IJMRT: Volume (1), Issue 3, May 2019

ISSN (Print) : 2663-2381
ISSN(Online) : 2663-4007

International Journal of Multidisciplinary Research Transactions

(A Peer Reviewed Journal)

www.ijmrt.in

Distributed Network Security through Log Analysis

Danamma Awati[1]*, Prof. Rajashree Shettar[2]

[1,2] Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bengaluru, Karnataka, India

*Corresponding Author

Abstract

In view of the increasing usage of networks, lot of security issues arise, the management of such issues is becoming tedious. To resolve such issues, network security log analysis becomes main objective. There are various log analysis algorithms and approaches available, but this paper proposes the design of managing and analyzing network security log analysis based on ELK (Elasticsearch Logstash and Kibana). ELK is mainly used to manage and analyze large amount of network logs.

A thorough analysis of information extracted from the logs, helps to identify the issues, to be aware of problems and vulnerabilities. By using ELK, the amount of time taken for analyzing the log files is reduced compared to manual methods.

Keywords: ELK (Elasticsearch Logstash and Kibana), Log analysis, Network security threats and vulnerabilities, SVM (Support vector machine), PCA (Principle component analysis).

Subject classification: Data Mining, Machine Learning

.

1. Introduction

Security has become a major concern, due to wide improvements and advances of the Internet. Collecting the logs accurately and efficiently is one of the major concerns for security purpose. With the increase in number of available tools, analysis of such large number of logs has simplified the efforts. Use of ELK for log analysis has reduced the time consumed compared to traditional methods of log analysis. The ELK stack is the technology which solves the problems of manual or traditional log analysis system. All the log data are distributed in each cluster node, which has good horizontal expansion capabilities. To handle millions of log data along with ELK, machine learning algorithms like PCA and one class-SVM to carry out the log analysis has been experimented.

Logging has become one of the major and essential elements of any networking application or any system. It helps to track about ongoing activities and events in the cyber world that helps to resolve issues related to any faults or any failures. Log analysis is the way of analyzing the computer-generated data to identify various networking activities. Most of the businesses and organizations are required to do log analysis because of security concerns and for compliances purpose. Log analysis will not only help for monitoring, but it also helps to measure the capabilities and weaknesses of the system. This paper concentrates on the internal threats and tries to cover all such kind of threats with the help of ELK stack.

Lot of research work has been carried out earlier in this regard. Various tools have been developed in the recent years. Before 2014, Facebook's Scribe, Cloudera's Flume and LinkedIn's Kafka, etc. have widely been used. Since 2014 Splunk's log system called Splunk has been widely used and has also becomes quiet famous for log system analysis. But, there are some disadvantages in Splunk log system. Basically, Splunk is economically very high to popularize the technology and product. ELK stack is open source compared to Splunk. ELK [2] stack consists of tools such as Elastic search, Logstash and Kibana. ELK has been widely used over the world and has become a very useful and recommended log system for many companies. Some machine learning algorithms like Support Vector machine (SVM) and Principle component analysis (PCA) have major contribution towards the log analysis system [3]. This research paper also compares the various parameters of algorithms and it explains the usage of ELK and the usage of log analysis system.

2. Existing Solution

There are various tools available in the market to handle log analysis and such tools or methods has their own importance and provide the better results. Splunk [4] is also one of the valuable technologies used by various organization to be monitor and analyzing the data. The drawbacks of Splunk logging system is basically it offers limited amount of usage and the cost varies based on the amount of usages. Splunk can collect, analyze, store, search and visualize the logs. Apart from financial issue, product should fulfill the need in a big amount and should be flexible. Hence, to over come the drawbacks of existing log system, this paper discusses the new technology for log analysis. The new log analysis mechanism is referred as ELK logging analysis system.

3. System Architecture

The system architecture for the new log analysis is as shown in figure 1. The architecture shows how the various log files flow from different network services and how it is gathered into Logstash.

