WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analyse of the Unstructured Logs Data Use Machine Learning Technique with Distributed Processing Tools

Parth Yadav*, Anshu Tiwari**, Manish Saxena***

[a]*Student , BST, Bhopal., M.P. ,India*
[b]*Asst. Prof, Bhopal., BST, M.P. ,India*

A B S T R A C T

Server log may be an easy document that records activity on the server. Computers generated logs contain information regarding various operations of a network. There are varieties of server log such as access logs and error logs. The web site owners are particularly curious about accessing logs that record hits and connected information. These logs are in quantity so ensuring assortment of enormous amount of knowledge known as Big Data. Massive knowledge or a giant knowledge, it is a set of enormous data sets that can't be processed by traditional computing techniques. It includes huge volume, high rate, and protractible style of knowledge. This knowledge is in structured, semi structured or in unstructured form.

Keywords: Hadoop, HDFS, Mapreduce, Log analyzer.

1. INTRODUCTION

(World Wide Web) is growing rapidly, so many number of users are connected to the internet and generate huge amount of log data every day. Identifying the user behavior is very important to utilize the information of web usage pattern. To achieve this, various data mining techniques are used to find out the hidden information and find patterns from the log data to find the user interest and the navigation pattern. Through this we can find the frequently accessed pages, ip hit ratio and the user navigation. But the data generated now-a-days is growing rapidly therefore analyzing these data by using traditional approach is not feasible.

Therefore we need a powerful tool to store and analyze these huge amount of log data. In this dissertation we present the Hadoop framework which is very reliable for storing such huge amount of data into HDFS and analyze the unstructured logs data

The most supply of information is that the net access log.net server logs area unit plain text (ASCII) files, freelance of server platform. There are unit some variations between server package, however historically there are unit four kinds of server logs [5]

1. Transfer (access) log
2. Error log
3. Referrer log
4. Agent log

Computer applications and browsing based majorly suffer from latency or web page prediction (user requirement) issues. Recently various efforts are made for finding a stable and efficient solution for improving the web page prediction. With the help of this prediction we can safe our customer or user can easily reach their own requirement as well as we can increase our own business based on web page prediction.

WEB PREDICTION

On the basis of user previous web behavior we can predict the next page this is known as web prediction. This prediction is done for improving the Web page pre-fetching. Today the massive use of the web has increased the traffic in the network as well as the load that the web servers manage.

LOG FILE ANALYSIS

The process of recording events in a file during the execution of the operating system, process, system, network, virtual machine, or application is called "logging" and the file is called a "log file". A Log File is composed of log entries and each log entry contains useful information associated with events that occur in the system, network, virtual machine, or application.

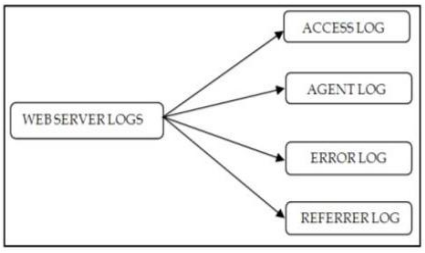

Fig.1 Web Server log files

Web Usage Mining

Web usage mining aims to determine some helpful patterns from the Web usage data, such as, click streams, user transactions and user¨s web access activities, which are often stored in server side Web logs. A server side Web log records user sessions of visiting Web-pages of a website day by day.For storing and analyzing large amount of log data we need a powerful tool, therefore we are using Hadoop. These data is analyzed by Machine Algorithm.

Mining and Analysis Requirement

Following requirement for mining and analysis is:- Apache Hadoop, Apache Hive, Apache Spark

HDFS

HDFS stands for Hadoop Distributed file system. It is designed in order that terribly massive size files will be keeping with streaming information access patterns, running on clusters of goods hardware.

MapReduce

MapReduce is such a framework that was initial represented by Jeffrey Dean and Sanjay Ghemawat (both functioning at Google).Hadoop MapReduce [6] is a computer code programming model to method giant datasets on clusters in associate degree economical manner. A MapReduce task divides the computer file into variety of chunks that area unit appointed to totally different DataNodes and area unit processed in parallel manner.

Methodology

For storing and analyzing these large and complex data we need a powerful tool, therefore we use Hadoop for storing large amount of log data into HDFS. These data is analyzed by using SQL so we uses spark-SQL which works on spark framework and HiveQL which works on MapReduce framework.

