ISSN 2395-1621

Enterprise Web Application Using Hadoop MapReduce System

[#1]Priyanka B.Mohite, [#2]Prof.A.R.Kulkarni

[1]priyankamohite77@gmail.com
[2]arkulkarni10@gmail.com

[#1]Department of Computer Science and Engineering
[#2]Prof. Department of Computer Science and Engineering

Walchand Institute of Technology, Sholapur.

ABSTRACT

Now-a-days there is boom of marketing online. Everyone is going online as it became a need in today's internet world. The web log file analysis is an important as well as a necessary task for analyzing the behavior of customers. It helps website owners to improve the advertising and business sales. One of the processes for discovering the knowledge from the web data is web mining. Log files are generated very fast i.e. 1-10Mb/s and a single data center can generate very huge of terabytes of data in a day. These generated data sets are huge and need to be handled. We need parallel processing and reliable data storage to analyze such large data set. The Hadoop framework works in an environment that provides distributed storage by Hadoop Distributed File System i.e. HDFS and MapReduce programming model which is a parallel processing systems. Hadoop is implemented to manage from single server to thousands of machines, each offering computation and storage. We design and implement the enterprise web log analysis system .By this system we can analyze daily web log records and performance of statistical report of customer and every actions of users request. The aim of this work is to provide web log analysis to analyze the data results. To get more precise and accurate user interest recommendations which they want to see. Log file analysis tool will provide us reports showing hits for dataset, user's activity, in which part of data users are interested and where is the traffic sources.

ARTICLE INFO

Article History

Received: 8th May 2016

Received in revised form :

8th May 2016

Accepted: 11th May 2016

Published online :

18th May 2016

Keywords: **Hadoop, Cloud computing, Hadoop Distributed File system, MapReduce, Pig Latin language.**

I. INTRODUCTION

Today's world mostly depends on the internet, PC's and mobiles. Nearly all of data get generated from these devices. Each and every person wants to complete the task just by sitting at home. Every Service provider is trying to put his own applications, business strategies on the internet. So seating at home we can do easily shopping, we get wheatear information, banking related work and so more services. In such a competitive world Service providers eager to know the choice of customers, user's activity in which part of web application user is interested, what is best selling in the market, are customers satisfied by their services, is the application is user friendly, what is customer purchasing etc. As well as they also need to know about problems such as how to make application more interesting, how to give proper service, how to make web application more popular by improving the advertisement strategies and can decide future marketing plan's .Log files is important factor for all these. All the actions are taking placed in log files whenever someone accesses your web application. The log file is huge file containing the information for service provider, considering these log files can give tons of insight that help understand website traffic patterns, user activity, their interest etc[10][2]. The huge file gets converted into big data. So we need Hadoop to handle the data and for storage purpose cloud. Thus, through the log file analysis we can get the information about all the questions as log is the record of people interaction with web application.

II. BACKGROUND

Thousands of terabytes or petabytes of data are get generated by a data centre in a day. Per record rate log files get generated. It's a challenging process to store and analyse such log files having huge volume. Analysing log files is not

an easy process not only due to huge volume but also because of the unrelated structure of log files. Usual database like SQL DBMS solutions are not suitable for analysing such log files. The SQL DBMS and Hadoop MapReduce are compared by Andrew Pavlo and Erik Paulson in 2009 [3] it is suggested that Hadoop MapReduce performs better than the SQL DBMS. It is mentioned that traditional DBMS cannot handle a large dataset. So we need to have Big Data technologies like Hadoop framework [6].

Hadoop was created by Doug Cutting and Mike Cafarella administrated by Apache Software Foundation. Hadoop MapReduce is a software framework for simply writing applications which procedure vast amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner[15].

The MapReduce framework consists of a single master and one save TaskTracker cluster node . The master is answerable for developing the jobs component tasks on the slaves, examining them and re-executing the failed tasks. The slaves perform the tasks as expressed by the master.

The Tom White[14] described Hadoop Cluster stores datasets with the Hadoop Distributed File System(HDFS). Run distributed computations with MapReduce. Use Hadoop data and I/O building blocks for compression, data integrity, serialization and persistence. Hadoop Load data from relational databases into HDFS, using Sqoop Performs large-scale data processing with the pig query language. Hadoop splits log files into blocks and these blocks are evenly distributed over thousands of nodes in a Hadoop cluster. MapReduce improves performance for large log files by breaking job into number of tasks by parallel computation.

