parallel. Hadoop breaks up log files into equal sized blocks and these blocks are evenly distributedover thousands of nodes in a Hadoop cluster.Also, it does the replication of these blocks over multiple nodes to provide features like reliability and fault tolerance. Parallel

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

computationof MapReduce improves performance for large log files by breaking job into number of tasks. The Hadoopimplementation shows that MapReduce program structure can be effective solution for analyzing very large weblog files in Hadoop environment [9]. Hadoop-MR log file analysis tool that provides a statistical report on total hits of a web page, user activity, traffic sources was performed in two machines with three instances of Hadoop by distributing the log files evenly to all nodes [10].A generic log analyzer framework for different kinds of log fileswas implemented as a distributed query processing to minimize the response time for the users which can be extendable for some format of logs [11].Hadoopframework handles large amount of data in a cluster for web log mining. Data cleaning, the main part of preprocessing is performed to remove the inconsistent data. The preprocessed data is again manipulated using session identification algorithm to explore the user session. Unique identification of fields is carried out to track the user behavior [12].

III. HADOOP MAP REDUCE

Hadoop is an open source framework for large scale computation and data processing on a cluster of commodity hardware. It allows applications to work with thousands of computational independent computers. 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.

MapReduceis a java based distributed programming model that consists of two phases: a parallel "Map" phase, followed by an aggregating "Reduce" phase. A map function processes a key/value pair (k1, v1, k2, v2) to generate a set of intermediate key/value pairs, and a reduce function merges all intermediate values [v2] associated with the same intermediate key (k2).

Map (k1, v1) → [(k2, v2)]

Reduce (k2, [v2]) → [(k3, v3)]

Maps are the individual tasks that transform the input records into intermediate records. A MapReduce job usually splits the input data set into independent chunks which are processed by the map tasks. The framework sorts the output of the map, which are then input to the reduce tasks. Both the input and the output of the processed job are stored in the Hadoop file-system.

The Hadoopcluster consists of a single NameNode, a master that manages the file system namespace and regulates its access to files by clients. There can be a number of DataNodes usually one per node in the cluster which periodically report to NameNode, the list of blocks it stores. HDFS replicates files for a configured number of times. It automatically re-replicates the data blocks on nodes that have failed. Using HDFS a file can be created, deleted, copied, but cannot be updated. The file system uses TCP/IP for communication between the clusters.

IV. PROPOSED METHODOLOGY AND DISCUSSIONS

Log files usually generated from the web server consist oflarge volume of data that cannot be handled by a traditional database or other programming languages for computation. The proposed work aims on preprocessing the log file using Hadoop is shown in Figure 1.The work is divided into phases, where the storage and processing is made in HDFS.

Web server log files are copied to Hadoop file system. The log file that resides in HDFS is loaded in to Hive table. Then data cleaning is done using Hive query Language. Data cleaning is the first phase carried out in the proposedwork as a pre-processing step in web server log files. The web server log files contains a number of records that corresponds to automatic requests originated by web robots, that includes a large amount of erroneous, misleading, and incomplete information. In the proposed work the web log file containing request from robots, spider and web crawlers are removed. Request created by web robots are not considered as used data, it is filtered out from the log data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

