ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

# Analysis of Web Log Files Integrating Hadoop MapReduce with Naive Bayes Algorithm

Priyanka B.Mohite, Prof. A.R. Kulkarni

M.E Student, Dept. of CSE, Walchand Institute of Technology, Sholapur, Maharashtra, India

Assistant Professor, Dept. of CSE, Walchand Institute of Technology, Sholapur, Maharashtra, India

**ABSTRACT**: An important task for E-Commerce companies is to analysing web log files to predict their customer behaviour and to improve their business. A new technique for analysing is presented in this paper for analysing the data known as weblog analysis using Hadoop MapReduce system. Hadoop is an open source framework used to store and process data which is very huge in volume. The speed of generating log files are very fast at the rate of 1-10 Mb/s per machine, a single data centre can generate tens of tera to peta bytes of log data in a day. These datasets are huge and in order to analyse such large datasets we need parallel processing and reliable data storage system. Hadoop framework provides both these requirements. MapReduce is a java based framework for parallel computation using key/value pair. Thus, the Hadoop MapReduce system helps to analyse the data which will provide the information of the potential users such as login time, credit worthiness, and many more in minimum response time. 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.

**KEYWORDS**: Hadoop; MapReduce; web log file analysis; Big Data; Naive-Bayes

## I. INTRODUCTION

Today's world mostly depends on the E-Commerce application such as online shopping, share markets, online banking, weather forecasting, Railway reservations etc, which creates a huge amount of data in tera bytes, these data are produced by different devices and modules of applications.

Analysis is necessary for decision making to provide better business strategies for service providers. Recommendation and reviews provided by the application helps the users to get the services easily from their location Viz: online shopping, net banking, wheatear forecasting, online reservation, Google maps etc. The competition among service providers has increased in terms of providing fast and best services , additionally they also analyse the area of interest of users, products maximum purchased by the users and log analysis, the queries posted by users, reviews given by users for different products. This analysis helps to provide better recommendation to other users during their purchase and also help in deciding future marketing plans.

Log files are important factor, to handle these large amounts of data we use Hadoop and MapReduce for parallelizing the computation required for analysis. Thus through the log file analysis, service provider can get all the information of users interacting with the application.

## II. RELATED WORK

In This section reviews some of the previous research works which are related to log file usage and the approach used in analysing them.

P.Saravana Kumar/ R.Iswarya and R.Vidhya, "Predictive Analysis of Users Behaviour in Web Browsing and Pattern Discovery Networks" There research focus to collect web access log files which are recorded in the server. On the basis of this, they attempt to predict the next set of web pages that a user may visit based on the knowledge of previously visited pages. The first is pre-processing state in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

contents of pages provide evidence that pages are related [11]. Suneetha and K.R., "Identifying User Behaviour by Analysing Web Server Access Log File" describes a method to improve the performance of application with the help of system administrator of web based application. The paper is concerned with analysis of the web log data of the NASA website. Using the approach of in-depth analysis important information can be found i.e. topmost errors, potential visitors of website through analysing the website's log file. In addition to this, to capture most active days and least active days of server the log files content was used. By above information, particular days for shutting down server can be resolute [12]. Lawrence McClendon and Natarajan Meghanathan., in their paper title, "Using Machine Learning Algorithms to Analyse Crime Data" they explained Data mining and machine learning is important part of crime detection and prevention. In this research the used WEKA, an open source data mining software, to relative study between violent crimes patterns from the society and crime dataset [13]. Ahmad Tasnim Siddiqui., in their paper title "Web Mining Techniques in E-Commerce Applications" explained today web is the best medium of communication in modern business. Now day's online purchase has been increased as compared to window shopping as it provides millions of ranges. As, companies are able to attract most of the customers because ecommerce is not just buying and selling over internet but it also act as to get advantage on big giants of market. For this purpose data mining sometimes called as knowledge discovery is used. As vast information has been provided on internet, it helps to improve e-commerce applications After that they explained the proposed architecture which contains mainly four components business data, data obtained from consumer's interaction, data warehouse and data analysis. After finishing the task by data analysis module it'll produce report which can be utilized by the consumers as well as the e-commerce application owners [14].