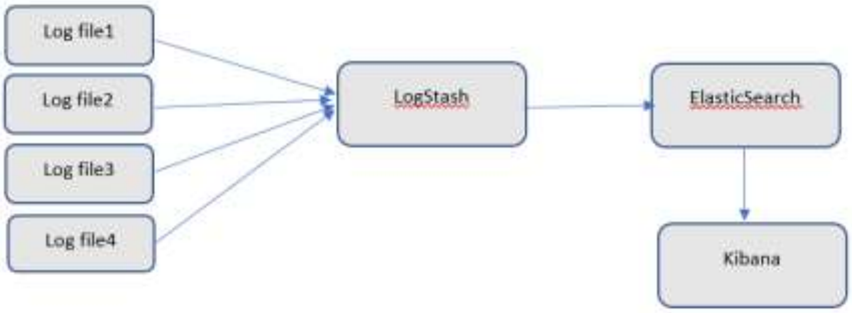

Figure 1: ELK architecture

From Logstash, logs are filtered based on the various parameters such as log levels, info, debug, error and warning. This output is forwarded into Elasticsearch. Elasticsearch indexes the logs and this is fed into Kibana where it will display the complete log analysis report. The log analysis report is detailed and provides various charts such as pi-chart, bar chart, etc. which helps analyze the data and take suitable action.

4. Proposed System

The proposed method consists of setting up of at least two virtual machines [6]. One of the virtual machines is used for log analysis work and another virtual machine is used as backup. ELK will form the cluster where, Logstash is mainly involved in the filtering of logs based on various parameters like log levels and priorities, etc. Once the logs are filtered and fed into Elastic search for indexing. The indexed data is fed into Kibana which displays the log analysis report. Basically, ELK is mainly used to analyses the log files and reduce the usage of unnecessary logs. The test of the log analysis system focuses on the accuracy and effectiveness of the log files by combined machine learning algorithms like PCA and one class-SVM. The log file data is imported, and this data is used by PCA and one class-SVM for log file analysis. This reduces the time and provides effective log analysis results. By comparing the results of two algorithm we got to know that one-class SVM will provide the better results than PCA [7] in log analysis system. The results are compared based on amount of time consumed by the algorithm to provide log analysis reports and by finding success rate which is based on number of false logs found. The lesser the false logs, better are the results.

5. Experimental setup and Results

The ELK stack can be deployed on a different operating system with various manner. Deploying ELK stack on Docker is less cumbersome and hence docker setup for ELK stack has been setup. The docker container is deployed and on top on docker container the ELK stack (Elasticsearch Logstash and Kibana) is deployed by enabling the required port for ELK. During the setup, make sure the ELK server is coming up with the available ports.

Docker container is used to extract the different logs hosting by Logstash. Logstash will filter the logs better and forwards the filtered logs to Elasticsearch for indexing. Once filtering and indexing is done, then Kibana is used to analyze and visualizes the data in various forms. There are other open sources such as Fluentd or Filebeat, to forward the logs to Elasticsearch.

The dataset taken for experimentation is webserver logs dataset from Kaggle.com which is in .csv format. The data set is basically explaining about web server visitors' interest logs. The data collected from the web server represents visitors' interests in different areas through web. Data collected consists of essential fields like IP address, Country, language and area of interest as a parameter and extracted those fields and parsed such data for experimentation. With the help of Logstash, dataset is filtered and forwarded to Elasticsearch for indexing. Before doing this, the logstash.conf file has to be setup, which gives the input, path and other parameter for Kibana to visualize the logs. Kibana is basically UI (User interface) where we can visualize the logs in the required pattern like bar charts or pie charts etc.

The analysis of logs is carried out using Principle Component Analysis (PCA) and Support Vector Machine (SVM) algorithms [7]. Log data set contains lot of duplicate logs entries i.e. similar logs, hence to reduce such duplication, PCA technique. The logs are analyzed for the amount of time consumed, accuracy and percentage of false rate when SVM, PCA and SVM, PCA and one-class SVM algorithms are executed. Table 1 shows the results of execution of these algorithms. Table 1 shows the results obtained for log analysis using PCA and one-class SVM is best when compared to using only SVM or using PCA and SVM techniques. The experiment is carried out by considering fields such as IP address, source port and log type from the log dataset. By considering IP address field and port, the authenticity of the user could be checked i.e. whether the user is a valid user or not. To check whether the user IP address is an authorized IP address or not, machine learning algorithms are applied on the dataset.

