Business intelligence By Log analysis in Hadoop using MapReduce

[1]Dipak R. Kawarkhe, [2]Nikhil B. Malwal, [3]Sagar A. Shirsath, [4]Abhishek A. Talekar
[5]Prof. Ratan Deokar

[1,2,3,4] Student of Information Technology
[5] Assistant Prof. of Information Technology MET BKC IOE Nashik,India
Email : dipakp1992@gmail.com[1], nikhilmalwalom@gmail.com[2], shirsathsagar889@gmail.com[3], talekarabhishek41@gmail.com[4]

Abstract—
A log file is a file that records the events that occur generally in the operating system or either event generated when software runs. Log is the main source of system operation status user behavior system actions etc. Log analysis system not only work on the massive data but also the data processing ability, the adaptation to a variety of scenarios under the efficiency, scalability and performance of quality data, which never be achieved from available standalone analysis tools or single computing framework. Now a day' s data is increasing day by day and there are huge numbers of log files are generated at the data centers and recommended companies. As this log file increases a scalable and advance system is necessary to efficiently handling the logs. Log files analysis becomes an important aspect for analyzing customer' s behavior regards to improve the sales of particular products. Hence we propose a log analyzer with the combination of Hadoop and Map-Reduce paradigm. The joint of Hadoop and MapReduce programming tools makes it possible to provide batch analysis in minimum response time and in memory computing capacity in order to process log in a high available, efficient and stable way.

Business intelligence is the variance in the software application use to analyze large scale raw data. It is nothing but the set of process architecture, different application strategies, and technological architectures. Business Intelligence is the way by which business trends can use this Log analysis structure and enhance their way to increase productivity and growth in revenue. Company can guess the user requirements by using their comments, reviews, opinions and many more by which it' s easy to identify and fulfill end users requirements. In our systems we are going to provide all these necessary things that can be identified using Log analysis which is used to improve productivity and these will totally turn into growth in revenue generation of any organization.

Keywords - Apache Hadoop, MapReduce, Logfiles, Parallel processing, Stemmer, Stopword.

I. INTRODUCTION

The need of today' s world is everything that going to be browsing something on the internet. Each field is having their own way by putting their applications on the internet. Through internet we do shopping recommending products and many more related works. This is everything about best services in the markets whether people purchasing the product and them always want the quality services of the product. They always try to know the problem occurred by giving opinions and reviews.

In a day, thousands of petabytes or terabytes of log files are generated by a data center. The data is available in extremely large amount so handling such a large amount of data become very crucial task. The data is not properly defined; it may be available in the structure as well as unstructured type of a format. It is very challenging to store and analyze this large volume of log files at a lower rate. The problem of log files analysis is very complicated because of not only its volume but also its changeable structure. The log data generated may be structured or unstructured. Log is the main source of system operation status, user behavior, system action etc. Log analysis system needs not only the massive and stable data processing ability but also the variety of scenarios under the requirements of efficiency and performance, which can' t be achieved from available standalone analysis tools or even single computing framework.

As the growth of data increases over the years, so that storage and analysis become incredible. Through various techniques and algorithms get available but the problems remains idle. To overcome the issues Hadoop and MapReduce functionalities are used, to process large files through parallel processing. Hadoop MapReduce tunes up the task faster and load data faster than DBMS. Hadoop MapReduce is applicable and gets also use in various areas of Big Data analysis. As log files is one of the type of Big Data which grows fastly and in increasing order so Hadoop is the best suitable platform for sorting log files and parallel implementation of MapReduce program for analyzing them. Hadoop is an open source software framework for distributed storage and processing of very large dataset on computer clusters build from commodity hardware. Hadoop has greater capability to handle such a large scale applications to work with thousands of nodes and terabytes or petabytes of data.

II. Literature Survey

A: Related Work

The literature survey state that the concept of lo analysis. Here, we found that user reviews are must to know that product is how much likely get used in the market. Any customer who wants to purchase any product the reviews and comments generated by the users so it will help to the customers to operate and popularate of that particular product.

