Map Reduced Based Log Analysis and Prediction Using Hadoop

Divyesh Bhoir, Amey Chavan, Aditya Patil
BE Students, Dept. of Information Technology
Bharati Vidyapeeth College of Engineering,
Navi Mumbai, India.

Prof. Vaibhav Pawar
Dept. of Information Technology
Bharati Vidyapeeth College of Engineering,
Navi Mumbai, India.

Abstract
Nowadays in internet world Logs are an significant part of any computing system. As logs grow and the number of log sources increases, a scalable system is necessary to efficiently process logs.so analyzing log file is becoming a essential task for analyzing the user's behavior to improve business as well as for datasets like social network, medical, banking system it is important to analyze the log data to get required knowledge from it.Log files are generated at a high velocity, these datasets has large volume.In order to analyze such huge datasets, we need parallel processing system and reliable data storage mechanism. So we provide them with an application that shows the log analysis with the help of graphs.

I. Introduction

Nowadays, everyone is using Internet for browsing something. Each and every field upload their applications on Internet. From home we can do shopping or can do work related to bank, we get weather information etc. And in such a competitive environment, service providers are curious to know about whether they are providing best service in the market, whether people are using their product, are they finding application unique and user friendly, or in the banking field they need to know about how many customers are looking forward to our bank scheme, they also need to know about problems occurred, how to solve them, how to make websites or web application interesting, which products people are not using and in that case how to improve advertising strategies to attract customer, what will be the future marketing plans.

Logs has variety of information, but as user using a particular application or website increases, then the data collected in a log file is huge that can be used useful to make a business strategy .

In similar way, they also need to know about problems that have been occurred, how to resolve them, how to make websites or web application interesting, which products people are not purchasing and in that case how to improve advertising strategies to attract customer, what will be the future marketing plans. To answer these entire questions, Logs come in all shapes, but as applications and infrastructures grow, the result is a massive amount of distributed data that's useful to mine[1].

One can use log analysis to know the count of errors or event occurred (such as login failures). One can also know the connections or transactions per second. One can also know the site visited by which user and when, by doing log analysis. This log analysis can also provide the no of unique user visits in addition to file access statistics.

As the rate of data increases over years, storage and analysis of the log data becomes difficult, this in turn increases the processing time and cost of processing. Various techniques and algorithms are used in distributed computing the problem remains still idle. To solve this problem Hadoop MapReduce is used, to process number of files in a parallel manner. The use of internet produce data in large quantity as users are more interested in performing their day to day activities through internet.

Users communication with an application or a website is analyzed through web server log files, a computer generates data in semi-structured format, so hadoop can be use to process the large amount of data in the log file. Mapreduce algorithm is used to process data in two phases map and reduce and the result is combined from various cluster and a report is generated.

Hadoop breaks up log files into numbers of blocks and these blocks are evenly distributed over cluster of thousands of nodes. Reliability and fault tolerance features implemented on account of replication of these blocks over the MapReduce paradigms are designed to compute large volumes of data in a parallel fashion, in which the workload is divided across a large number of machines or nodes. MapReduce works by breaking the

processing into map and reduce phase .Each phase has key value pairs as input and output, the types of which may be chosen by the programmer.[8]

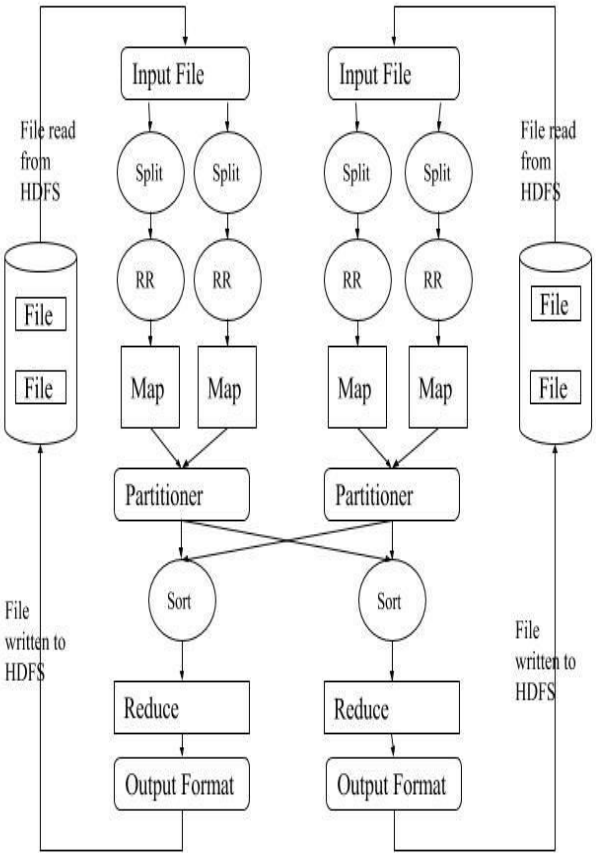

Fig.1 Map reduce data flow

II. Proposed System

The proposed system is used to analyze large datasets, with parallel processing system and reliable data storage mechanism.

The Hadoop framework provides reliable data storage by HDFS and MapReduce programming model which is a parallel processing system for large datasets.

HDFS splits input log data and sends part of the data to several machines in Hadoop cluster to hold blocks of data.

The result will be displayed in graphical form so that it will help the user can make their business strategies.

III. Architecture Diagram

The user enter the credentials i.e Username and password. This verification is done . After login the user browse and select the log file. These weblog file is analysed by using Mapreduce algorithm using Hadoop and sorted with respect to different parameter and the Mapreduce data is saved in HDFS and after sorting the graphical representation of ip-addresses and its statistics which will help user to make business strategies and grow their business.

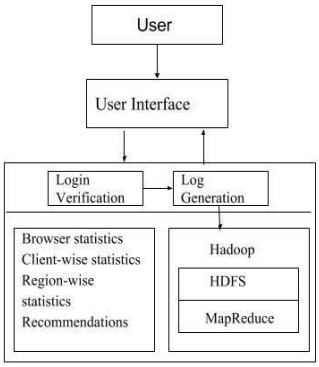

Fig.2 Architecture Diagram

V. Conclusion

Analyzing the logs dataset by using Apache Hadoop will solve all the issues caused by the existing systems. Performance will increase rapidly. From the report business owners can evaluate which module of the website need to be improved, which are the frequent visiting customers, from which geographical region website is getting maximum hits, etc., which will help in designing business strategies. Hadoop Map Reduce framework provides parallel distributed processing and reliable data storage for large volumes of log files. So that access time

References

[1]Sayalee Narkhede and Tripti Baraskar, "HMR LOG ANALYZER: ANALYZE WEB APPLICATION LOGS OVER HADOOP MAPREDUCE"

[2]Hemant Hingave, Prof. RasikaIngle "An approach for Map Reduce based Log analysis using Hadoop"

[3] 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.

[4] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27

[5]http://hadoop.apache.org

[6]https://en.wikipedia.org/wiki/Apache_Hadoop

[7]http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoophdfs/ HdfsDesign.html

[8]Roshini R and Manjunath Raikar, "Map Reduce based Analysis of Live Website Traffic Integrated with Improved Performance for Small Files using Hadoop''