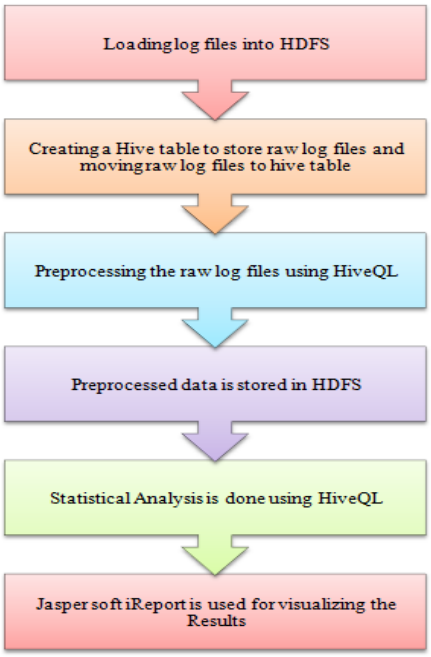

Fig.1. Flow chart describing the methodology

In the preprocessing step the entries that have status of "error" or "failure" have been removed. Also some access records generated by automatic search engine agent is identified and removed from the access log. The important task carried out in data cleaning is the identification of status code. Only the log lines holding the status code value of "200" is identified as correct log. So only the lines having value "200" in status code field are extracted and stored in a Hive table for further analysis.

Than the identification of unique user, unique fields of date, URL referred, and status code are identified. These unique values is retrieved and used for further analysis in order to find the total URL referred on a particular date or the maximum status code got successes on specific date.

In this research Hadoopframework is used to compute the log processing in pseudo distributed mode of cluster. The web server logs of www.ubdtce.org for a period of five months from December 2014 to March 2015 are used for processing in Hadoop environment. The log files are analysed in Centos 6.6 OS with Apache Hadoop 1.1.2 and Apache Hive 0.10.0.

A. ***Pseudo Distributed Mode***

Hadoop framework consist of five daemons namely Namenode, Datanode, Jobtracker, Tasktracker, Secondary namenode. In pseudo distributed mode all the daemons run on local machine simulating a cluster.

B. ***Apache Hive***

Apache Hive [13] is an essential tool in the Hadoopecosystem that provides a Structured Query Language called HiveQL for querying data stored in theHadoop Distributed File system.The log files stored in the HDFS are loaded in to a hive table and cleaning is performed. The cleaned web log data is used to analyse unique user and unique URLs, daily statistics, monthly statistics etc.

C. ***JasperSoftiReport Designer***

JasperSoftiReport Designer is a powerful graphical design tool for report designers. iReport can help to design reports to meet the most complex reporting demands. iReport is built on the NetBeans platform and is available as a standalone application or as a Netbeans plug-in. After pre-processing,by making a JDBC connection to Hive jaspersoft'siReport 5.6 the results stored in HDFS is visualized in the form of graphs and tables.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

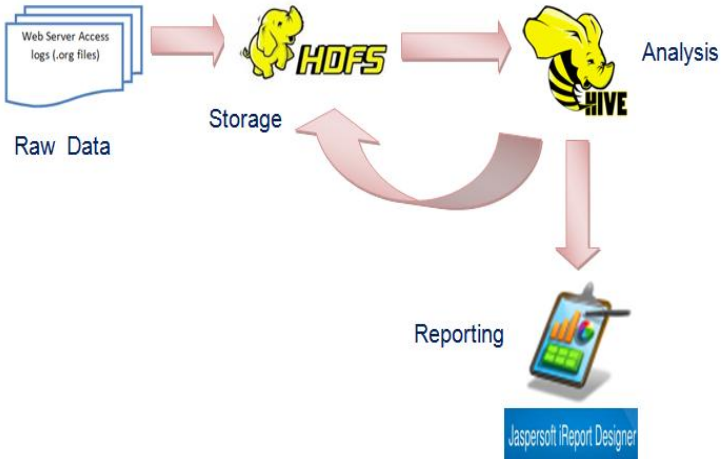

Fig.2. Raw log data processing and visualizing

Figure 2 illustrates copying raw log files into HDFS and then preprocessing is done using Apache Hive data warehouse tool. Then JasperSoft'siReport tool is used to generate the analysis results in the form of graphs and tables.

V. EXPERIMENTAL RESULTS

The major advantage of data cleaning is to produce a quality result and increase in efficiency. Afterperforming Pre-processing step results are shown in table 1.It shows how much reduction happened in the size of data after pre-processing.