II. PROPOSED WORK

The aim of this work is to provide web log analysis to analyze the data results. To get more precise and accurate user interest recommendations which they want to see. Log file analysis tool will provide us reports showing hits for dataset, user's activity, in which part of data users are interested, traffic sources, etc. From these reports business communities can evaluate which parts of the website or data need to be improved, which are the potential customers, from which geographical region website is getting maximum hits, etc, which will help in designing future marketing plans. The Hadoop will help to manage the big data by parallelization of data sets.

III. SYSTEM ARCHITECTURE

Following system architecture shown in Figure1 consists of major components like Client, Admin and User Application, Server implementing Hadoop storage and MapReduce programming model.

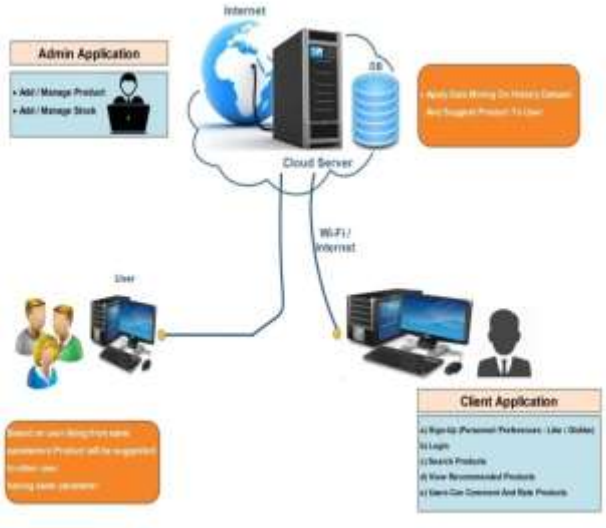

Fig 1: System Architecture

Modules of project

1. Client Application

Client can Sign up through the client application. After Login there are various options for user as per requirement. User can search products, Like/Dislike , can make comments on products, View recommended products. Client Application is connected to the server and User application through the web Services. The Client/Server communication is completed through SOAP/XML.

1.1SOAP

SOAP, originally defined as **Simple Object Access Protocol**, Is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks. It relies on Extensible Markup Language (XML) for its massage formats, and usually relies another application layer protocols, most notably Remote Protocol Cell (RPC) Hypertext Transfer Protocol (HTTP), for massage negotiation and transmission. SOAP can form the foundation layer of a web services protocol stack, providing a basic massaging framework upon which web services can be built. This XML based protocol consists of three parts, an envelope which defines what is in the massage and how to process it, a set of encoding rules for expressing instances of application-defined data types, and a convention for representing procedure cell and responses.

1.2 XML

Extensible Markup Language (**XML**) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. It is defined in the XML 1.0 Specification[19] produced by the W3C, and several other related specifications,[20] all gratis open standards.[21] The design goals of XML emphasize simplicity, generality, and usability over the Internet.[22] It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

2. Admin Application

Admin Application Add/Manage Product, all stock. The all managed data get sent to HDFS.

2.1 Hadoop

Using the solution provided by Google, **Doug Cutting** and his team developed an Open Source Project called **HADOOP**. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

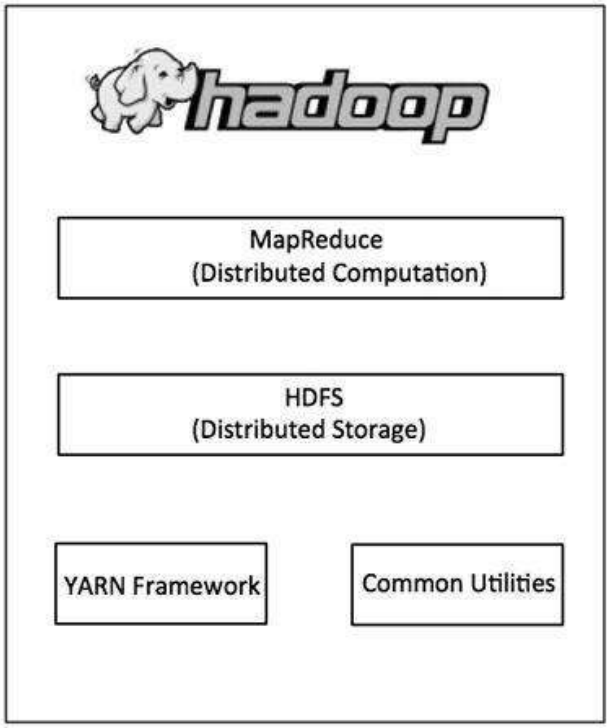

Fig 2: Hadoop Architecture

2.2 HDFS

A HDFS instance may consist of thousands of server machines, each storing part of the file system's data. Since we have huge number of components and that each component has non-trivial probability of failure means that there is always some component that is non-functional. Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

