ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

Operational-Log Model for Big Data Systems

[1]JASMINE, [2] M. ARATHI

[1]M. Tech Student, Department of SE, School of Information Technology (JNTUH), Kukatpally, District RangaReddy, Telangana, India

[2]Assistant Professor, Department of CSE, School of Information Technology (JNTUH), Kukatpally, District RangaReddy, Telangana, India

ABSTRACT— ***Big knowledge systems (BDSs) are advanced, consisting of multiple interacting hardware and computer code elements, like distributed computing nodes, databases, and middleware. Any of those elements will fail. Finding the failures' root causes is extraordinarily grueling. However, as years glided by, the degree of log knowledge will increase at the side of the scale of the system yet because the range of users concerned. Ancient or existing log instrument tools don't seem to be able to handle the large quantity of knowledge. Therefore, massive knowledge is that the resolution to beat this issue. The most purpose of this paper is to gift a review of log file analysis in massive knowledge surroundings supported previous analysis works. This paper conjointly highlights the characteristics of huge knowledge yet as Hadoop Framework that has been wide used as massive knowledge application. Results from the papers reviewed shows that majority researchers applied MapReduce because the main element of Hadoop for analyzing the log files and HDFS because the knowledge storage. Previous researchers have conjointly used alternative tools and algorithms alongside the Hadoop Framework for analysis functions.***

1. INTRODUCTION

Since decades past, log information has been enjoying a vital role in automatic data processing system. There are varied kinds of log that record totally different forms of activities for automatic data processing system, applications, network traffic or maybe internet servers. each details of the log information are crucial in deciding the standing and condition of a running system. Therefore, log information is one in all the most sources in observation and analyzing numerous systems and there are several tools designed specifically for analyzing them

To pinpoint a problem's root cause, analysts usually examine operational information logs and traces generated by the BDS parts. A log or trace could be a sequence of temporal events captured throughout a specific execution of a system. As an example, a log will contain software system execution methods, events triggered throughout software system execution, or user activities. No clear distinction exists between logs and traces. Often, the term "log"

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

represents however a program is employed (such as security logs), whereas "tracing" captures program's parts that are invoked in a very given execution of the system.

Tracing is employed for debugging and program understanding. During this article, we have a tendency to primarily use the term "log." Logs share many characteristics that build operating with them troublesome in industrial settings:

***Velocity*.** The data (in some cases) requires real-time processing.

***Volume*.** Logs can contain huge amounts of historical data.

***Variety*.** The captured data can be structured or unstructured.

Veracity. The captured data requires cleaning.

***Value*.** Not all the captured data is useful.

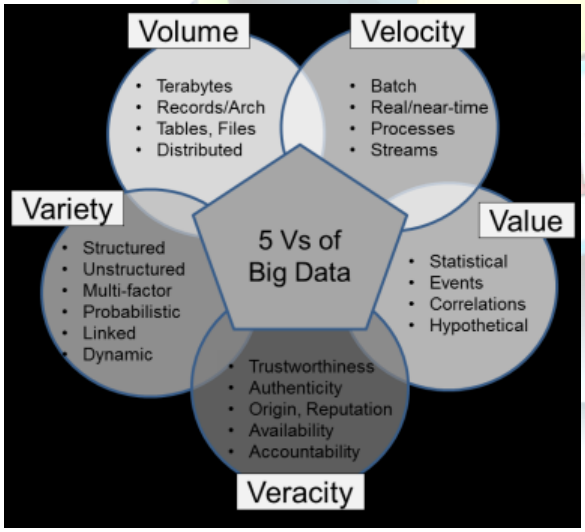


Fig 1: 5Vs of Big Data

These characteristics also describe big data. Essentially, BDSs designed to process big data usually emit big data (captured in logs) themselves.1 Of course, not all BDSs generate large volumes of logs. Also, small systems might generate big data. However, most BDS-emitted logs will exhibit at least one big data characteristic. To leverage log data, developers need ways to effectively deliver, store, and crunch large volumes of data.

In recent research the Big Data has been suggested to have 2 more characteristic which are

Visualization: The data needs to be readable and easily understood while various optimization algorithms will provide an advantage of providing an optimal review of the data analyzed.

Variability: This characteristic allows Big Data to handle uncertainty in data with changing of data helping in prediction of future behavior of the subjects.

