LOG ANALYSIS TECHNIQUES: A BRIEF STUDY

Jui Pattanaik
Computer Science & Engineering,
Aryan Institute Of Engineering & Technology, Bhubaneswar
Santosh Kumar Sharma
Computer Science & Engineering,
Nm Institute Of Engineering & Technology, Bhubaneswar
Purnya Prava Nayak
Computer Science & Engineering,
Capital Engineering College, Bhubaneswar
Manisha Pradhan
Computer Science & Engineering,
Raajdhani Engineering College, Bhubaneswar

Abstract— Log Analysis is a critical procedure in most framework and system exercises where log information is utilized for different reasons, for example, for execution checking, security examining or notwithstanding for revealing and profiling. Nonetheless, as years cruised by, the volume of log information increments alongside the span of the framework just as the quantity of clients included. Customary or existing log analyser instruments are not ready to deal with the huge measure of information. Thusly, Big Data is the answer for defeated this issue. The principle motivation behind this paper is to introduce a survey of log document investigation in Big Data condition dependent on past research works. This paper likewise features the qualities of Big Data just as Hadoop Framework that has been generally utilized as Big Data application. Results from the papers assessed demonstrate that dominant part analysts connected MapReduce as the principle segment of Hadoop for investigating the log records and HDFS as the information stockpiling. Past analysts have likewise utilized different instruments and calculations together with the Hadoop Framework for investigation purposes. The discoveries of this paper will give an intelligible audit of Hadoop use execution in breaking down various kinds of log records and prescribe justifiable outcomes for end clients to use in future work.

Keywords— Hadoop; HDFS; log analysis; log files; Map Reduce

I. INTRODUCTION

Since decades back, log information has been assuming a significant job in PC framework. There are different sorts of log that record various types of exercises for PC framework, applications, arrange traffic or even web servers. Each detail of the log information is pivotal in deciding the status and state of a running framework. In this way, log information is one of the principle sources in checking and investigating assorted frameworks and there are numerous devices planned explicitly for breaking down them.

As years cruised by, the volume of log information increments alongside the measure of the frameworks just as the quantity of clients included. This is when information transformed into Big Data and all the huger information issue emerges as far as volume, speed, or potentially assortment that surpasses the capacities of most current innovation (Joshi, 2013). Huge information is really alluding to the engineering and advancements that are built up to catch, store, procedure and run better quality volumes of information yet with less measure of time or continuously. In spite of the fact that the monstrous volume of log can contribute unmistakably progressively extensive and important data, breaking down them is certainly an incredible test. In the customary methodology, the accessible information is handled utilizing a ground-breaking PC. In any case, there is normally a breaking point to the span of information being prepared, as it isn't versatile, while Big Data extends with extraordinary speed and assortment.

In the event that massive amount of information confines the adequacy of log record examination, at that point Big Data would be the arrangement. In this manner, this paper will talk about the log records and the normal for Big Data just as other research works that actualize Hadoop system for investigating different kinds of log documents. Hadoop is presented in this paper as an open-source system that permits appropriated handling of gigantic informational indexes on groups of PCs which can conquer the log examination issue. The fundamental segments of Hadoop, MapReduce and Hadoop Distributed File System (HDFS) are likewise examined in this paper.

II. RELATED WORKS UTILIZING HADOOP

The quantity of research proposing the utilization of Hadoop Framework in comprehending Big Data issue is expanding every year. A large number of them have been investigating and prescribing new log examination approach utilizing Hadoop in particular spaces or zones. This paper examined the related works by different specialists that centers around log examination in Big Data condition utilizing Hadoop structure.

1. (Vernekar and Buchade, 2013) proposed the utilization of Hadoop structure utilizing MapReduce on log examination for framework dangers and issue distinguishing proof. The MapReduce calculation comprises of Map stages and Reduce stages. The log record that should be prepared will be the contribution to the Map stage while the yield of each guide stage will at that point be given to specific keys. The Reduce capacity will at that point give the last outcome or log report. The proposed framework gives a productive method for gathering and corresponding log so as to recognize the framework dangers and issues. The scientists found that the proposed framework has huge improvement accordingly time, which is accomplished with the utilization of MapReduce.