	Raw Data	**After Cleaning**
File Size	108.4 MB	9.3 MB
No. of Rows	4, 66,621	47, 039

Table 1. Results Before and After Pre-processing

In the current research web access logs were taken from www.ubdtce.org website for the time period 31/Oct/2014 to 31/Mar/2015and the following results were obtained:

1. ***General Statistics:*** In this section we get general information pertaining to the website like how many times the website was hit, total visitors, bandwidth used etc. It enlists all the general information which one should know related to a website.Table 2. Shows the hits, visits and bandwidth usage of ubdtce.org website for a period of five months.

Summary	
Hits	
Total Hits	466621
Visitor Hits	422591
Visitors	
Total Visitors	47039
Total Unique IPs	4560
Bandwidth	
Total Bandwidth	8698.03 MB
Visitor Bandwidth	8219.00 MB

Table 2. General Statistics obtained after analysing web logs

2. ***Activity statistics:***It provides the statistics on daily and monthly basis. It gives on which days the website was visited maximum. Figure 2 and figure 3 shows the daily and monthly access statistics of www.ubdt.orgwebsite.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

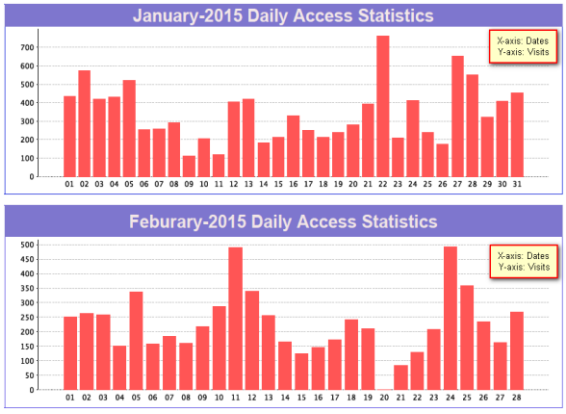

Fig.3. Daily Access Statistics

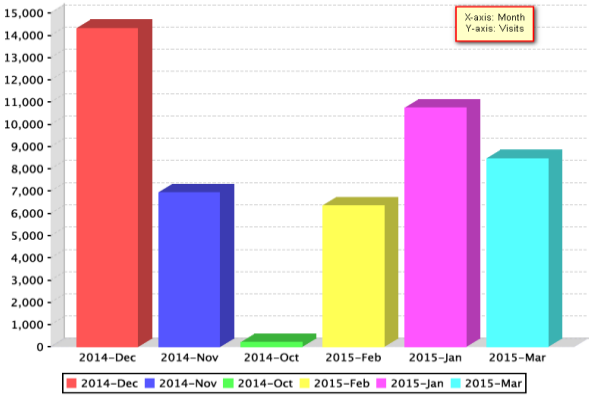

Fig.4. Monthly Access Statistics

Figure 3 shows that more number of visits are on 22nd, 27th, 28th of January and 11th, 12th, 24th February and very less visitors on 9th, 11th of January and 20th of February. Figure 4 shows that more number of visitors are in the month of December and very less visitors in the month of October.

3. ***Access Statistics:***This part of the analysis can be considered the most important as it provides which IP is producing more hits and more visits and which IP is using high bandwidth. It helps in determining that who all accessed the website. The table3 shows a list of IP addresses that hit the website along with how many times the website was visited by a particular user and how much bandwidth used by each user.

Host	Hits	Visitors	Bandwidth (MB)
14.139.152.34	29772	4371	826
216.158.82.218	9391	9262	118
14.139.155.178	1805	143	34
71.198.24.238	1604	93	6
117.241.0.112	1165	214	12
14.141.216.130	1133	180	19
112.133.192.42	1029	150	27
117.240.86.5	811	101	15
37.228.105.7	792	30	7
117.211.56.9	768	208	11

Table 3. Access Statistics

4. ***Visits-per-Country:***The table shows Number of visits to the website based on countries.