2. RELATED WORK

Issues arise after you should store and compare an outsized volume of logs. One issue arises whereas you're uploading a log to an overseas storage facility for processing. Playing the analysis onsite is typically difficult thanks to the dearth of resources and tools that can diagnose the problem's cause onsite. A log, even compressed by a thought repository utility such as nada, will reach tens of gigabytes. If the log is collected in-house, repeating the fle from the machine on that the log was collected to the storage facility is quick and simple as a result of internal networks are generally quick.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

3. PROPOSED WORK

3.1 Log Files

Log files are record files that are generated mechanically by the supply system in nearly all digital devices. They contain vast quantity of data that is important for creating business selections or troubleshooting. as an example, log files can keep data of everything that gets in and out of the online servers. the online servers shall record within the log files the quantity of clicks, visits or different relevant net users' records that are sometimes hold on in predefined file format Analysis of log files has been important in breakdown several problems. The contribution of the log analysis is categorised into four totally different areas

a) Performance: Log analysis is employed in optimization or debugging method for activity the system performance. Logs within the case of performance facilitate the administrator to grasp however resources of explicit system are used.

b) Security: .Logs for security functions square measure normally won't to discover breaches or misdeed and to perform postmortem investigation of security incidents. as an example, intrusion discover ion wants the reconstruction of sessions from logs so as to detect unauthorized access into a system.

c) Prediction: Logs are notable to be able to turn out prediction info. There are prophetic analysis tools that use log information to assist in selling strategy, promotion placement or inventory management.

d) Reporting and profiling Analyzing: .logs is additionally required in identification resource utilization, employment or user behavior. For instance, logs can record the tasks' characteristics from a cluster's employment so as to profile resource utilization for giant information center.

3.2 Log Files Analysis

When the log information has been collected, it's to travel through a preprocessing stage before continuing to the log analysis stage. The info preprocessing shows the ETL (Extract, Transform, and Loading) is truly a neighborhood of the ETL that rework information to a desired format. It is crucial for the information to endure preprocessing operation so as to touch upon numerous imperfections in raw collected data because it might contain noise like errors, redundancies, outliers and different ambiguous information or missing values.

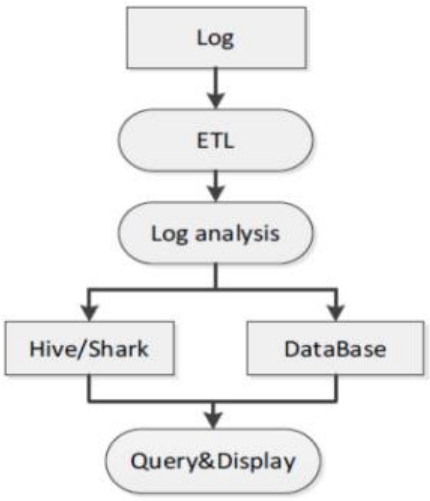


Fig 2: Log Files Analysis

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

The most operations in information preprocessing square measure mainly: information improvement information refinement: Handle missing values and noise yet as information inconsistency. Information integration: A method of group action duplicated information. information transformation: The collected information are going to be regenerate to the format of the destination system advised a number of steps of preprocessing within the analysis of journal analysis that involve removing moot attributes or records that have missing valuable information and reworking URLs into code numbers so as to urge clean information. Not the entire log records square measure helpful or necessary. Therefore, before the method of journal information analysis is being done, the info improvement part must be applied. The info improvement method involves removing:

Records that have missing worth information, as an example, once the execution method are suddenly terminated; the log file record isn't fully recorded. outlawed records that have exception standing numbers as an example four hundred or 404 that caused by protocol consumer errors, unhealthy requests or a call for participation not found. moot records that don't have any vital URLs. There square measure some files that square measure generated mechanically once website is requested, as an example .txt, .jpg, .gif or .js extensions. Therefore, in Log Analysis, the aim of implementing log preprocessing is to enhance the log quality and to extend the results accuracy. The preprocessing part helps to filter and to arrange solely acceptable data that is employed before applying the Map scale back formula in order that it's going to not have an effect on the analysis result. [2] proposed a system which is an innovative congestion control algorithm named FAQ-MAST TCP (Fast Active Queue Management Stability Transmission Control Protocol) is aimed for high-speed long-latency networks. Four major difficulties in FAQ-MAST TCP are highlighted at both packet and flow levels. The architecture and characterization of equilibrium and stability properties of FAQ-MAST TCP are discussed. Experimental results are presented comparing the first Linux prototype with TCP Reno, HSTCP, and STCP in terms of throughput, fairness, stability, and responsiveness. FAQ-MAST TCP aims to rapidly stabilize high-speed long-latency networks into steady, efficient and fair operating points, in dynamic sharing environments, and the preliminary results are produced as output of our project. The Proposed architecture is explained with the help of an existing real-time example as to explain why FAQ-MAST TCP download is chosen rather than FTP download.

