Generating Analytics from Web Log

Vempaty Prashanthi, Srinivas Kanakala

Abstract: ***Recent engineering trends integrate clever expertise in every factors of our lifes. Today's technologies generate terra bytes of recorded messages every day to record their data. It is difficult to research on such recorded messages and represent usable records such as patterns to directors, so as to manipulate and reveal those technology. Patterns minimally characterize huge corporations of recorded messages and allow the manager to do future analysis of data, along with variance detection and event forecast. Even though patterns exist typically in automatic recorded messages, spotting them into large collection of recorded messages from different resources lacking any past information is a widespread responsibility. We aim a bigdata using hadoop so as to extract very pleasant styles for the set of recorded messages which are given. Our approach is high-speed, memory efficient, exact, and expandable. Hadoop is implemented in map reduce framework intended for disbursed platforms to procedure hundreds of thousands of recorded messages in a moment. Hadoop is a strong technology which is used for different recorded messages produced in a wide style of systems. Present technique uses algorithmic procedures to limit the additional over- head based totally on the truth that recorded messages are continually routinely produced. We examine the accuracy of Log Mine of huge units of recorded messages produced in commercial appliances. It has efficiently created styles which might be as exact as the styles produced by genuine and un expandable procedures.***

Keywords **: *BigData, HDD'S. Weblog, MapReduce, Hadoop.***

I. INTRODUCTION

At present analyzing data is one of the major challenges in various heterogeneous disciplines even though specialization exists in each of their respective. Data analytics is an effective technique for analyzing the data for different kinds of business in the internet. Big data is is used to analyze the data and this how all business systems in the internet are along the way of success. As the features and technologies in the industries expand the data pertaining to the industries also increase. Therefore difficulty in handling
such huge amount of data also increases in any fields of business system.

The main aim of big data is to analyze the data and return meaning full information for making decisions. The gathered data is stored, processed and then analyzed for gaining meaning full data. Hadoop is used for analysis of web log in big data. The log files generated in the internet are large and complex to analyze. These log files generated are also error prone. As the days pass on the generation of number of log files increases and the storage devices are filled with files which are nowhere used. This happens in logs such as call logs, web logs etc. Big data is a technique used to analyze such data and give valuable information for future decision making. It can also be used in the concept of network coding. Clusters using network coding , to find energy efficient path in network.

Big data is a word which is used to store & process huge volumes of data(structured as well as unstructured) which helps a business in a regular or daily basis. Bigdata mainly consists of 5v's - Volume, Velocity, Variety, Veracity and Value. It is not concerned with the amount of data but how efficiently it can process that data and extract the required information. That extracted information or insights can help the organizations in better decision making and management.

When a web user surfs a specific webpage or website, the server records the small amounts of it within the web access log format. In the web access request log, you will see the types of files users measure accessing the situation from wherever the request has been created and alternative data like what browsers they are using and device access points. An access log could be a list of all the entries users have requested from an internet website. Such log files measure semi-structured data that is so hard to store, method and analyze visitors or accessed person's previous information from a warehouse system.

II. RELATED WORK

In [1][2] E. and S. Kasetty et al have discussed the importance of Map Reduce compared when compared to RDBMS for log processing. Different processing methods used in map reduce for logged data have been explained. These two works were extended to apply clusters on these data[3]. For heterogenous applications C. Ding [4][5] describes a unified logging structure. With slight modifications our model works on top of the above structures.

C. Faloutso et al [6] showed that in High Performance Computing logs can be utilized for identifying failures and resolving in systems which are large. These techniques mainly concentrate on differentiating the deleted log messages into a linear of failure events, and utilize the order for identifying the main reason for the problem. Two components are needed for log analyzer. One for recognizing the patterns in recorded messages and other for identifying the events and problems by matching the patterns toward in ward flow of messages[7][8]. The properties of log analyzer are :

No-supervision: Pattern Recognizer works from the staring as it does not have any previous knowledge. The pattern recognizer needs to be working from the scratch without any prior knowledge or individual administration. It does not take inputs given from individual administrator for every latest recorded message.

Revised Manuscript Received on March 17, 2020.
***** Correspondence Author
Vempaty Prashanthi, Assistant Professor Department of Information Technology, GRIET
Dr.Srinivas Kanakala, Assistant Professor, Department of CSE, VNRVJIET, Hyderabad,India.

Heterogeneity: Dissimilar appliances generate dissimilar messages[11-14] which are logged. Every appliance has its own format of generation of logged messages. Every format is to be recognized by the automated recognizer.