3. Server

The server is the important component in whole process. When the data is send to the HDFS it transfers the data to the server. The server receives the data, process data. After processing data apply analysis on dataset and suggests product. As the java programming is used for the developing of the application the GlassFish server is used for connectivity purpose. MapReduce technique are applied on the dataset for data mining. The algorithm used is Naïve Bayes. The result will get at last after data mining process.

3.1 GlassFish Server

GlassFish is a server that supports Java EE API such as JDBC, RMI, e-mail, JMS, web services, XML and defines how to coordinate them. Java EE also features some specifications unique to Java EE for components. GlassFish is based on source code released by Sun and Oracle Corporation's Top Link persistence system. It uses a derivative of Apache Tomcat as the servlet container for serving Web content, with an added component called Grizzly which uses Java New I/O (NIO) for scalability and speed.

3.2MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets with parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations. A MapReduce program is composed of a **Map()** procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a **Reduce()** method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance. The models are inspired by the map and reduce functions commonly used in functional programming although their purpose in the MapReduce framework is not the same as in their original forms.

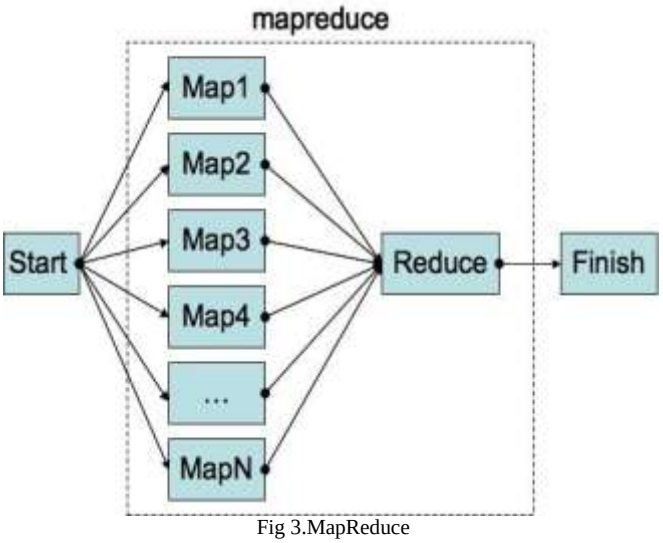

Fig 3.MapReduce

3.4 Naïve Bayes

Build the vocabulary as the list of all distinct words that appear in all the documents of the training set. Classification: Analysis of data by correlation i.e. for eg: Fruit may be considered to be an apple if it is red, round& about to 10 cm in diameter. Naïve Bayes considers each of the features to contribute independently to the probability i.e. fruit is an apple, regardless of any possible correlations between the color, roundness & diameter feature. In this project the algorithm can be used for recommendations and advertisement whether will the customer buy or like the products.

4. User Application

User Application is the application used by user's who don't have authenticated login id but want to view products, suggestions etc.

5. Cloud Server

Cloud computing is a phrase used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. In science, cloud computing is a synonym for distributed computing over a network, and means the ability to run a program or application on many connected computers at the same time. The phrase also more commonly refers to network based services, which appear to be provided by real server hardware, and are in fact served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up (or down) on the fly without affecting the end user - arguably, rather like a cloud.

IV. MATHEMATICAL MODEL

1. Set Theory :

Let G be the global set,
$G = \{U, S, A, A_t, P, A_d\}$

Where,
U= Set of Users.
$U= \{U_1, U_2, \ldots\ldots, U_n\}$
Where n is infinity.

S= Set of Servers.
$S= \{S_1\}$

A= Set of Algorithms
$A= \{a_1, a_2, \ldots. ak\}$ where k not equal to infinity.