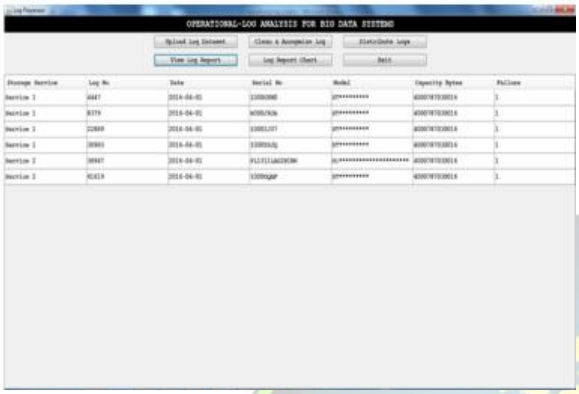
4. EXPERIMENTAL RESULTS

In this experiment, we have to upload the log dataset into the log processor. The uploaded dataset contains the information like, date of mfg, serial number, model and failure information. After successfully loading the dataset we have to clean and anonymize data. Here, cleaning means here it removes the unwanted fields from dataset and for privacy issues will anonymize the data and save in clean.txt. And after, the uploaded log information will be distributed

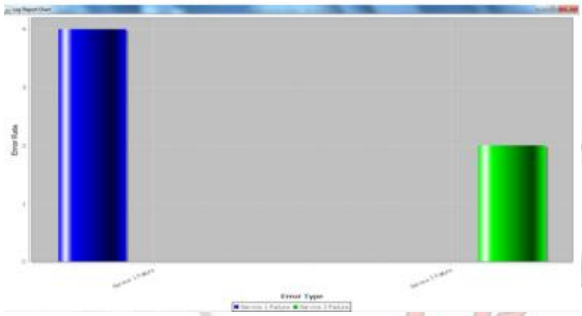
ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

among different storage devices to find the failure devices. Here we are using Map Reducer for parallel processing of data.



Log report chart:



5. CONCLUSION

The bigger in size the log records are, the more valuable the know-how they are able to furnish. Nevertheless, processing colossal sized log knowledge can be very challenging and it's definitively no longer an easy challenge. For this reason, the log files should be analyzed in colossal data atmosphere utilizing the enormous data utility reminiscent of Hadoop. The issues and solutions we discussed here should be of interest to practitioners because they can readily leverage existing techniques to build their own solutions. Our findings should also be of interest to the academic community because they highlight unsolved practical problems.

REFRENCES

[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] Christo Ananth, S.Esakki Rajavel, I.AnnaDurai, A.Mydeen@SyedAli, C.Sudalai@UtchiMahali, M.Ruban Kingston, "FAQ-MAST TCP for Secure Download", International Journal of Communication and Computer Technologies (IJCCTS), Volume 02 – No.13 Issue: 01 , Mar 2014, pp 78-85

[3] Collins, E. 2014. Intersection of the Cloud and Big Data. IEEE Cloud Computing 1 (1): 84–85.

[4] Gupta, A. 2015. Big Data Analysis Using Computational Intelligence and Hadoop: A Study. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 1397–1401.

[5] Hingave, H., and R. Ingle. 2015. An Approach for MapReduce Based Log Analysis Using Hadoop. In 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 1264–68.

[6] Joshi, Sanat. 2013. Big Data. InTech 60 (3): 40–43.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 17, March 2017

[7] Kim, Yong-Hyun, and Eui-Nam Huh. 2014. A RuleBased Data Grouping Method for Personalized Log Analysis System in Big Data Computing. In 2014 Fourth International Conference on Innovative Computing Technology (INTECH), 109–14.

[8] K P, Ajay, Dr K. C. Gouda, and Dr Nagesh H R. 2015. A Study for Handling of High-Performance Climate Data Using Hadoop. IJITR 0 (0): 197–202.

[9] Li, Meijing, Xiuming Yu, and Keun Ho Ryu. 2014. MapReduce-Based Web Mining for Prediction of Web-User Navigation. Journal of Information Science 40 (5): 557–67.

[10] Lin, Xiuqin, Peng Wang, and Bin Wu. 2013. Log Analysis in Cloud Computing Environment with Hadoop and Spark. In 2013 5th IEEE International Conference on Broadband Network Multimedia Technology (IC-BNMT), 273-76.