IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 01-04

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

Sunil Bhishe[1], Aman Gaikwad[2], Kunal Kshirsagar[3], Nikhita Vaidya[4], Priyanka Hatwar[5] Neha Mogre[6]
[1,2,3,4,5]UG Student, [6]Assistant Professor
Department. of CSE, TGPCET, Nagpur, Maharashtra
[1] sunilbhishe07@gmail.com, [2]amangaiwad786@gmail.com, [3]kunalkshirsagar05@gmail.com,
[4]vaidyanikhita@gmail.com, [5] priyankahatwar317@gmail.com, [6] neha.cse@tgpcet.com

Abstract: *The growth of websites and the Internet has opened up new research, social, entertainment, education and business opportunities. With the fast growth of the Internet, the digital data generated by the websites is becoming so massive that the traditional text software and relational database technology faces a bottleneck while processing such massive data and the results generated by these technologies are not satisfactory. Cloud computing offers a good solution for this problem. Cloud computing is not only capable of storing such massive data but also capable of processing and analyzing such voluminous data faster, by making use of distributed storage and distributed computing technology. A weblog is a group of connected web pages that consists of a log or daily record of information, particular fields or views which is altered, every now and then, by owner of site, other websites or by website users. An enterprise weblog analysis system based on Hadoop architecture with Hadoop Distributed File System (HDFS), HadoopMapReduce Software Framework and Pig Latin Language aids the business decision-making process ofthe system administrators and helps them to collect and identify the potential value which is hidden within such huge data generated by the websites. Such a weblog analysis includes the analysis of an Internet site's entry log as well as provides information about the amount of visitors, days of week and rush hours, views, hits, very often accessed pages, application server traffic trends, performance reports at varying intervals and statistical reports which indicate the performance of program.*

Keywords*: Big data; hadoop; mapreduce; web server logs; log analysis; hive.*

I. Related Works

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 [3] 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 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 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 [3],

II. Introduction

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. Due to the advancement of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. Website statistics are based on serverlogs.IT organizations analyze server logs to answer questions about security and compliance. A server log is a simple text file which records activity on the server. Computer generated logs that capture data on the operations of a network**.** Useful for managing network operations, especially for security andregulatory compliance. There are several types of server log — website owners are especially interested in access logs which record hits and related information. These logs are in large amount thus resulting collection of large amount of data i.e Big Data. Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. It includes huge volume, high velocity, and extensible variety of data. This data can be in structured, semi structured or in unstructured form**.** The fast development of the Internet has led to increased usage of the Internet in people's daily life, which has led to accelerated growth of web logs. This has posed various problems as regards to handling of the weblogs in a timely manner, extracting of the information required by the people from the massive weblogs generated. A single computer cannot handle weblog satisfactorily to meet the needs of the people. A Hadoop-based Weblog Analysis System combines Cloud Computing and Hadoop technology processes all the gathered logs as well as carries out a distant parallelization study that solves the difficulties of conventional setups such as simultaneous data handling and gathering of data.

III. Background

The Hadoop framework gives us dependable data storage with the help of Hadoop Distributed File System as well as MapReduce calculating standard that acts as a parallel operating setup over huge sets of data. A Hadoop distributed file system splits the initial data as well as provides the initial data fragments across various computers across HDFS cluster. These machines across the hadoop cluster carry blocks of data which enables the processing log data in parallel and evaluates the result efficiently. The superior hadoop methodology is to “Save initially and query afterwards”. Initially, Hadoop puts entire data over the Hadoop Distributed File System. After this is completed, Hadoop executes the queries which are in Pig Latin language which enables to lessen the time for reply and the load against the user setup .Pig is modelled to blend well within explanatory methodology of SQL as well as the procedure-oriented map-reduce methodology associated with either the machine-code or an assembly language which is inflexible, resulting into a large amount of tailor-made consumer computer program that proves difficult for managing as well as reutilizing. After being completely executed, Pig performs the task of compiling Pig Latin in the form of concrete designs. These concrete designs get implemented across Hadoop. Hadoop is a publicly accessible, implementation of map-reduce. An Apache-incubator project such as Pig is open-source and accessible for public usage. Pig drastically lessens the time which is needed for performing generation as well as implementation related to their data study activities, in contrast with the time taken when Hadoop is utilized alone. A new debugging environment is obtained when Hadoop is integrated with Pig which results in a very high efficiency leading to increased profitability.

IV. Existing System

The current processing of log files goes through ordinary sequential ways in order to perform preprocessing, session identification and user identification. The non-Hadoop approach loads the log file dataset, to process each line one after another. The log field is then identified by splitting the data and by storing it in an array list. The preprocessed log field is stored in the form of hash table, with key and value pairs, where key is the month and value is the integer representing the month. In existing system work is possible to run only on single computer with a single java virtual machine (JVM). A JVM has the ability to handle a dataset based on RAM i.e. if the RAM is of 2GB then a JVM can process dataset of only 1GB. Processing of log files greater than 1GB becomes hectic. The non-Hadoop approach is performed on java1.6 with single JVM.We use Excel or similar software to produce statistical information and generate reports.