Country Code	Visits
IN	25465
US	11099
FR	547
CN	297
UA	124
CA	115

Table 4. Visits per Country

5. ***Errors:***The last feature isfinding out what kind of errors people face when they access the website. The figure 5 shows the errors users encountered when they accessed the website.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

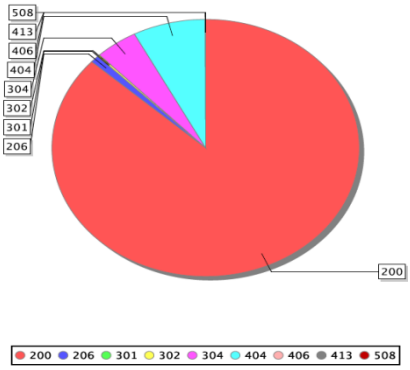

Fig.5. Pie chart showing the errors that occur frequently

VI. CONCLUSIONS

In this paper we applied HadoopMapReduce programming model for analyzing web server log files where data get stored on multiple nodes in a cluster so that access time required can be reduced and MapReduce works for large datasets giving efficient results. In order to have summarized results for a particular web application, we need to do log analysis that will help to improve the business strategies as well as to generate statistical reports. Using Visualization tool for log analysis will provide us graphical reports showing hits for web pages, user's activity, in which part of website users are interested, traffic sources, etc. From these reports business communities can evaluate which parts of the website need to be improved, which are the potential customers, from which geographical region website is getting maximum hits, etc., which will help in designing future marketing plans. Log analysis can be done by various methods but what matters is response time. HadoopMapReduce framework provides parallel distributed processing and reliable data storage for large volumes of log files. Here hadoop's characteristic of moving computation to the data rather moving data to computation helps to improve response time.

REFERENCES.

1. SayaleeNarkhede and TriptiBaraskar, "HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.
2. Liu Zhijing, Wang Bin, (2003) "Web mining research", International conference on computationalintelligence and multimedia applications, pp. 84-89.
3. Yang, Q. and Zhang, H., (2003) "Web-Log Mining for predictive Caching", IEEE Trans.Knowledge and Data Eng., 15(4), pp. 1050-1053.
4. Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden,Michael Stonebraker, (2009)"A Comparison of Approaches to Large-Scale Data Analysis", ACMSIGMOD'09.
5. Mr.YogeshPingle, VaibhavKohli, ShrutiKamat, NimeshPoladia, (2012)"Big Data Processingusing Apache Hadoop in Cloud System", National Conference on Emerging Trends inEngineering & Technology.
6. Tom White, (2009) "Hadoop: The Definitive Guide. O'Reilly", Scbastopol, California.
7. Jeffrey Dean and Sanjay Ghemawat., (2004) "MapReduce: Simplified Data Processing on LargeClusters", Google Research Publication.
8. Apache-Hadoop,http://www.hadoop.apache.org
9. Chen-Hau Wang, Ching-Tsorng Tsai, Chia-Chen Fan, Shyan-Ming Yuan, "A Hadoop Based Weblog Analysis System", 2014 7th International Conference on Ubi-Media Computing and Workshops.
10. SayaleeNarkhede and TriptiBaraskar, "HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.
11. MilindBhandare, VikasNagare et al., "Generic Log Analyzer Using HadoopMapreduce Framework," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, September 2013.
12. Savitha K, Vijaya M S, "Mining of web server logs in a distributed cluster using big data technologies", International Journal of Advanced Computer Science and Applications, Vol.5, NO.1, 2014
13. Edward Capriolo, Dean Wampler, and Jason Rutherglen, (2012) "Programming Hive. O'Reilly"

BIOGRAPHY

Harish Sis currently pursuing his M. Tech. in Computer Science & Engineering degree from University B.D.T College of Engineering, Davangere, India. He has received B.E degree in Computer Science and Engineering from Siddaganga Institute of Technology, Tumkur under Visvesvaraya Technological University, Belgaum. His research interests are Big Data Analytics and Web Mining.

Mrs.Kavitha G is working as Assistant Professor in Department of studies in CSE, University B.D.T College of Engineering, Davangere, India.