WEB LOG ANALYSIS

FIRST REVIEW REPORT

Submitted by

LAVANYA BANDLA (19MCB0008)

KALYANI ANJANKAR (19MCB0011)

D.SANDEEP KUMAR (19MCB0023)

Prepared For

BIG DATA FRAMEWORK (CSE6001) – PROJECT COMPONENT

Submitted To

RAMESH BABU K

Associate Professor

School of Computer Science and Engineering

Table of Contents

2.PROBLEM STATEMENT

A sample data-set is given of the websites and browsers that people visit on a daily basis. For the first problem query, find out the most viewed page and for the second, find total hits per unique day as the project is all about the analysis using Apache Pig.

3.OBJECTIVE

A significant development in the field of technology in sectors such as business, public and private has been observed leading to accumulation of large data over the web. Information acquired from the web are used to describe the exponential growth and availability of data, both structured and unstructured. As data over the web is heterogeneous in nature, analyzing such data is necessary in order to gain acquaintance wherein log file analysis is an effective solution. Log files are the files that list the actions that have been occurred and reside in web server. There prevails a need to process and store log files using traditional techniques however in the enterprise scenario the data from these log files is out-sized due to which processing capacity of conventional approaches becomes incompetent for gaining information for processing.

4.LITERATURE REVIEW

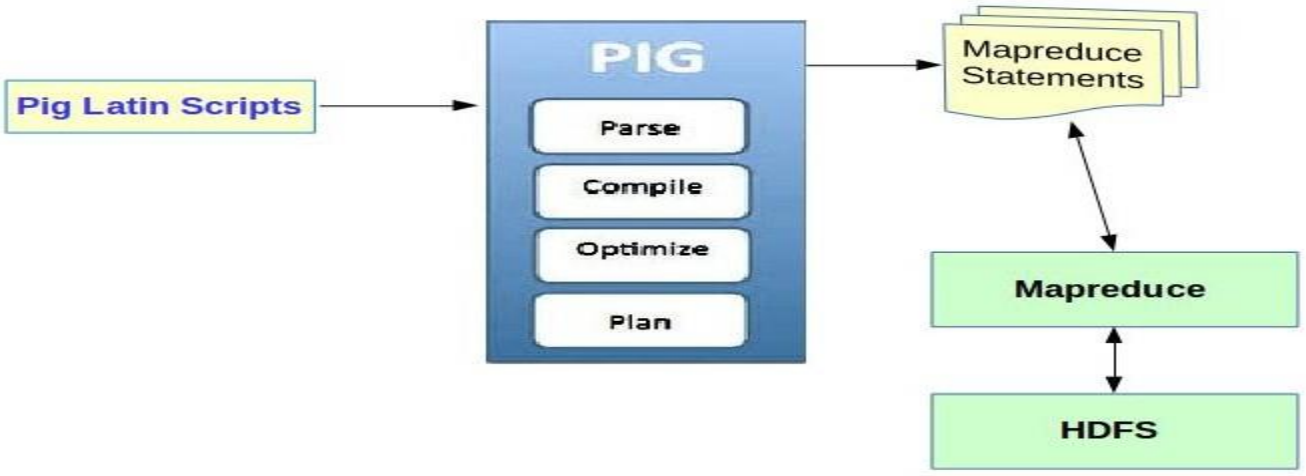
Sayalee Narkhede and Tripti Baraskar, "hmr log analyzer: analyze web application logs over hadoop mapreduce", International Journal of UbiComp (IJU), Vol.4, No.3, July 2013: Web server logs stores click stream data which can be useful for mining purposes. The data is stored as a result of user's interaction with a website. Web usage mining is an application of data mining which can be used to discover user access patterns from weblog data. The obtained results are used in different applications like, site modifications, business intelligence, system improvement and overspecialization. In this particular study they have analyzed the log files of smart sync software web server to get information about visitors; top errors which can be utilized by system administrator and web designer to increase the effectiveness of the web site[1].

] Navin Kumar Tyagi, A. K. Solanki and Manoj Wadhwa, "Analysis of Server Log by Web Usage Mining for Website Improvement", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010:In today's Internet world, log file analysis is becoming a necessary task for analyzing the customer's behavior in order to improve advertising and sales as well as for data-sets like environment, medical, banking system as it is important to analyze the log data to get required knowledge from it. Web mining is the process of discovering the knowledge from the web data. These data-sets are huge. In order to analyze such large data-sets we need parallel processing system and reliable data storage mechanism. Virtual database system is an effective solution for integrating the data but it becomes inefficient for large datasets. The Hadoop framework provides reliable data storage by Hadoop Distributed File System and MapReduce programming model which is a parallel processing system for large datasets. Hadoop distributed file system breaks up input data and sends fractions of the original data to several machines in hadoop cluster to