Efficiency: Records that are generated day to day in the internet is very fast in the systems such as internet of things, web logs, etc. Therefore, the processing of such fast generating records should also be very fast so that the rate at which the messages are processed is more compared to generation.

Scalability: The system should be expandable as huge recorded messages are generated and need to be processed without any memory issues.

The existing system [9][10] uses Relational Data Base Management System. It is a type of Database management system which store up the data or tuples in format of rows and columns. RDBMS is more dominant as it needs less number of guesses to know data can be retrieved out from the specific database. Therefore database is able to be seen in wide variety of ways or in different perspectives.

The RDBMS had been the one of the best solutions for all the things the database requires. RDBMS uses structured query language (SQL) to store, query, update and delete the contents in that specific database. However, the volume and velocity of this raw data have changed drastically in the past few years. It's continuously increasing every minute by minute.

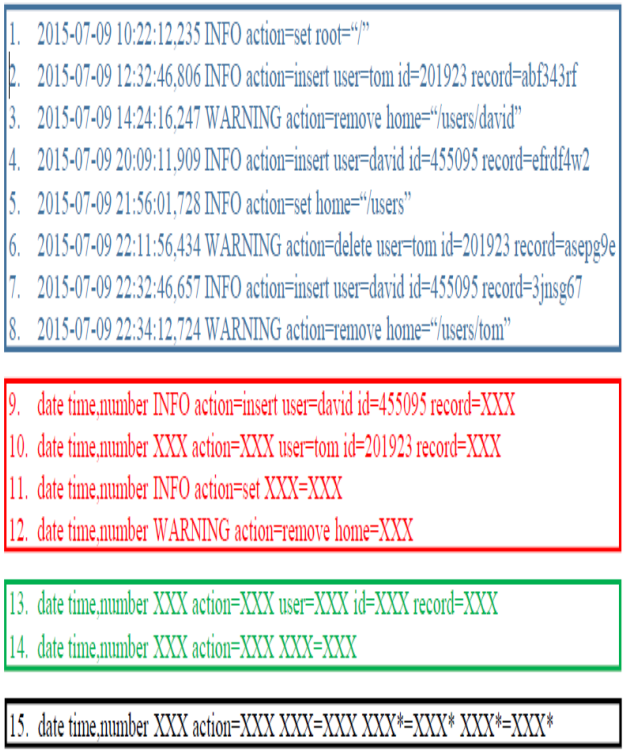

Fig 1: Patterns of logs.

*Limitations of using RDBMS for analysis:*The size of data has been increased rapidly to the range of pet bytes where one pet byte = 1,024 terabytes in number. Here, the RDBMS cannot handle large amounts of data. To address this issue, RDBMS added a greater number of CPUs to the DBMS to increase its capability.

Another limitation is that the majority of the data that comes from social media, audio, video is in a semi-structured or unstructured format. However, the RDBMS cant process these unstructured data. To handle such a huge amount of data high velocity is required. RDBMS doesn't support the high velocity data because it is designed not for the rapid growth but for the study data growth. Even if RDBMS tries to store and process these data, it may turn out to be much expensive.

III. PROPOSED SYSTEM

The proposed gem is by using "HADOOP Ecosystems". Big data terminology is used for huge data or data sets which are so large that normal processing applications cannot handle it. Big Data is a phrase used to mean a massive volume of all structured, semi structured and unstructured data. The data is generating at a rapid rate and in different number of formats that we cannot handle them. The social networking and mobile are the one contributing the highest amount of data generation. These above factors have progressed towards the term "big data". With the fast emergence of these data, traditional data processing techniques are unable to catch up with them. These factors have contributed for the adoption of Big data. To know why big data is much better compared to RDBMS for data analytics, advantages of big data for data analytics should be known.

Advantages of using hadoop for analytics:

- Identify the main causes of failure.
- Processing large volumes of data and extracting insigts.
- Understanding the usage of insights developed.
- Understanding the usage of marketing process through data driven process.
- Offering discounts or offers to the customers based on their buying habits.
- Improving the relationships between customer and vendor.
- Re evaluating the risk associated with that business.
- Enhancing the experience of customer.
- Enhancing the interactions or values for both online or offline customers.