From literature survey we studied the above techniques there are some time consuming, they will be applicable on limited data, static size sets. In this dissertation, the drawback is overcome by using Hadoop MapReduce and Naive Bayes algorithm. Hadoop has been proposed as a solution to make the analysis faster and it is also able to analyse larger dataset. Hadoop framework enables to successfully minimize the analysis process response time as well as analysing larger dataset (TBs). Naive Bayes algorithm is easy to build.

## III. ARCHITECTURE OF HADOOP MAPREDUCE

### A. *HADOOP:*

Hadoop [10] uses MapReduce algorithm for parallelization of the data. Hadoop applications are developed for the complete analysis of big data. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage [10]. Hadoop has two main layers:

1. MapReduce
2. HDFS

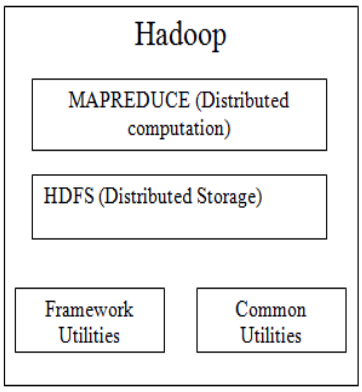


Fig.1. Hadoop Architecture

The **Hadoop Distributed File System (HDFS)** is fully based on the Google File System (GFS) and it is designed to run on large clusters that provide distributed file system in reliable and fault tolerant manner [10]. Master/Slave

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

architecture is used by HDFS where Master have single NameNode that manages file system data and have one or more slave DataNodes that stores actual data.

The code is run on all clusters of Hadoop. The following tasks are performed by Hadoop:

1. Initially data get divided in directories and files.
2. Files are divided into blocks of size 64M or 128M (preferably 128M). These files are distributed in all clusters for further process.
3. Management is done by the HDFS, as it is presented on top of local file system
4. Replication of block is done to handle hardware failure.
5. Execution of code is checked.
6. Map and reduce will perform sorting.
7. The sorted data get send to certain computers.

### B. *MAPREDUCE:*

The MapReduce is distributed computing programming model [4]. Map and Reduce are two important tasks in MapReduce algorithm. In this algorithm Map receives collection of data and converts it into another collection of data, where each component is broken into tuples. Map sends input to the Reduce from its output and joins those input tuples into smaller collection of tuples. The Reduce task is always executed after Map job as the name implies MapReduce [4]. The Fig. 2 illustrates the work of MapReduce.

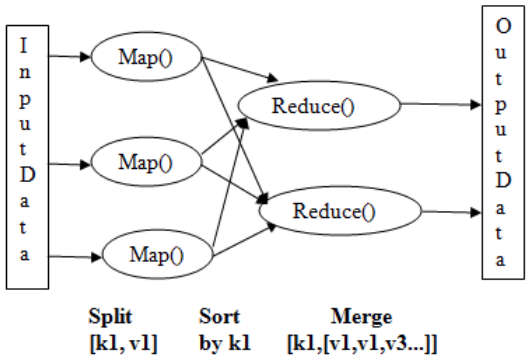


Fig.2. MapReduce

The main benefit of MapReduce is that it is simple to extent data to handle out on many nodes. The primitives processing data are called mappers and reducers in MapReduce model. Sometimes is not easy to decompose data processing into mappers and reducers. Once, the application is written in MapReduce form, it can be run over hundreds, thousands or even ten thousands machines in cluster. Many programmers use MapReduce model because of this simple scalability. MapReduce paradigm/concept is based on where the data is present. It executes three stages i.e. Map stage, Shuffle stage and Reduce stage. The MapReduce algorithm consists of following steps.

Step 1: Map Stage:

The input data is processed by mapper's job. The input data is in type of files or directories stored in HDFS. Mapper function receives line by line input. After processing several chunks of data gets created by mapper.

