Customer behavior analysis of web server logs using Hive in Hadoop Framework

Lavanya KS, Srinivasa R
Dept. of Studies in CSE, RRCE College of Engineering, Bangalore, Karnataka, India
Email: lavanyaa.ks@gmail.com , srinivasa.r@gmail.com

Abstract: Web log file is a log file created and stored by a web server automatically. Analyzing such web server access logs files will provide us various insights about website usage. Due to high usage of web, the log files are growing at much faster rate with increase in size. Processing this fast growing log files using relational database technology has been a challenging task these days. Therefore to analyze such large datasets we need a parallel processing system and a reliable data storage mechanism (Hadoop). Hadoop runs the big data where a massive quantity of information is processed via cluster of commodity hardware. In this paper we present the methodology used in pre-processing of high volume web log files, studying the statics of website and learning the user behavior using the architecture of Hadoop MapReduce framework, Hadoop Distributed File System, and HiveQL query language.

Keywords: big data, customer behavior analysis, hadoop, log analysis, web server logs

1. INTRODUCTION

In today's competitive environment, manufacturers/service providers are keen to know whether they provide the best service/product to customers or whether customers look forward to get their service or buy product. Service providers should need to know how to make their websites or web application interesting to customers and how to improve advertising strategies to attract them. All these questions can be answered by Log files. Log files contain a list of actions that occurred whenever customer accesses the service provider's website or web application. Every ―hit‖ to the Website will be logged in a log file. These log files are stored in web servers. The raw web log file is one line of text for each hit to the website and contains information about who visited the site, where they came from, and what they did on the website. These log files carry a useful information for service providers so analyzing these log files can give them insights about website traffic patterns, user activity, customer interest etc.

2. RELATED WORK

As data center generates thousands of terabytes or petabytes of log files a day it is highly challenging to store and analyze such high volumes of log files. Analyzing log files looks complicate because of their high volume and complicate structure. Traditional database techniques have failed to handle these log files efficiently due to large size. In 2009, Andrew Pavlo and Erik Paulson compared the SQL DBMS with Hadoop MapReduce and suggested that Hadoop MapReduce loads data sooner than RDBMS. Also traditional RDBMS cannot handle large datasets.

This is where big data technologies play a major role in handling large sets of data. Hadoop is the best suitable platform that stores log files and does parallel implementation of MapReduce program. For enterprises Apache Hadoop is a new way to store and analyze data.

Hadoop is an open-source project created by Doug Cutting under the administration of the Apache Software Foundation. Hadoop enables applications to work with thousands of nodes with petabytes of data. While it can be used on a single machine, its true capability lies in scaling to hundreds or thousands of computers. Tom White describes Hadoop is specially designed to work on large volume of data using commodity hardware in parallel. Hadoop breaks log files into equal sized blocks and these blocks are evenly distributed among thousands of nodes in cluster. Further, it does the replication of these blocks over multiple nodes to provide reliability and fault tolerance.

In case of large log files parallel computation of MapReduce improves performance by breaking job into many tasks. Hadoop implementation shows that MapReduce program structure can be an effective solution to analyze large volume of weblog files in Hadoop environment. In my project Hadoop-MR log file analysis tool, which provides a statistical report on total hits of a web page, traffic sources and user activity, was performed on two machines with three instances of Hadoop by distributing log files evenly among all nodes. A generic log analyzer framework for different kinds of log files was executed as distributed query processing to minimize response time for the users that can be extendable for some format of logs. Hadoop framework handles large volume of data in a cluster for web log mining. Data cleaning, major part of preprocessing was performed to remove inconsistent data. The preprocessed data was again manipulated using session identification algorithm to explore the user session. Unique identification of fields was carried out to track the user behavior.

3. HADOOP MAP REDUCE

Hadoop is an open source framework used for large scale computation and processing on a cluster of commodity hardware. It permits applications to work along with thousands of independent computers. The important characteristic of Hadoop is to move computations on the data rather than move data for computation. Hadoop is mainly used to breakdown large amount of input data into smaller chunks and each can be later processed separately on different computers. It implements a MapReduce programming model to achieve parallel execution.

MapReduce is a java based distributed programming model consisting of two phases: a parallel ―Map‖ phase, followed by an aggregating ―Reduce‖ phase. Map function processes a key/value pair (k1, v1, k2, v2) to create a set of intermediate key/value pairs. On the other hand, reduce function merges all intermediate values [v2] that are associated with the same intermediate key (k2).Map (k1, v1) → [(k2, v2)],Reduce (k2, [v2]) → [(k3, v3)].Maps are the individual tasks that convert input records into intermediate records. A MapReduce task usually splits the input data into individualistic chunks that are then processed by the map tasks. The framework sorts the output data of the map that are then sent to reduce tasks. Both the input and the output data are stored in the Hadoop file-system.