System architecture:

Map Reduce is one of the most frequently utilized frame work to process huge amounts of structured or unstructured data stored in hadoop cluster. It was developed under Google to provide correspondence as well as reduce fault tolerance in data. Map reduce process large data in the form of key value pair. The key value pair is choosed depending upon our choice. These key value pairs are used for map reduce process as our system is not static. For static systems columns are used for analysis of data. Map reduce API provides the subsequent options such as instruction execution, parallel processing of huge amounts of data and high availability. Mapreduce work flow undergoes different stages which stores the output in hdfs with replications at the end. A Job tracker checks all the map reduce jobs which are working on hadoop cluster. A Job tracker has an important task in scheduling jobs and keeps track of every map and reduces jobs. Map reduce contains two processing stages map stage and reduce stage. Between these two stages there is one more stage called intermediate stage which takes the input from the mapper perform shuffling. Sorting and combining.

Three phases exists in this system.
1. Mapper phase 2. Intermediate Phase 3. Reducer Phase

1. Mapper phase :

Record reader contributes the inputs to the Mapper phase. The record reader is responsible to send key value pair to mapper. The inputs collected by the mapper is spit into keys values pairs. Depending upon the keys and partition constraints input is distributed to the specified reducer. The output generated is also a pair of keys values pairs. This is termed as intermediate key value pair.

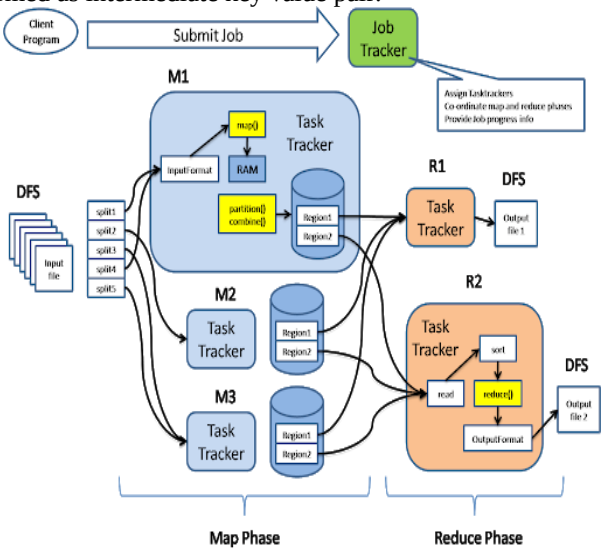

Fig 2: Map Reduce Frame Work.

2. Intermediate Phase

This stage comes in the middle of the guide and diminish stages. Right now activities are done dependent on the outcomes required. Right now same key qualities from various mappers will get into one mapper. Operations like shuffling, sorting and combing are done in this phase. It utilizes the Round Robin calculation to compose the middle of the road key qualities sets into the nearby plate.

3. Reducer Phase

This is the second phase of the Map Reduce information stream. Right now gets the contribution from the specialist and combiner. The reducer's rationale will start with the activities performed by the mapper. It creates the yield documents like part records which contains the real yield of the examined information. Each time when the activity is run reducer shows the quantity of reducers required for the activity for execution. As the reducer performs equal handling and subsequently the presentation and throughput of framework is expanded.

IV. METHODOLOGY

Hadoop distributed file system (HDFS) utilized to store gigantic informational collections or information and stream these enlightening lists at an extremely rapid exchange to different applications. HDFS facilitates easy access of data. As a single machine cannot hold gigantic or large information, the records of this data are stored in different machines. This data is stored in a redundant style to safe guard the data for any attacks or occurrence of disappointments. HDFS likewise makes applications accessible to parallel processing.

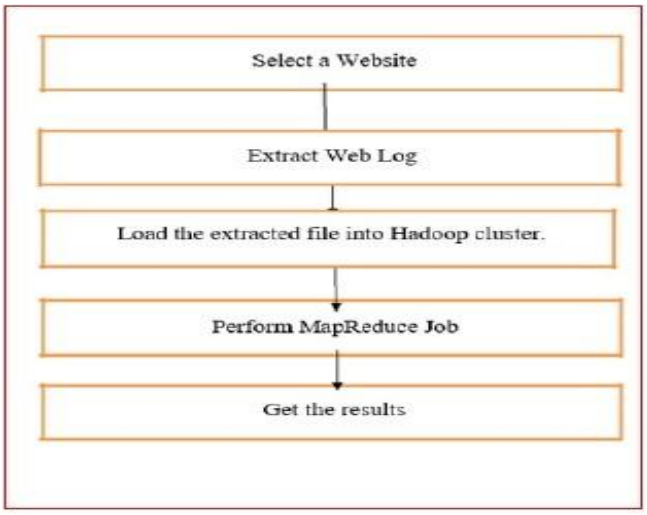

Fig 3: Process Flow.