Step 2: Reduce Stage:

This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The data came from mapper is processed by Reducer's job. HDFS stores new set of output after all processing.

Hadoop sends map and reduce task to appropriate servers in cluster during MapReduce jobs. Hadoop framework supervises all details of data-passing such as taking tasks, verifying completion of task and replicating data around cluster between nodes. The network traffic get reduce because most of task is computed on nodes with data on local disks. The cluster collets and reduces data to form correct result at the end after completion of task and sends it to Hadoop server.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

### C. *HADOOP MASTER/SLAVE ARCHITECTURE:*

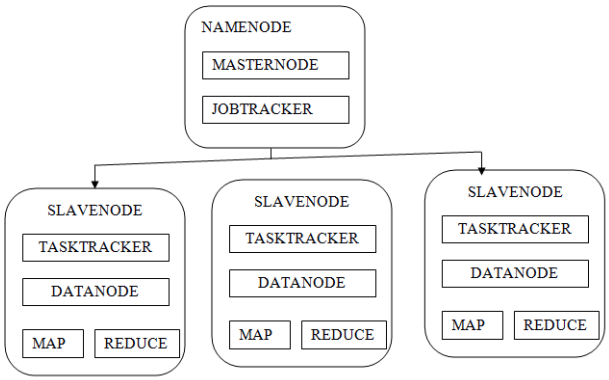


Fig.3. Hadoop Master/Slave Architecture

Hadoop have Master/Slave Architecture.
Master: Namenode, JobTracker
Slave: {DataNode, TaskTracker}..... {DataNode, TaskTracker}

HDFS and MapReduce are components of Hadoop cluster and both of them have Master/Slave architecture.

1. HDFS:
   Master: NameNode
   Slaves: {DataNode}..... {DataNode}

- File system operations such as opening, closing and renaming files and directories are managed by Master (NameNode). It also manages mapping of blocks to DataNode with regulating access to files by clients.
- Slaves (DataNodes) read and write request from file system's client and also perform deletion, creation and replication of block as per master's (NameNode) instruction.

2. MapReduce:
   Master: JobTracker
   Slaves: {TaskTracker}..... {TaskTracker}

- The interconnection between users and MapReduce framework is done by Master (JobTracker). When map/reduce job is submitted, JobTracker executes jobs on first-come/first-served basis and assigns map/reduce tasks to Task trackers.
- Slaves (TaskTracker) executes task as per masters instructions and also handle data motion between maps and reduce phases.

## IV. SYSTEM ARCHITECTURE

According to author who has used Naive Bayes method for log analysis has less time efficiency as compared to naive bayes with hadoop map reduce [9]. This data is big data requires multinode to handle the data and process it parallel. Our work focuses on integrating hadoop system with naive bayes algorithm.

Following system architecture shown in Fig.4 consists of major components like Client, Product Admin, Guest user application, Hadoop Framework and MapReduce programming model used for distribution purpose.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

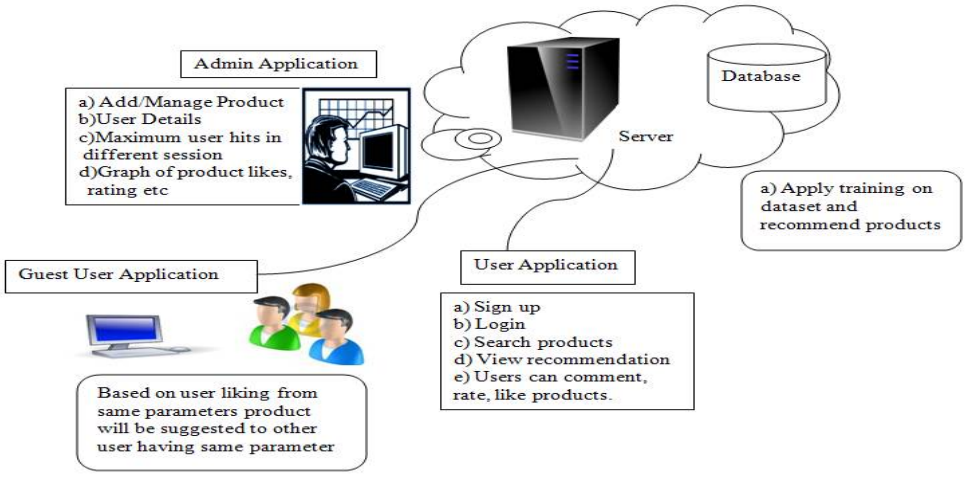


Fig.4. System Architecture

The Modules of system architecture shown in Fig.4 are explained as follow:

