IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Big Data Analysis of Fetched Logs using Hadoop Framework

Prof. Girija Chiddarwar[1], Sanket Chhajed[2], Sameer Deshmukh[2], Pranoti Dongre[2], Bharti Nile[2]

Associate Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India[1]

Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India[2]

Abstract: With every second slipping away, there are thousands of activities carried out on the internet. The big data generated out of logs is very crucial and important to provide a superior backbone to the whole IT infrastructure. It is not an easy task to analyse the logs generated by system as there are numbers of discrete type machines available in a network. The growing size and complexity of log files has made the manual analysis of system logs by administrators prohibitive. This fact makes it important for tools and techniques that will allow some form of automation in management and analysis of system logs to be developed. The proposed technique uses Hadoop framework to process huge amount of data. To extract useful information data mining technique is used, among which clustering is most popular. Various types of clustering like Connectivity-based clustering, Centroid-based clustering (K-Means Clustering), Density-based clustering, etc. can be used to cluster the log files. Centroid-based (K-Means Clustering) is the most effective technique amongst all. The input logs will be clustered using K-Means clustering and then processed in Hadoop framework. The analysis result can be used for detecting security threats in network and inform the administrator about the Intrusion Detection, SQL-Injection, system error, etc.

Keywords: Big Data, Hadoop, Log Analysis.

I. INTRODUCTION

Big data analytics has the power to change the world. Hadoop, is the platform for the distributed big data, also plays an important role in big data analytics. Organizations now realize the inherent value of transforming these big data into actionable insights. Data science is the highest form of big data analytics that produce the most accurate actionable insights, identifying what will happen next and what to do about it. Looking at the runtimes for analytical algorithms, it can be easily seen that limitations in terms of data set sizes have vanished today – but at the price of larger runtimes. This is in all cases prohibitive for interactive reports, but likely also for predictive analytics if the model creation has to be done fast or in real-time. In those cases, an in-memory engine is still the fastest option. The in-Hadoop engine is slow for smaller data sets but is the fastest and sometimes the only option when data sets are really big in terms of volume.

Web log files records activity information when a web user submits a request to web server. The main source of raw data is the web access log which is known as Log file. A log files contain various parameters which are very useful in recognizing user browsing patterns. The request information sent by the user via protocol to the web server is recorded in log file. The log file entry contains IP address of the computer making the request, the visitor data, line of hit, the request method, location and name of the requested file, the HTTP status code, the size of the requested file and etc. Log files can be classified into categories depending on the location of their storage that is Web Server Logs and Application Server logs. Web servers maintain at least two types of log files: Access logand Error log. The access log records all requests that were made on this server.

II. RELATED WORK

Yan liu, Ning Cao, Wei pan, Guangwei Qiao[1]proposed a system for anomaly detection which is very important for development, maintenance and performance refinement in large scale distributed system. In this paper, map-reduce based framework is implemented to analyze the distributed logs for detecting anomaly. K-means clustering algorithm is used to integrate the collecting logs and then map-reduce based algorithm is used to parse the collected cluster log files.

Nadeem Akhtar, Mohd Vasim Ahamad, Shahbaaz Ahmad[2] implemented a Hadoop distributed file system and map reduce programing model is used for storage and retrieval of big data. In this paper, experiment show work done on Hadoop by applying number of files as input and then analyzing performance of Hadoop system.

According to the words of **Amrit Pal, Kunal Jain, Pinki Agrawal, Sanjay Agrawal[3]** log is the main source of system operation status, user behavior and systems actions. Log analyses helps to improve the business strategies as well as to generate statistical report. The joint of Hadoop and map-reduce programing tool make it

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

possible to provide batch analysis in minimum response time and in-memory computing capacity in order to process log in high level efficient and stable way.

Jie Yang, Yansheen ZZhang, Shuo Zhang, Dazhong HE [4] proposed that the digital world is impractical and inefficient to use the traditional data base management techniques on big data. Hadoop is an open source framework, which can be used to process the huge amount of data in parallel. To extract useful information, data mining techniques can be used. Among many techniques of data mining clustering is most popular. In this paper, K-means clustering method by using improved initial center is proposed.

Hemant Hingave, Prof.Rasika Ingle[5]introduced the idea and process of k-mean algorithm and focus on integration of location and property of k-mean clustering algorithm. the paper analyzes the parallel design and implementation process of algorithm in hadoop platform.

III.PROPOSED TECHNIQUE

The proposed technique consists of scanning logs in real time as well as in off time which can give the best result of analysis of logs that can be studied by IT admin to take preventive measures for the attacks.

The real-time logs can be stored and processed by using flume. The in-memory engine for analytics is still the better option for faster processing such big data of logs. The attacks like SQL Injection, Cross-Site Scripting (XSS), URL Injection, etc. are the major attacks that threatens the security of any IT Infrastructure in today's world. To detect this attack there are various techniques available, like, to detect URL Injection we can look over the status code in the log file and conclude that there has been an attack performed. Similarly, for SQL Injection we can see the queries of modification that are used in SQL database and forXSS "<script>" can be seen in the log file where attacker have embedded the malicious code.

Hadoop, being an open framework which stores and process big data can be easily used to detect attacks by using some form of regular expressions. Once the all-day log is collected by an admin, then it can be uploaded to Hadoop HDFS which can be then analysed by firing mapper, reducer and driver commands manually.

The collective result of real-time and manually performed analysis will be then displayed on the web page for effective and easy interaction.

IV. CONCLUSION

The analysis of log files in real time and off time plays an important role in any IT Infrastructure so as to make them strong enough to withstand against the hackers. The real-time log files can be used to ingest the logs in Hadoop for storage and analysis purpose. Once the analysis is done the result can be displayed on web page which is more interactive as compared to Hadoop interface. After the real time the all-day logs can be also added to Hadoop manually in which analysis of selected attacks can be performed by an admin which can be useful in finding out the attacks which were missed during real-time. Hence the attacks can be doubly verified, giving a strong support to IT Infrastructure.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude towards our guide **Prof. G. G. Chiddarwar** for her invaluable guidance and supervision that helped us in our research. She has always encouraged us to explore new concepts and pursue newer research problems. I credit our project contributionto her. Collectively, we would also like to thank our project committee members Prof. Bhojane and **Prof. S. D. Wable** for their time, suggestions, and for graciously agreeing to be on our committee, and always making themselves available.

REFERENCES

[1]. Yan liu, Ning Cao, Wei pan, Guangwei Qiao, "System anomaly Detection in Distributed System through MapReduce-Based Log Analysis", 2010 3rd International conference.
[2]. Nadeem Akhtar, Mohd Vasim Ahamad, Shahbaaz Ahmad, "Map-Reduce Model of Improved K-Means Clustering Algorithm Using Hadoop MapReduce", 2016 Second International Conference.
[3]. Amrit Pal, Kunal Jain, Pinki Agrawal, Sanjay Agrawal, "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop", 4th international conference.
[4]. Jie Yang, Yansheen ZZhang, Shuo Zhang, Dazhong HE, "Mass flow logs analysis system based on Hadoop", Proceedings of IEEE.
[5]. Hemant Hingave, Prof.Rasika Ingle, "An approach for Map-Reduced based Log Analysis using Hadoop", IEEE sponsored 2nd international conference.
[6]. http://hadoop.apache.org/