Table 1 Comparison results of various algorithms

Type	**Parameters**	**Result**
SVM	Time consumption (%)	3.61
	Accuracy (%)	87.01
	False rate (%)	1.28
PCA & SVM	Time consumption (%)	1.92
	Accuracy (%)	92.012
	False rate (%)	0.95
PCA & one-class SVM	Time consumption (%)	1.22
	Accuracy (%)	98.16
	False rate (%)	0.51

The resulting table 2 indicates:

- Number of Entries in the log table with similar interest.
- False entries indicates in each interest, number of anomalous IPs.
- Execution time indicates the amount of time taken in identifying anomalous IPs in the log entries
- Accuracy obtained after applying machine learning algorithm

Table 2 Algorithm Results

Log File numbers	No of Entries in the log files	False Entries	Execution time	Accuracy
1	200	10	0.16	93
2	300	20	0.32	97
3	500	17	0.53	100
4	700	15	0.34	87
5	1000	12	0.28	96

With the consideration of results one-class SVM will gives the better results compared to others. The time required for calculating of log dataset and the percentage rate of detecting false rate gives the performance of the algorithm.

Following bar graph in figure 1 gives the representation of performance parameters of various algorithms.

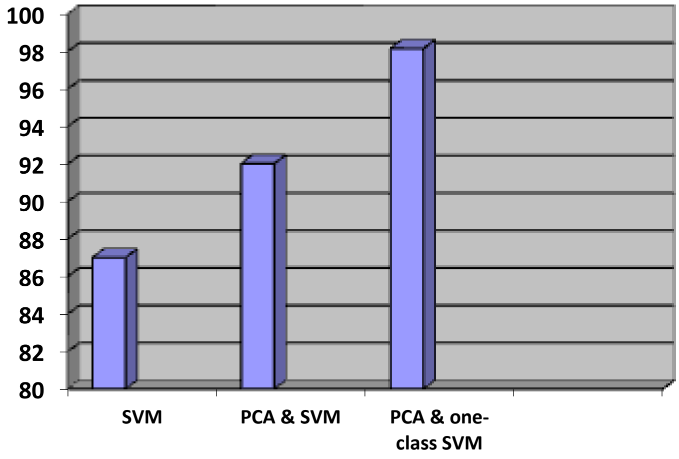

Figure 1 Representation of accuracy for different algorithms

The main objective is to manage the network security log management analysis system with application of machine learning algorithms SVM and PCA and with the help of ELK. The limitation of traditional log system with unstructured logs is solved.

5. Conclusion

The main objective of research work is to handle various network logging problems with the help of latest technologies like ELK and reducing the time consumption for log analysis by using machine learning algorithms. ELK stack is not only overcoming the shortcomings of legacy logging analysis system it also provides the better monitoring results of log analysis. With the help of PCA and one-class SVM the performance of log analysis has improved.

REFERENCES

[1]. Ibrahim Yahya Mohammed AL-Mahbashi, Mr. Prashant Chauhan, & Miss. Shivi Shukla. (2016). Review on Efficient Log Analysis to Evaluate Multiple Honeypots using ELK. Internation Journal Of Advance Research And Innovative Ideas In Education, 2(6), 492-504.

[2]. Chenlin Rao. Authority guide of ELK stack [M]. Beijing: Machinery Industry Press, 2015: 70-335.

[3]. Experience Report: System Log Analysis for Anomaly Detection. In Software Reliability Engineering (ISSRE), 2016 IEEE 27th International Symposium on,pp.207-218.

[4]. Carasso, D. (2012). Exploring splunk. published by CITO Research, New York, USA, ISBN, 978-0.

[5]. P. Winter, E. Hermann and M. Zeilinger. Inductive intrusion detection in flowbased network data using one-class support vector machines [A]. 4th IFIP International Conference on New Technologies, Mobility and Security Piscataway: IEEE Press, 2011: 1-5.

[6]. M., Delibaş, E., Karanlık, B., İnal, A., &Aytekin, T. (2016, April). Real time distributed analysis of MPLS network logs for anomaly detection. In Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP, pp. 750-753.

[7]. Yannan Li, Weihong Jiang, Qun Zhao. Host log analysis and research [J]. China High-tech Enterprises, 2010, 139 (4): 193-194.

[8]. Sharma, V. (2016), Getting Started with Logstash. In Beginning Elastic Stack, pp. 1-15, Apress.