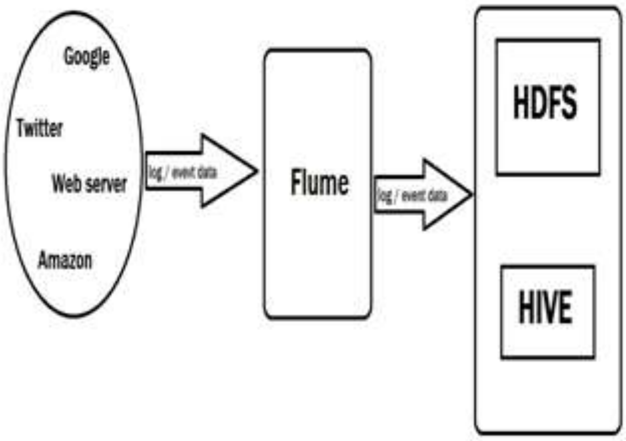

Fig. System Architecture

V. Proposed System

Proposed solution is to analyze web log generated by Apache Web Server. This is helpful for statistical analysis. The size of web log can range anywhere from a few KB to hundreds of GB. Proposed mechanism design solution that based on different dimensions such as timestamp, browser, and country. Based on these dimension, we can extract pattern and information out of these log and provides vital bits of information. The technologies used are Apache Hadoop framework, Apache flume Proposed system uses four node environments where data is manually stored in local hard disk in local machine. This log data will then be transferred to HDFS using FLUME framework. FLUME has agents running on Web servers. This log data is processed by MapReduce to produce Comma SeparatedValues.

VI. Methodology

APACHE WEB SERVER

A Web server is a program that uses HTTP (Hypertext Transfer Protocol) to serve the files that form Web pages to users, in response to their requests, which are forwarded by their computers' HTTP clients. Dedicated computers and appliances may be referred to as Web servers as well.The process is an example of the client/server model. All computers that host Web sites must have Web server programs. Leading Web servers include Apache (the most widely-installed Web server), Microsoft's Internet Information Server (IIS) and nginx (pronounced engine X) from NGNIX. Other Web servers include Novell's NetWare server, Google Web Server (GWS) and IBM's family of Domino servers.

Web Server Logs

A server log is a log file (or several files) automatically created and maintained by a server consisting of a list of activities it performed. A typical example is a web server log which maintains a history of page requests. The W3Cmaintains a standard format (the Common Log Format) for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested.

Hadoop Distributed File System (HDFS)

HDFS provides a means for greater rate of processing over the data belonging to a program. Hadoop File System which was created by utilizing a DFS design is run over affordable and easy to obtain physical components of a computer. HDFS is different from other distributed systems due to its great fault-resilient nature and its design which utilizes cheap hardware. HDFS carries huge data as well as gives a simpler access mechanism. To save so much large amount of data, the files are saved throughout numerous computers. The acquired files are saved in repeating style to save the system against probable failure of data. HDFS makes programs ready for managing in a parallelmanner.

Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

Hive

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Hive is a data warehouse infrastructure software that can create interaction between user and HDFS. The user interfaces that Hive supports are Hive Web UI, Hive command line, and Hive HD Insight (In Windows server). Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping. HiveQL is similar to SQL for querying on schema info on the Metastore. It is one of the replacements of traditional approach for MapReduce program. Instead of writing MapReduce program in Java, we can write a query for MapReduce job and process it.

VII. Conclusion & Future Scope

7.1 CONCLUSION

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, clients 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.

7.2 FUTURE SCOPE

The cloud computing research is progressing extremely fast to put together Cloud Computing with Hadoop architecture containing Hadoop Distributed File System (HDFS), Hadoop MapReduce Software Framework as well as the Pig Latin Language to expose novel trends for performing more productive as well as more quick collection, storage and analysis of huge weblogs produced by a huge number of Internet users on a daily basis. We hope to see better requirement analysis of logs and a better performance by such a Weblog Analysis System based on Hadoop that gives a reference for doing performance-tuning that enables us to foretell the future trend of a system.

References

[1]. Jeffy Dean, Sanjay Ghemawat. “MapReduce: Simplified Data Processing on Large Clusters”, OSDI04: Sixth Symposium on Operating System Design and Implemention, Ssn Francisco, CA, December, 2016.

[2]. L.K. JoshilaGrace, V.Maheswari, Dhinaharan Nagamalai ,“ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING”,International Journal of Network Security & Its Applications (IJNSA), Vol.3,No.1, January 2016

[3]. Chen-Hau Wang, Ching-Tsorng Tsai, Chia-Chen Fan, Shyan-Ming Yuan, “A Hadoop Based Weblog Analysis System”, 2016 7th International Conference on Ubi- Media ComputinandWorkshops.

[4]. J. Dean.S. Ghemawat, ”MapReduce: Simplified Data Processing on Large Clusters”, Int’l Conf of Operating Systems Design and Implementation (OSID), San Francisco, CA, pp. 137-150.

[5]. Milind Bhandare, Vikas Nagare et al., “Generic Log Analyzer Using HadoopMapreduce Framework,” International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, September 2015.

[6]. Thanakorn Pamutha, Siriporn Chimphlee and Chom Kimpan, “Data Preprocessing on Web Server Log Files for Mining Users Access Patterns” .International Journal of Research and Reviews in Wireless Communications of Vol. 2, No. 2, ISSN: 2046- 6447 ,June 2012.

[7]. M. Zaharia, D. Borthakur, J. S. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, “Job scheduling for multi-user map reduce clusters,”EECS Department, University of California, Berkeley, Tech. Rep.

[8]. Mackey, G. Sehrish, S. Jun Wang, “Improving metadata management for small files in HDFS” IEEE International Conference on luster Computing and Workshops, pp.1–4, Aug. 2014.

[9]. Jie Yang, Yansheen ZZhang, Shuo Zhang, Dazhong HE, “Mass flow logs analysis system based on Hadoop”, Proceedings ofIEEE.

[10]. Bharat Parte, Umesh Jamdade, Pranita Sonavane, Sheetal Jadhav, “SQUID Log Analyzer Using Hadoop Framework”, Journal of Innovative Research in Engineering & Management (IJIREM), Volume-2, Issue-1, January-2015