The Hadoop cluster carry a single NameNode, a master that manages the file system namespace and synchronize its access to files by clients. There can be number of DataNodes (usually one per node in the cluster) that report the list of blocks it stores to NameNode periodically. HDFS replicates files for a configured number of times and re-replicates automatically the data blocks on nodes that have failed. With the help of HDFS any file can be created, copied and deleted but cannot be updated. The file system uses TCP/IP to communicate between the clusters.

4. PROPOSED METHODOLOGY AND DISCUSSIONS

Log files normally generated from the web server consist of large volume of data that cannot be handled by a traditional database or additional programming languages for computation.

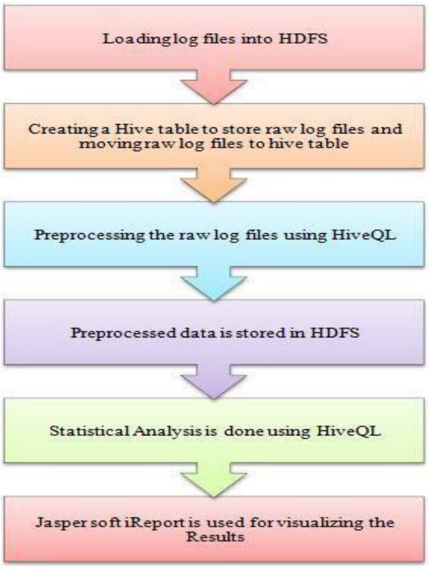
Loading log files into HDFS
Creating a Hive table to store raw log files and moving raw log files to hive table
Preprocessing the raw log files using HiveQL
Preprocessed data is stored in HDFS
Statistical Analysis is done using HiveQL
Jasper soft iReport is used for visualizing the Results

Fig.1 flow chart describing the methodology

The proposed work aims on preprocessing the log file using Hadoop is shown in Fig 1.The work is split into phases, where the storage and processing is made in HDFS.

Web server log files are first copied into Hadoop file system and then loaded to Hive table. Data cleaning, which is done using Hive query Language, is the first phase carried out in our project as a pre-processing step. Web server log files consist of number of records that correspond to automatic requests generated by web robots. The records usually carry a large volume of misleading, erroneous, and incomplete information. In our project web log files carrying requests from robots, spider and web crawlers are removed. Notably, requests created by web robots are not considered as useful data and are filtered out from the log data

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

The next step is to identify unique user, unique fields of date, status code, and URL referred. These unique values are retrieved and used for further analysis to find the total URL referred on a particular data or the maximum status code with successes on precise date.

In our project Hadoop framework is used to compute the log processing through 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 framework.the log files are mainly analyzed with the help of Centos 6.6 OS with Apache Hadoop 1.1.2 and Apache Hive 0.10.0.

Pseudo Distributed Mode.

Hadoop framework incorporate five daemons namely Namenode, Datanode, Jobtracker, Tasktracker, Secondary namenode. In case of pseudo distributed mode all the daemons are run on local machine that actually stimulates a cluster.

Apache Hive.

Apache Hive [13] is an important tool in the Hadoop ecosystem that provides a Structured Query Language called HiveQL to query the data stored in Hadoop Distributed File system. The log files that are stored in the HDFS are loaded in to a hive table and cleaning will be performed. The cleaned web log data is used to analyze daily statistics ,unique user and unique URLs, monthly statistics etc.

JasperSoftiReport Designer.

JasperSoftiReport Designer is a powerful graphical design tool used by report designers. iReport can help to design reports to meet a myriad of retorting needs. iReport is mainly 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.

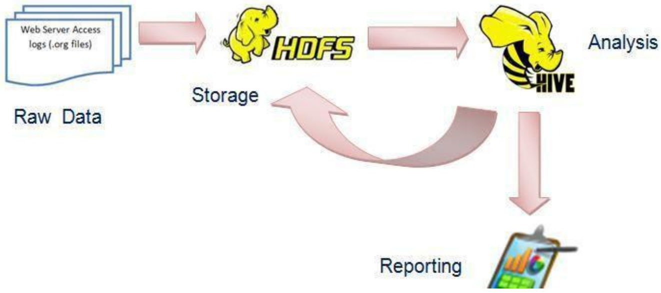

Fig.2 Raw log data processing and visualizing

The above flow chart illustrates copying raw log files in to HDFS and then preprocessing is done with the help of Apache Hive data warehouse tool. Next JasperSoft‘s iReport tool is used to produce the analysis results in the form of graphs and tables.

5.EXPERIMENTAL RESULTS

One of the main advantage of data cleaning is it produces quality results with increased efficiency. The results from Pre-processing step are shown in table below. The results indicate how much data reduction in size has taken place.

	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 our project web access logs were taken from www.ubdtce.org website for time period of 31/oct/2014 to 31/mar/2015 and the following results were obtained.