V. IMPLEMENTATION

Step 1: Create a Webpage &host the webpage using bluehost hosting service.
Step 2: Extract the weblog data from the webpage
Step 3; Process the raw weblog data to obtain a cleaned data set(CSV format)
Step 4: Make another catalog with same name weblog investigation in the group.
Step 5: Write the map reduce process in Eclipse
Step 6: Make a jar file and duplicate the container document to nearby edge hub utilizing WinScp
Step 7: Login into the bunch utilizing clay and duplicate the information record from neighborhood to group.
Step 8: Run the mapreduce program.
Step 9: Result is seen through command interface.

VI. EXPERIMENTS AND RESULTS

In this experiment sequential and map reduce technologies are compared with each other. Synthetic data is produced by varying the count of log messages (12 millions by default) and count of patterns (1600 by default). The count of workers related to map reduces is also changed (9 by default) . Every worker contains 1GB amount of memory and single core of processor. As seen in Fig 4 the processing time of map reduce increases gradually contrast to execution of sequential. When 9 workers are used the speed of execution of map reduce raises to 5X when contrast to sequential procedure. It is to be observed that there are constant patterns here in. The map reduce process is capable of handling huge numbers of logs as the amount of patterns will not increase with respect to amount of logs in real human appliances. Fig 5 depicts that as the count of patterns is increased the processing time of map reduce process and sequential process is also increased. From Fig 6 depicts that as the count of workers is made double it decreases the execution time by 40 percentage.

In Fig 6, we show that doubling the number of workers reduces the running time by 40%. selected.

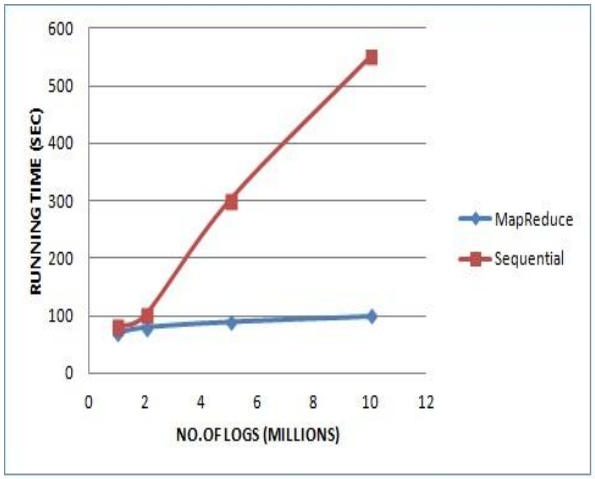

Fig.4. Runtime wrt no. of Logs.

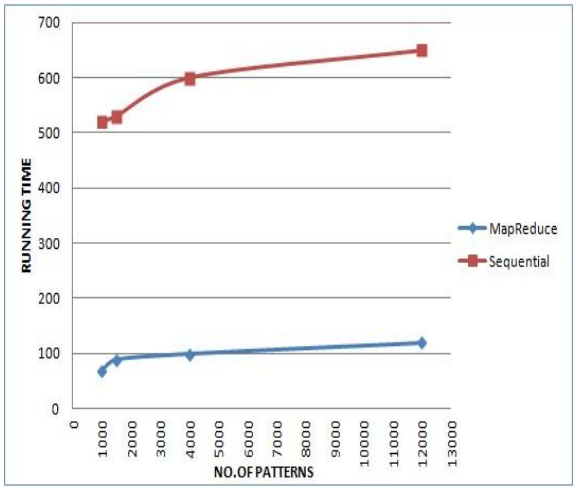

Fig.5.Runtime wrt no. of patterns.

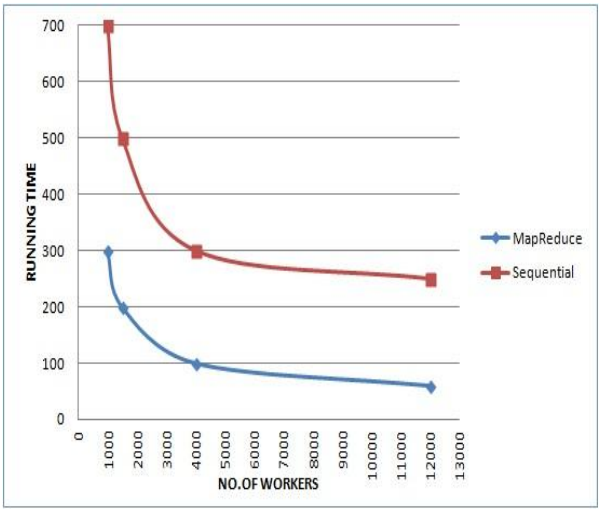

Fig.6: Runtime wrt no. of workers.