2. (Wang et al., 2014) have structured and implemented and undertaking Weblog examination frameworks dependent on the architecture of Hadoop with HDFS (Hadoop Distributed File System) and MapReduce just as Pig Latin Language. The primary motivation behind their framework is to help framework heads to rapidly catch and examine information covered up in the gigantic potential esteem, in this manner giving a significant premise to business choice. The exploration that has been done demonstrated that the structure of MapReduce program is a compelling answer for exceptionally expansive Weblog records in the Hadoop condition. Other than that, the log prerequisite is anything but difficult to break down utilizing Pig programming language that additionally gives better execution. The framework prevailing with regards to giving AP server traffic insights that assistance the framework directors to distinguish potential issues and anticipate the future pattern. Figure-7. Weblog investigation framework flowchart utilizing Hadoop (Wang et al., 2014).

3. (Nandimath et al., 2013) have proposed a plan to conquer the issue of investigation of enormous information utilizing Apache Hadoop in their examination. The preparing includes four stages which incorporate making a server of required design utilizing Amazon web administrations, bringing in information from a database to Hadoop, performing employments in Hadoop and trading information back to the database. In stage two, information is put away in Mongo DB, which is a NoSQL database. At that point, MapReduce is utilized to perform six Hadoop occupations that are actualized in spring structure. A Hadoop work comprises of the mapper and reducer capacities. The created yield of information preparing in the Hadoop work must be traded back to the database. The old qualities in the database must be refreshed promptly so as to avoid loss of significant information. The analysts concurred that the application can play out a task on enormous information in ideal time and produce a yield with least use of assets.

4. (Therdphapiyanak and Piromsopa, 2013) proposed applying Hadoop for Log Analysis of Apache web servers. They utilized circulated K-Means grouping calculation dependent on Mahout/Hadoop Map-Reduced model to break down high volume of log documents. Their discoveries demonstrated that the execution was superior to an independent log analyzer as it was equipped for supporting an immense size of log. Their strategy turned out to be ready to extricate information for a million passages of logs as it can't be acquired without the adaptability of Hadoop and the proposed investigation.

5. (Hingave and Ingle, 2015) proposed a log analyzer with the mix of Hadoop and MapReduce worldview (3) Pig Program MapReduce Job Reports Result Analysis (1) Data Pre-handling (2) Upload Weblog Raw Data Hadoo MapReduce HDFS MapReduc HDFS Pig (4) VOL. 10, NO. 23, DECEMBER 2015 ISSN 1819-6608 ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). All rights saved. www.arpnjournals.com 17782 utilizing NASA's web log document of size 77MB of 445,454 records. They directed a trial to complete an examination between MySQL (RDBMS) and Hadoop in dissecting the clients' action of the web. The outcomes acquired, demonstrated that the proposed log analyzer improved reaction time as the time required for ETL procedure and investigation utilizing Hadoop is estimated multiple times not exactly the MySQL.

6. (Yang et al., 2013) proposed an examination of framework for stream log which focused on the system traffic follows in China to defeat the growing size of the stream logs which have expanded to 870GB every day for single city. Hadoop has been proposed as an answer for make the examination quicker and it is additionally ready to dissect bigger dataset. The framework utilizes HDFS for the logs stockpiling and MapReduce Framework for investigation work and furthermore for making their own content called Log-QL.

The consequences of the investigation demonstrate that the new framework empowered them to break down TBs of information contrasted with the current brought together framework that can just process up to 10GB of information. Nonetheless, the framework will just perform better as the span of information develops.

7. (Narkhede, Baraskar, and Mukhopadhyay, 2014) connected Hadoop MapReduce programming model for examining web log records in distributed computing condition so as to recover the hit mean explicit web application. The test utilizes HDFS to store the web log document and MapReduce programming model is utilized to compose application for dissecting log record. The log documents that have been utilized contain 100,000 records with each log having diverse fields of URL, date, hit, age and others. The connected model of utilizing HDFS and MapReduce has given investigated results in insignificant reaction time. While the execution test results against number of records, the quantity of hubs in the group demonstrate that the execution of the framework will increment alongside the expansion in number of hubs.