Once the product get release in the market, the people firstly checks the specification and other related thing of the product. If the specification meets user' s requirements then customers are ready to buy that product. The existing customer leaves their comments and reviews about that particular product. Which is useful for new user? In log analysis the logs are in the form of bad, good, positive, negative, etc. Feedbacks of existing users by which the popularity of the product get known to new users.

By this the organization gate known that which product is much trending in the market by that they can increase productivity of that particular product and can generate high revenues. Also from negative feedback from user they can easily know the bugs and faults of that product. So it' s helpful to overcome the bugs easily. In this way the log analysis helps to know this much of related information.

B: Existing System:

Our existing system state that, until now the product is rated on its cost specification features. In existing system there is no provision of application further proceed to use of log analysis in existing system. The comments of users are not analyzed. When the term log analysis is specified then the greater extent of large databases get used where the lacks of comments and reviews of users are given but existing system not works on comments or reviews of existing users.

The existing system goes through disadvantage:

1. Not a real time application.
2. Comments and reviews are not analyzed.
3. Exact popularity and productivity is not analyzed.

Till now, we defined the trend in any particular product by using only their news, promotions, advertisement, newspaper etc. But now we are using log analysis on hadoop by which the comments and reviews by users who are existing users of the particular product. They can share their feedback which is very useful for next buyers.

The existing sytem only finds the logs by MapReduce system which will used to do only certain operations like the mostly viewed logs by users and the total visits on the basis of weeks, days, months etc. This only tends to only have a overview of visitors and to have a approximate overview.

The popularity and productivity is important aspect in every type of product. The existing system is only assign with general activities only. But it does not satisfy the user' s expectations.

III. Proposed System

The problem of emerging logs from the big data sets of data get overcome using MapReduce function in Hadoop, is the ultimate solution for business which are depend on par diagrams of graphical log based analysis. The working of proposed system is as following:

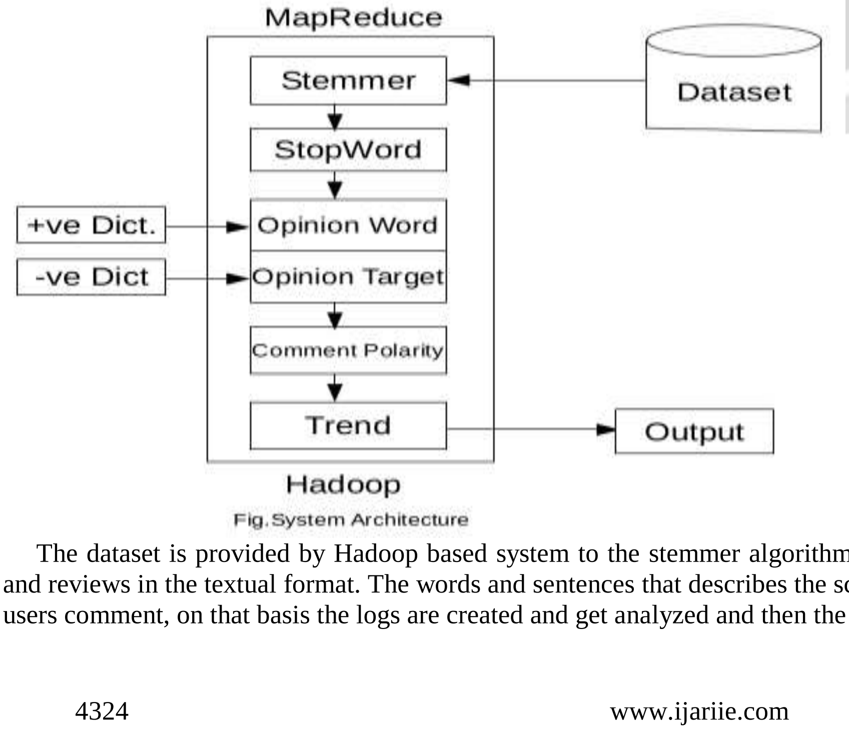

Fig.System Architecture

The dataset is provided by Hadoop based system to the stemmer algorithm. The dataset contains of existing users comments and reviews in the textual format. The words and sentences that describes the scenario of that particular product is described in the users comment, on that basis the logs are created and get analyzed and then the judging of that particular product is done.