General Statistics

In this section we get general information related to the website like how many times the website was hit, total number of visitors, bandwidth used etc. It lists out all the information that one should know about the websites. The below table indicates number of hits, visits and bandwidth consumption 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 accessed After Analyzing Web Logs

Activity statistics

This section provides statistics on daily and monthly basis. It also gives on what days the website had visited maximum. Fig.2 and Fig.3 shows the daily and monthly obtain statistics of www.ubdt.orgwebsite.

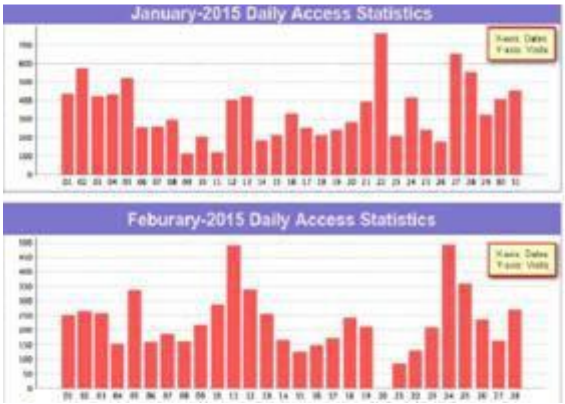

Fig.3 daily access statistics

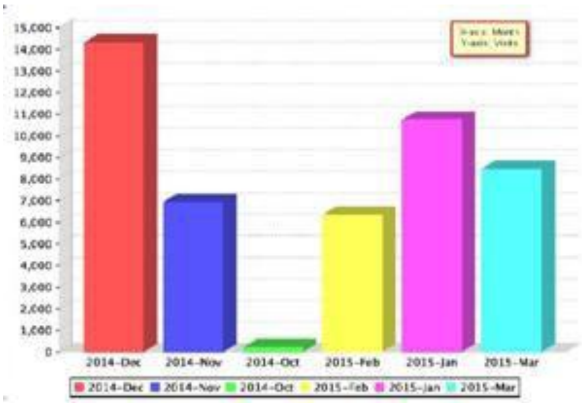

Fig.4 monthly access statistics

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

Access Statistics

This part of our project can be considered as the most important part as it provides which IP produces more hits and more visits and which IP uses high bandwidth. It also helps in determining who all accessed the website. The below table indicates 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 each user used.

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

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

Errors

The last feature is find what kind of errors people encounter when they access the website. The below chart indicates the errors users encountered while they accessed the website.

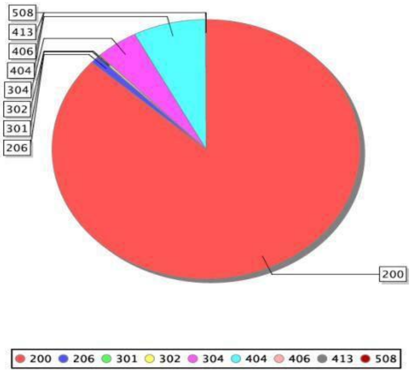

Fig.5 pie chart showing the errors that occur frequently

6.CONCLUSIONS

Web sites are one of the important means for organizations for making advertisements. In order to get outlined results for a specific web site, we need to do log examination that helps enhance the business methodologies and also produce measurable reports. In this project with the help of Hadoop framework web server log files are analyzed. Data gets stored on multiple nodes in a cluster so the access time required is reduced. MapReduce works for large datasets giving efficient results.Using visualization tool for log analysis will give us graphical reports indicating hits for web pages, client‘s movement, in which part of the web site clients are interested. From these reports business groups can assess what parts of the site need to be enhanced, who are the potential clients, what are the regions from which the site is getting more hits, and so on. This will help organizations plan for future marketing activities. Log analysis can be done using many different techniques however what is important is response time. Hadoop MapReduce model provides parallel distributed processing and reliable data storage for huge volumes of web log files. Hadoop‘s ability of moving processing to data rather than moving data to processing helps enhance response

REFERENCES.

[1]. SayaleeNarkhede and TriptiBaraskar, –*HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce*,‖ International Jour nal 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] 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.

[9] SayaleeNarkhede and TriptiBaraskar, –*HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce*,‖ International Jour nal of UbiComp (IJU) vol.4, No.3, July 2013.

[10] 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.

[11] 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

BIOGRAPHY

Lavanya KS Sis currently pursuing his M. Tech. in Computer Science & Engineering degree from University RRCE College of Engineering, Bangalore, India. She has received B.E degree in Computer Science and Engineering from HMS Institute of Technology, Tumkur under Visvesvaraya Technological University, Belgaum. Her research interests are Big Data Analytics and Web Mining.

Mr.Srnivasa R is working as Associate Professor in Department of studies in CSE, University RRCE College of Engineering, Bangalore, India