Table-1. Summarized related research of log analysis using Hadoop

Authors	Hadoop component	Other tools or Algorithm	Type of log analysis	Results of Research
(Vernekar and Buchade, 2013)	MapReduce		Analyzing sys log to identify system threats	Improve response Time
(Wang *et al*., 2014)	HDFS, MapReduce	Pig language	Weblog analysis	Better performance able to predict trend
(Nandimath *et al*., 2013)	MapReduce, HDFS	MongoDB	Amazon log files	Optimal operation Time
(Therdphapiya nak and Piromsopa, 2013)	MapReduce Mahout	K-Means	Apache web server log	Support huge file Size
(Hingave and Ingle, 2015)	MapReduce	MySQL (RDBMS)	NASA's web log files	Improve response Time
(Yang *et al*., 2013)	HFDS, MapReduce	Log-QL (script)	Flow logs of China network traffic	Able to analyse larger dataset (TBs) Perform better as the data size grows
(Narkhede, Baraskar, and Mukhopadhya, 2014)	HDFS, MapReduce		Web log files in cloud computing	Minimal response Time Perform better as the number of nodes increase

III. RELATED WORK USING OTHER FRAMEWORKS

In this section, studies conducted, optimizations made and applications built using Elasticsearch, Logstash and extensions of the same are discussed.

1. In Sun, M., Convertino, G., & Detweiler, M. (2016), the plan tends to two basic prerequisites: empowering non-specialized partners to investigate logs and enabling them to do job explicit examination and offer the outcomes. To satisfy these prerequisites, they have previously created and approved a theoretical system with 4W dimensions to test the feasibility of a unified platform. At that point, the authors made proper tests to approve the 4W theory and instantiated the 4W system as a data model.
2. In Debnath, B., Solaimani, M., Gulzar, M. A. G., Arora, N., Lumezanu, C., Xu, J., ... Khan, L. (2018), LogLens gives an outline to executing an ongoing log examination framework by utilizing unsupervised AI based strategies. It can distinguish oddities with no (or insignificant) human inclusion and can without much of a stretch adjust to the framework conduct change. It sends log analysis as a service using the Spark Big data handling framework.
3. Traditional anomaly detection that relies heavily on manual log inspection becomes impossible due to the sharp increase of log size. However, developers are still not aware of the state-of-the art anomaly detection methods, and often have to re-design a new anomaly detection method by themselves, due to the lack of a comprehensive review and comparison among current methods. In He, S., Zhu, J., He, P., & Lyu, M. R. (2016), They fill this gap by providing a detailed review and evaluation of six state-of-the-art anomaly detection methods which include 3 supervised and 3 unsupervised methods to detect anomalies in the log files.
4. In Delic, K. A., & Riley, J. A. (2015), the system they've given is named smart Log Analytics (SLA) – as they aim to supply an involuntary, effective and economical system from that they expected to develop an applicable set of technologies to change the price effectiveness equation for the particular target of Enterprise Applications. However, they didn't develop to cloud-resident applications. For those varieties of applications, the quantity of logs is going to be a lot larger and completely different from legacy enterprise applications, however their hope is that experiences gained with SLA can alter some advantages for more analysis within the space of cloud-resident applications.