VII. CONCLUSION

Now a day's storing the information has turned into a major issue. Traditional database systems do not support large volumes of data. Therefore, a conventional database system called Hadoop is used for managing huge volumes of data. Number of elements and changed include in enormous information like web based life and cloud based life. These mechanical changes are putting weight on the appropriation of huge information. Big data is vastly improved than RDBMS for information investigation. From the outcomes we can analyze diverse sorts of IP addresses used timestamps, number of references by every client to the site and locate the top N users. Based on Number of Bytes Used information like most visited user can be identified. False snaps and unknown Ip Addresses can be blocked giving a safe domain to clients.

REFERENCES

1. E. and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery, 7(4):349{371, 2003.
2. M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. In ACM Sigmod Record, volume 28, pages 49{60. ACM, 1999.
3. S. Blanas, J. M. Patel, V. Ercegovac, J. Rao, E. J. Shekita, and Y. Tian. A comparison of join algorithms for log processing in mapreduce. In SIGMOD, pages 975{986. ACM, 2010.
4. C. Ding and J. Zhou. Log-based indexing to improve web site search. In SAC, pages 829{833. ACM, 2007.
5. M. Eltahir and A. Dafa-Alla. Extracting knowledge from web server logs using web usage mining. In Computing, Electrical and Electronics Engineering (ICCEEE), pages 413{417, Aug 2013.
6. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases, volume 23. ACM, 1994.
7. G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy. The uni_ed logging infrastructure for data analytics at witter. Proceedings of the VLDB Endowment, 5(12):1771{1780, 2012.
8. K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon. Parallel data processing with mapreduce: a survey. ACM SIGMOD Record, 40(4):11{20, 2012 .
9. A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover. Exact discovery of time series motifs. In SDM, pages 473{484. SIAM, 2009.
10. X. Ning and G. Jiang. HLAer: A system for heterogeneous log analysis, 2014. SDM Workshop on Heterogeneous Learning.
11. R. Rajachandrasekar, X. Besseron, and D. K. Panda. Monitoring and predicting hardware failures in hpc clusters with ftb-ipmi. In IPDPSW Workshops, pages.1136-1143. IEEE, 2012.
12. V.Prashanthi, K.Srinivas,"Identification of Opportunities for Coding in a Network", International Journal of Recent Technology and Engineering,Volume7, Issue5,Pages.140-144 ISSN: 2277-3878, January 2019.
13. V.Prashanthi, D.Suresh Babu, C.V.Guru Rao, Network Coding aware Routing for Efficient Communication in Mobile Ad-hoc Networks, "International Journal of Engineering & Technology (UAE)", ISSN: 2227-524X, Vol.7, No.3(2018), pp.1474-1481.
14. K.Srinivas, V.Prashanthi, "Energy-Efficient Cluster Based Routing Protocol in Mobile Ad Hoc Networks Using Network Coding", Journal of Computer Networks and Communications, Vol. 2014, Article ID 351020, 12 pages, 2014. ISSN: 2090-715X.
15. K.Srinivas, A.Venugopal Reddy," Connected Dominating Set-based Broadcasting in Mobile Ad-Hoc Networks using Network Coding", International Journal of Applied Engineering Research.

AUTHORS PROFILE

Vempaty Prashanthi, Assistant Professor Department of Information Technology, GRIET, pursuing Ph.D(Thesis submitted) from JNTU Hyderabad and has 14 years of academic and research experience. Her Ph.D work was on network coding aware routing techniques in mobile ad hoc networks. Prior to PhD, she had earned M.Tech in CSE from JNTUH in the year 2010 and B.Tech in computer science and information technology from JNTUH in the year 2005.Her research interests include developing algorithms and models for building systems and applications in areas of Data Mining, Machine Learning and Big-data analytics. She has around 15 publications, in various journals and conferences.

Dr. Srinivas Kanakala, Assistant Professor, Department of CSE, VNRVJIET, Hyderabad, completed Ph.D from Osmania University , Hyderabad and has 15 years of academic and research experience. Her Ph.D work was on network coding aware routing techniques in mobile ad hoc networks. Prior to PhD, he had earned M.Tech in CSE from JNTUH in the year 2008 and B.Tech in computer science and information technology from JNTUH in the year 2004.Hisr research interests include developing algorithms and models for building systems and applications in areas of Data Mining, Machine Learning and Big-data analytics. She has around 15 publications, in various journals and conferences.