A_t = Set of Attributes (Naive bayes)
$A_t = \{a_{t1}, a_{t2}, \ldots\ldots, a_{tk}\}$

P= Set of Products.
$P= \{ P_1, P_2, \ldots\ldots, P_n\}$

A_d= Set of Admin.
$A_d = \{a_{d1}, a_{d2}, \ldots.., a_{dk}\}$

2. Morphisim:

- **Admin:**

<DB> ← Add Products (Product Details);

<DB> ← Manage Products (Product ID);

- Client:

Server <DB> ← Registration (User Details);

Yes/No ← Login (userId, Password);

void ← Buy Products(Product Details);

<details> ← View Product(details);

<DB> ← giveRatings(Rate);

<DB> ← Provide liks (Product);

- **Server:**

(Naive Bayes model) ← Train Naive Bayes(Normalize dataset);

<Products> ←Naive Bayes Suggestions (Current User Details);

<Vector> ← Normalize dataset (User History dataset);

V. CONCLUSION

The proposed work analyzes the data used by users. In similar way, web log analysis is to know about problems that have occurred, how to resolve them, how to make 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, log files are helpful. Log files contain list of actions that have been occurred whenever someone accesses to website or application. The project is to analysis the log files to improve the businesses as well as to generate statistical reports.

REFRENCES

[1]Chen-Hau Wang,Ching-Tsorng Tsai, Chia-chen Fan, Shyan-Ming Yuan (2014) "Hadoop based Web log analysis system", IEEE International Conference.

[2]Yang, Q. and Zhang, H., (2003) "Web-Log Mining for predictive Caching", IEEE Trans.Knowledge and Data Eng., 15(4), pp. 1050-1053.

[3]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", ACM SIGMOD'09.

[4]Liu Zhijing, Wang Bin, (2003) "Web mining research", International conference on computational intelligence and multimedia applications, pp. 84-89.

[5] S.Sathya Prof. M.Victor Jose, (2011) "Application of Hadoop MapReduce Technique to Virtual Database System Design", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), pp. 892-896.

[6] Mr. Yogesh Pingle, Vaibhav Kohli, Shruti Kamat, Nimesh Poladia, (2012)"Big Data Processing using Apache Hadoop in Cloud System", National Conference on Emerging Trends in Engineering & Technology.

[7] C.Olston, B.Reed, U.Srivastava, R.Kumar, and A.Tomkins, (2008) "Pig latin: a not-so-foreign language for

data processing", ACM SIGMOD International conference on Management of data, pp. 1099– 1110.

[8] Sayalee Narkhede, Tripti Baraskar, (2013) "HMR log analyzer : Analyze web application log over Hadoop mapreduce", International Journal of ubi comp.

[9] Jeffrey Dean and Sanjay Ghemawat., (2004) "MapReduce: Simplified Data Processing on Large Clusters", Google Research Publication.

[10] Pig Latin://wiki.apache.org/pig/PigLatin

[11] "3 approaches to big data analysis with Apache Hadoop"by DaveJaffe.http://www.dell.com/learn/us/en /19/ power/ ps1q14-20140158-jaffe

[12] "Why Big Data is a must in E-Commerce", Guest post by Jerry Jao, CEO of Retention Science. http://www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce.

[13] Apache-Hadoop,http://Hadoop.apache.org

[14]Tom White, (2009) "Hadoop: The Definitive Guide. O'Reilly", Scbastopol, California.

[15] Guru Prasad M S, Nagesh HR, Deepthi M (2014)"Improving the performancfe of processing for small files in Hadoop: A case study of weather Data Analytics".IJCSIT.

[16] D.sridevi, Dr. A. Pandurangan, Dr. S.Gunasekaran(2014) "Survey on the latest trends in web mining". International Journal of research in Advent Technology. Vol.2,No.3 E-ISSN:2321-9637.

[17]Hadoop Distributed File System://Hadoop.apache.org/ hdfs/

[18] Cloud Computing://www.wikinvest.com/concept/clou computing.

[19]"XML 1.0 Specification" (http:/ / www. w3. org/ TR/ REC-xml). W3.org.

[20] "XML and Semantic Web W3C Standards Timeline" (http:/ / www. dblab. ntua. gr/ ~bikakis/ XML and Semantic Web W3C Standards.

[21] "W3C DOCUMENT LICENSE" (http:/ / www. w3. org/ Consortium/ Legal/ 2002/ copyright-documents-20021231).

[22] "XML 1.0 Origin and Goals" (http:/ / www. w3. org/ TR/ REC-xml/ #sec-origin-goals).

[23] http:/ / glassfish. java. net/