5. There was also a study done on the importance of log parsing and its advantages in analysis of the log entries to draw insights in P. He, J. Zhu, S. He, J. Li, and R. Lyu(2016). The concepts behind the log parsing is making a sense out of the log entries which have multiple fields arranged in a different sequence depending on the type of the service writing the logs. This paper reviews the various log parsing techniques and draws conclusion that there cannot be a standard parser as the log messages are different for different type of services.
6. In Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu(2018) ,They first present a portrayal investigation of the present cutting-edge log parsers and assess their adequacy on five genuine world datasets with more than ten million log messages. They confirm that, despite the fact that the general precision of these parsers is high, they are not strong overall datasets. At the point when logs develop to an extensive scale (e.g., 200 million log messages), which is regular by and by, these parsers are not sufficiently effective to deal with such information on a solitary PC. To address the above constraints, they have structured and executed a parallel log parser (to be specific POP) over Spark, a vast scale information preparing stage. Far reaching tests have been directed to assess POP on both engineered and genuine world datasets. The assessment results show the ability of POP as far as exactness, proficiency, and viability on consequent log mining assignments.

IV. CONCLUSIONS

The study on various existing frameworks and systems using Hadoop and other frameworks gives the importance in analyzing the log files to draw insights. The study also reveals various techniques that can be employed to analyze log files and extract actionable insights from the log files.

REFERENCES

1. Bhandare, Milind, Kuntal Barua, and Vikas Nagare. 2013. Generic Log Analyzer Using Hadoop Mapreduce Framework. International Journal of Emerging Technology and Advanced Engineering 3(9): 603-7.
2. Nandimath, J., E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi. 2013. Big Data Analysis Using Apache Hadoop. In IEEE 14th International Conference on Information Reuse and Integration (IRI), 700–703.
3. Narkhede, S., T. Baraskar, and D. Mukhopadhyay. 2014. Analyzing Web Application Log Files to Find Hit Count through the Utilization of Hadoop MapReduce in Cloud Computing Environment. In 2014 Conference on IT in Business, Industry and Government (CSIBIG), 1-7.
4. Therdphapiyanak, Jakrarin, and Krerk Piromsopa. 2013. Applying Hadoop for Log Analysis toward Distributed IDS. In Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, 3:1–3:6. ICUIMC '13. New York, NY, USA: ACM.
5. Vernekar, S.S., and A. Buchade. 2013. MapReduce Based Log File Analysis for System Threats and Problem Identification. In Advance Computing Conference (IACC), 2013 IEEE 3rd International, 831–35.
6. Wang, Chen Hau, Ching TsorngTsai, Chia Chen Fan, and Shyan Ming Yuan. 2014. A Hadoop Based Weblog Analysis System. In 2014 7th International Conference on Ubi-Media Computing and Workshops (UMEDIA), 72-77.
7. Yang, Jie, Yanshen Zhang, Shuo Zhang, and Dazhong He. 2013. Mass Flow Logs Analysis System Based on Hadoop. In 2013 5th IEEE International Conference on Broadband Network Multimedia Technology (IC-BNMT), 115-18.
8. Sun, M., Convertino, G., & Detweiler, M. (2016). Designing a Unified Cloud Log Analytics Platform. 2016 International Conference on Collaboration Technologies and Systems (CTS).doi:10.1109/cts.2016.0057.
9. Debnath, B., Solaimani, M., Gulzar, M. A. G., Arora, N., Lumezanu, C., Xu, J., … Khan, L. (2018). LogLens: A Real-Time Log Analysis System. 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS).doi:10.1109/icdcs.2018.00105.
10. He, S., Zhu, J., He, P., & Lyu, M. R. (2016). Experience Report: System Log Analysis for Anomaly Detection. 2016 IEEE 13th International Symposium on Software Reliability Engineering (ISSRE).doi:10.1109/issre.2016.21 .
11. Delic, K. A., & Riley, J. A. (2015). SLA : Smart log analytics. 2015 XXV International Conference on Information, Communication and Automation Technologies (ICAT) .doi: 10.1109/icat.2015.7340517.
12. P. He, J. Zhu, S. He, J. Li, and R. Lyu. An evaluation study on log parsing and its use in log mining. In DSN'16: Proc. of the 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2016.
13. Pinjia He, Jieming Zhu, Shilin He, Jian Li, and Michael R. Lyu, Towards Automated Log Parsing for Large-Scale Log Data Analysis. IEEE Transactions on Dependable and Secure Computing (Volume: 15 , Issue: 6 , Nov.-Dec. 1 2018).