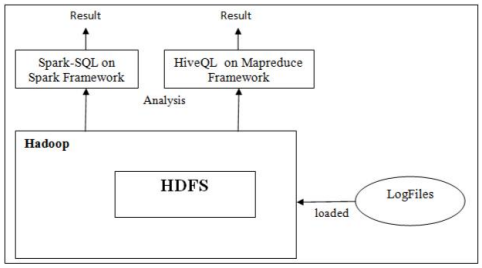

Figure.2. Workflow Diagram

Proposed Algorithm

The key step in Apriori algorithm is to find the frequent itemsets. In the kth iteration, it computes the occurrences of potential candidates of size k in each of the transactions. It is obviously that the occurrences counting of candidate itemsets in one transaction is irrelevant with the counting in another transaction in the same iteration. Therefore, the occurrences computation process in one iteration could be parallel executed.

Apriori algorithm, distributed which is denoted as Apriori. The steps are as follows.

Step1. Use MapReduce model to find the frequent 1-itemsets.

Step2. Set k = 1.

Step3.If the frequent (k+1)-itemsets cannot be generated, then goto Step6.

Step4. According to the frequent k -itemsets, use MapReduce model to generate the frequent (k+1)-itemsets.

Step5. If k is less than the max iteration times, then k + +, goto Step3; Otherwise, continue to the next step.

Step6. According to the frequent itemsets L, generate the strong rules.

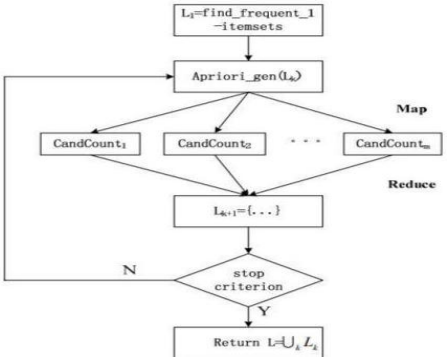

Figure.3 Flow Chart of Algorithm

Map-function

The input dataset is stored on HDFS as a sequence file of pairs, each of which represents a record in the dataset. The key is the offset in bytes of this record to the start point of the data file, and the value is a string of the content of this record.

The result analysis of all the experiments were performed using an i3-2410M CPU @ 2.30 GHz processor and 2 GB of RAM running ubuntu14.Here we are using Hadoop for reliable storage as well as parallel processing. So, to achieve this we are going to follow the following methods:

- Loading Data into HDFS.
- Analysing using spark-SQL.
- Analysing using HiveQL.
- Compare the performance.

Loading Data into HDFS

We load the access logs dataset in to HDFS. Figure shows loading a log file into HDFS. In this figure we can clearly see that there is no structure between the data of these logs file.

Figure .4 logs data stored into HDFS

Analyzing using Mapreduce

After the analysis is done by using spark-SQL we analyse the same dataset by using another SQL like engine which works on mapreduce framework. We can write sql on and compiler validate the sql query and convert sql into mapreduce program which is executed by mapreduce framework and we can get the output.

CONCLUSION and future work

There is a huge growth in log data which is generated from the various websites. So storing and processing these log data using the traditional approach is not feasible so we need a powerful tool to store and analyze these huge amount of log data. For storing such a huge amount of data we have discussed an HDFS which is very reliable in terms of storage on a cluster of commodity hardware. In this dissertation, data storage is provided using HDFS and analysis by Machine learning techniques. We can say that spark holds better efficiency as compared to Mapreduce. In future field of web page ranking for web personalization to understand the available methods and assist to perform their research in right direction. For future work, the application logs are analyze real time, because real time analysis give quick result to take any decision.

REFERENCES

Dr.S.Suguna, M.Vidhya,Mr. J.I.Christy Eunai, “Big Data Analysis in E-commerce System using hadoop mapreduce” in IEEE 2016 .

Bing Liu. Web data mining; Exploring hyperlinks, contents, and usage data, chapter 11: Opinion Mining. Springer, 2006.

Sandeep Kumar Dewangan, Shikha Pandey and Toran Verma, “A Distributed Framework for Event Log Analysis using MapReduce” at International Conference on Advanced Communication Control and Computing Technologies(ICACCCT) in 2016.

J.Srivastava et al, “Web usage Mining: Discovery and Applications of usage patterns from Web Data “, ACM SIGKDD Explorations, vol.1, Issue. 2, pp.12-23,2000.

Mr.Amin Nazir Nagiwale, Mr. Manish R. Umale, Mr. Aditya Kumar Sinha, “Design of Self-Adjusting algorithm for data-intensive MapReduce Applications” at International Conference on Energy Systems and Applications (ICESA 2015), Dr. D. Y. Patil Institute of Engineering and Technology, Pune, India 30 Oct -01 Nov, 2015.

Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet kaur, Navjot Kaur,“Research Paper on Big Data and Hadoop”, IJCST, V.7, Issue 4.

Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD Record. ACM. 1993;22: 207–216

Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.