*A.* ***PRODUCT ADMIN APPLICATION*:**

Product Admin Application will handle data such as add/manage products, show user profiles, purchased history till date, area of interest of user, maximum user hits in different session, the graph showing product reviews, likes, product sold.

*B.* ***CLIENT APPLICATION:***

Client can Sign up through the client application. After Login there are various options. User can search products, Like/Dislike, make comments on products, rate the products, View recommended products. Client Application is connected to the server and User application through the web Services.

*C.* ***GUEST USER APPLICATION:***

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

## V. EXPERIMENTAL SETUPS AND RESULTS

The system was developed using Hadoop 1.2.1 running on stand-alone mode on multiple machine running Ubuntu 15.04 operating System. The algorithm Mapreduce based Naive Bayes algorithm is implemented using JAVA with Jdk version 1.8.

To evaluate the performance of the system observational studies were conducted on real time dataset. In this experiment we compare our Naive Bayes Algorithm without MapReduce and with MapReduce in the primary factor appropriate execution time.

**Table 1:** Comparison of required time in milliseconds without and with MapReduce.

| **Users** | **Without Hadoop MapReduce(**time in ms**)** | **With Hadoop MapReduce(** time in ms**)** |
|---|---|---|
| Aa | 15607 | 7685 |
| Bb | 9139 | 8500 |
| Kk | 11957 | 11851 |
| Ee | 14925 | 7426 |
| Cc | 9633 | 8932 |

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

The above showed Table1 is comparison of the execution time with and without hadoop mapreduce. Here, with mapreduce the execution time gets reduced while without mapreduce it gets increased because of the parallel processing and distributed system. As the parallel processing distributes the data between nodes, so here the time difference is seen. The Fig.5 shows comparison graph of required time in milliseconds with and without Mapreduce of Table1. Without Mapreduce time extends while with hadoop MapReduce time needed is minimum.

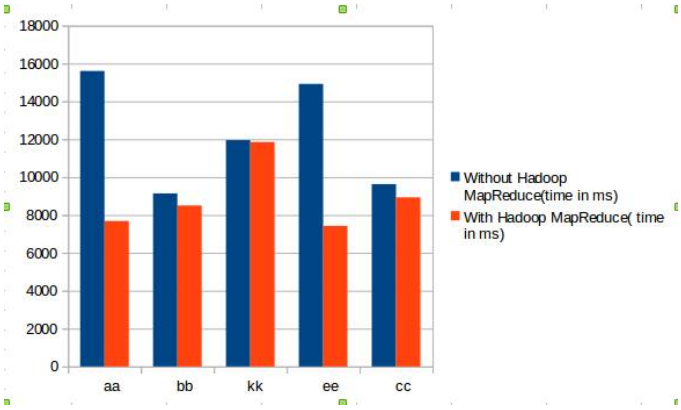


Fig.5. Comparison graph of time in milliseconds with and Without Mapreduce

For finding the relevant products that are retrieved we use Recall and Precision formulas. The Table2 shows Recall, Precision and F-Measure for retrieved relevant products. Here, we have collected 5 set of users and calculated each sets ratio for relevant products. Each set consists of 10 users. The number of available products, purchased products and retrieved products calculation is done to prove the recall, precision and f-measure.

Recall is the ratio of the number of relevant products retrieved to the total number of relevant products in the database. Precision means fraction of retrieved products that are relevant. F-Measure means ratio of recall and precision.