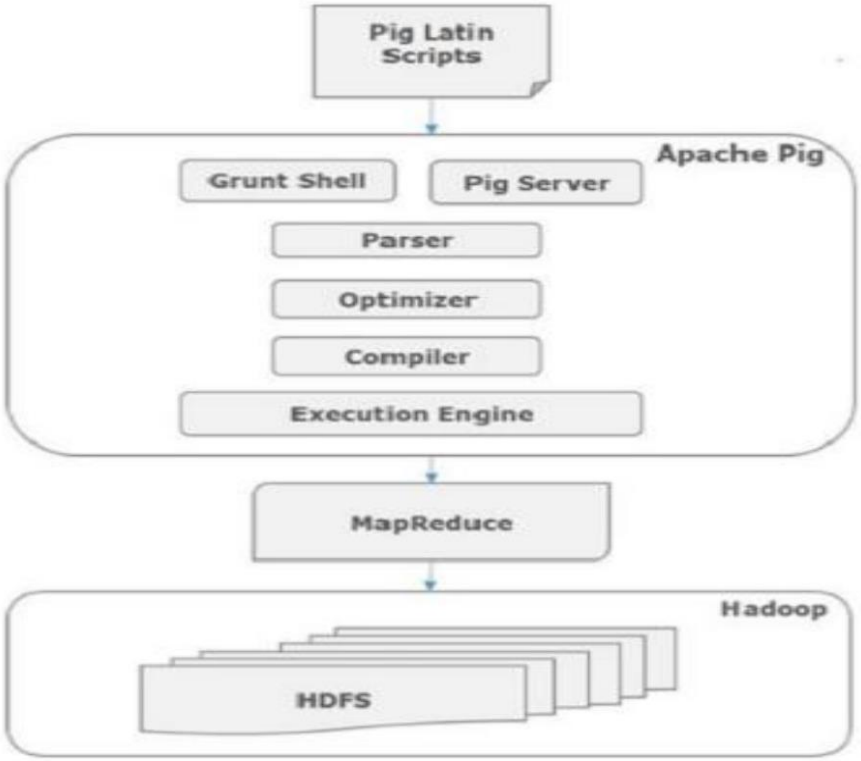
hold blocks of data. This mechanism helps to process log data in parallel using all the machines in the hadoop cluster and computes the result efficiently. The dominant approach provided by hadoop to "Store first query later", loads the data to the Hadoop Distributed File System and then executes queries written in Pig Latin. This approach reduces the response time as well as the load on to the end system. This paper proposes a log analysis system using Hadoop MapReduce which will provide accurate results in minimum response time [2].

5. PROPOSED ARCHITECHURE /FRAMEWORK

Pig Architecture consists of Pig Latin Interpreter and it will be executed on client Machine. It uses Pig Latin scripts and it converts the script into a series of MR jobs. Then It will execute MR jobs and saves the output result into HDFS. In between, it performs different operations such as Parse, Compile, Optimize and plan the Execution on data that comes into the system.



Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer's job easy.



Apache Pig Components

As shown in the figure, there are various components in the Apache Pig framework. Let us take a look at the major components.

Parser

Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG

(directed acyclic graph), which represents the Pig Latin statements and logical operators. In the DAG, the logical operators of the script are represented as the nodes and the data flows are represented as edges.

Optimizer

The logical plan (DAG) is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown.

Compiler

The compiler compiles the optimized logical plan into a series of MapReduce jobs.

Execution engine

Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

6.SOFTWARE AND HARDWARE REQUIREMENTS

Hardware requirements:

- Intel Core 2 Duo/Quad/hex/Octa or higher end 64 bit processor PC or Laptop (Minimum operating frequency of 2.5GHz)
- RAM > 4 GB
- Memory - 150GB (for Ubuntu hadoop and other components)

Operating system:

Linux (Ubuntu)

Software requirements:

- Hadoop
- Java
- Apache Pig
- Python3
- Perl
- Tableau

7.POSSIBLE INPUT

Based on each unique day we need to find the total hits. For example, on 24th of a particular month, there were X hits, on 27th of the month, there can be Y hits. The assumption has been made that logs are of a single month. To solve this problem, we have to use DateExtractor() available in Piggybank jar. This will take the timestamp as input and will give corresponding "day" against each timestamp. The input data is a CSV file.

Sample Attributes:

IP address, LogName, Date, User Id, TimeStamp,Request, PageLink

8.EXPECTED OUTPUT

- Pre-processed data with useful instances
- The most viewed page based on the number of people visited(IP address).
- No of hits per unique day.

9.REFERENCES

[1] Sayalee Narkhede and Tripti Baraskar, "hmr log analyzer: analyze web application logs over hadoop mapreduce", International Journal of UbiComp (IJU), Vol.4, No.3, July 2013

[2] Navin Kumar Tyagi, A. K. Solanki and Manoj Wadhwa, "Analysis of Server Log by Web Usage Mining for Website Improvement", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010

3. Milind Bhandare, Prof. Kuntal Barua, Vikas Nagare, Dynaneshwar Ekhande, Rahul Pawar, " Generic Log Analyzer Using Hadoop Mapreduce Framework", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 9, September 2013

4. L.K .Joshila Grace, 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 2011

5. Praveen Kumar, Dr Vijay Singh Rathore, "Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014

6. Anuja Pandit, Amruta Deshpande, Prajakta Karmarkar, "Log Mining Based on Hadoop's Map and Reduce Technique", International Journal on Computer Science and Engineering (IJCSE) Vol. 5 No. 04 Apr 2013