The Stemmer algorithm states that A consonant in a word is a letter other than A, E, I, O, U and other than Y preceded by a consonant. So on consonant are T and Y and in SYZYGY they are S, Z, and G. If a letter is not a consonant it is a vowel.

We are also using the Stopword algorithm. In computing, stopword are words which are filtered out before or after processing of natural language data (text) Though stop words usually refer to the most common words in a language,there is no single universal list of stiop list, Some tools specifically avoid removing these stopwords to support phrases search.Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are the some of the most common. Short function words such as the, is, at, which, and so on. In this case stopwords can cause problems when searching for phrases include them, particularly in name such as " The whom" , " Take That" .

Opinion Word and Opinion Target are the words from the positive and negative so we can judge the comment on the basis of these words. In this way finally we got the trend for the Business Intelligence and can specify the trending product to the organization and may more things.

Advantages:

1. Real time application.
2. Beneficial to increase revenue.
3. Productivity will increase through log analysis.
4. Easy to identify trending product according to customer reviews

IV. ALGORITHM USED

1. STEMMER ALGORITHM:

Stemming is the process of reducing inflected words to their word stem generally a word written format. The stem needs to be identical to the morphological root of the word. A more complex approach to the problem of determining a stem of word is lemmatization.

A more complex approach to the problem of determining a stem of a word is lemmatization. This process involves first determine the part of speech of a word ,and applying different normalization rules for each part of speech .the stemming rule change depending on a words part of speech .

Porter stemmer steps:

1. Gets remove of plurals and – ed or – ing suffixes.
2. Change y to i when there is another vowel in the stem.
3. Mapping of double suffixes to single ones – ization.
4. Deals with suffixes, -full, -ness, etc.
5. Takes off – ant, -ence,etc.
6. Removes a final e.

2. STOPWORD Algorithm.

In computing, stop words which are filtered out before or after processing of natural language deta. Though stop words usually refer to the most common word in a language, there is no single universal list of stop words use by all natural language processing tools and indeed not all tools even use such a list.some tools specifically avoid removing these stop words to support phrase search.

Stopword steps :
Step1: Load data stractucture having list of special symbol like fullstop(.),comma(,).
Step2: Read comment on normalize word.
Step3: Remove special symbol and replace it by space

V. CONCLUSION

Our proposed system makes user queries efficiently using Intelligent MapReduce technique. I will improve thre performance of log maintenance, indexing, ranking and retrieving result through different types of algorithms and mainly the two dictionaries. This will provide the faster result than other conventional datasets. With the increase in the capability of the distributed system, It proposes that the effective information retrival techniques to improve the performance of short-term information retrieval in a system.

VI. References:

[1] Savitha K, Vijaya MS, “Mining of Web Server Logs in a Distributed Cluster using Big Data Technologies,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5, 2014.

[2] Sayalee Narkhede and Tripti Baraskar, “HMR Log Analyzer: Analyze Web Application Logs Over Hadoop MapReduce,” International Journal of UbiComp (IJU), vol.4, 2013.

[3] Milind Bhandare, Vikas Nagare, “Generic Log Analyzer Using Hadoop Mapreduce Framework,” International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, 2013.

[4] Kanchan Sharadchandra Rahate, “A Novel Technique for Parallelization of Genetic Algorithm using Hadoop,” International Journal of Engineering Trends and Technology (IJETT), vol.4, issue 8, 2013.

[5] T. K. Das., “BIG Data Analytics: A Framework for Unstructured Data Analysis,” International Journal of Engineering and Technology (IJET), vol 5, No 1, 2013.

[6] Wang, Peng, Wu, Bin, “Log analysis in cloud computing environment with Hadoop and Spark,” 5th IEEE International Conference on Broadband Network & Multimedia Technology (IC-BNMT), pp. 273 –276, 2013.