**Table 2:** Recall, Precision and F-Measure for retrieved relevant Products.

| **Users** | 1-10 Users | 11-20 Users | 21-30 Users | 31-40 Users | 41-50 Users |
|---|---|---|---|---|---|
| **Recall** | 0.0395480226 | 0.0564971751 | 0.1073446328 | 0.0677966102 | 0.056497175 |
| **Precision** | 0.28 | 0.4166666667 | 0.4634146341 | 0.5454545455 | 0.4545454545 |
| **F-Measure** | 0.06930693 | 0.0995 | 0.17431 | 0.1206 | 0.1005 |

Table 3 shows comparison of traditional naive bayes algorithm with map reduce model. We implemented this concept in our project. Traditional naive bayes have some disadvantages which we have overcome by using Mapreduce model. As traditional naive bayes have low performance on large datasets while by using hadoop Mapreduce it process on large datasets. Here comparison shows the difference of traditional naive bayes algorithm with Mapreduce model.

**Table 3:** Comparison of Traditional Naive bayes Algorithm with MapReduce Model.

| Classification | Traditional | Based On MapReduce |
|---|---|---|
| **Naive Bayes** | 1. Low performance in large datasets.<br>2. It has strong feature independence assumptions. | 1. Improves the performance.<br>2. Reduce the training time.<br>3. Able to process large database. |

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 11, November 2016**

## VI. CONCLUSION AND FUTURE WORK

Various researchers have worked on this big data using different methods and naive bayes. Mostly, all this work has been done on small system or centralized system and time efficiency is less. To improve time efficiency we have used naive bayes algorithm with integration of hadoop system.

The Web Applications are becoming most popular now a day. Web applications creates big amount of data. So, analysing of these data is very important. To analyse these data here we used the Web Log Analysis using Hadoop MapReduce System. Naive Bayes algorithm is implemented based on MapReduce so it increased performance and reduced training time. The Log files will generate Statistical reports and will help to improve business strategies. As our system uses multiple nodes we can extend data in range of terabytes or more. As a future work to this idea here we can add more features to the application to improve business strategies by a parallel approach using the features provided by Hadoop..

## REFERENCES

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. Priyanka B.Mohite, Prof.A.R.Kulkarni, (2016) "Enterprise Web Application Using Hadoop MapReduce System"IERJ Enterprise Web Application Using Hadoop MapReduce System" vol2 Issue 3, 1145-1149.
5. Sara Landset,Taghi M. Khoshgoftar, Aaron N. Richter and Tawfiq hasanin.,(2015) "A Survey of open source tools for machine learning with big data and Hadoop ecosystems", Journal of big data.
6. Yogang Dai and haosheng sun, (2014) "The naive Bayes text Classification algorithm based on rough set in cloud platform".
7. S Saravanan, B Uma Maheswari, (2014) "Anlaysis large Web Log Files in a Hadoop Distributed Cluster Environment" IJCTA vol5 (5), 1677-1681.
8. Katheen Ericson and shrideep pallickara, "On the performance of high Dimensional data clustering and classification algorithms."
9. Jeffrey Dean and Sanjay Ghemawat., (2004) "MapReduce: Simplified Data Processing on Large Clusters", Google Research Publication.
10. Apache-Hadoop, http://Hadoop.apache.org
11. P.Saravana Kumar/ R.Iswarya, R.Vidhya. (2014) "Predictive Analysis of Users Behaviour in Web Browsing and Pattern Discovery Networks", IJLTET.
12. Suneetha, K.R., (2009). Identifying User Behaviour by Analysing Web Server Access Log File. 327–332.
13. Lawrence McClendon and Natarajan Meghanathan., (2015) "Using Machine Learning Algorithms to Analyse Crime Data" (MLAIJ) Vol.2, No.1.
14. Ahmad Tasnim Siddiqui., "Web Mining Techniques in E-Commerce Applications".
15. Songtao Zheng., (2014) "NAÏVE BAYES CLASSIFIER: A MAPREDUCE APPROACH"