International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Traffic Statistics Determination of Unified Threat Manager to Provide Threat Perception

Neelutpol Gogoi[1], M. N. Sushmitha[2]

[1]M.Tech, Department of Computer Science and Engineering, Hindustan University, Chennai-603103, India

[2]Assistant Professor, Department of Computer Science and Engineering, Hindustan University, Chennai-603103, India

Abstract*: In the recent time with the increasing rate of data, the security threats on these data have also increased. So, the volume of alerts generated by unified threat manager (UTM) becomes very large. Using conventional methods to analyze a lot of data would drag down the system performance. This paper is a proposal to a system, which will take the UTM logs as the input and analyze these logs using Hadoop Map-Reduce programming mode. Thereby, presenting some threat perceptions by determining the traffic statistics and produce a summarized alarm reports for discovering, predicting and stopping the security threats in the system at a comparatively faster rate. The system would analyze the logs directly at the time of some alerts occurrence and also historical data (logs) from the firewall in order to search for attack signatures (anomaly and patterns) for predicting attacks.*

Keywords: Data security, Hadoop, Map-Reduce, Traffic Analysis, Unified threat manager, Zero-day attack

1. Introduction

Security of data is a major concern in the present time. With the increasing complexity of computer system, databases and the networks around the globe, the attacks on these systems have also grown very complex. The use of the traditional antivirus, firewall etc individually could no longer provide proper security to our valuable data. So, unified threat management (UTM) was being introduced, a complex system which merges antivirus, firewalls, IDS, VPN etc that will together provide protection to the system and data.

But every system has vulnerabilities. Even UTM cannot provide full security from the new and complex types of attacks. The proper analysis of the UTM logs will help us to understand and recognize these new and complex threats. Data mining is a powerful technology with great potential to help companies focus on the most important information in the data they have collected. Data mining (also called as data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing the results into useful information - information that can be used to identify patterns or signatures. It allows users to analyze data, categorize it, and summarize the identified relationships. In technical terms, data mining is the process of finding correlations (or patterns) among dozens of fields in large relational database. There are several traditional techniques and software that were being used for log analysis. But, with the ever increasing number of threats and attacks, the volume of logs generated by the various components of the UTM becomes very large. The use of the conventional methods to analyze this huge volume of logs has several problems. *Firstly*, it will drag down the system performance. *Secondly*, it is easy to ignore the crucial information in large amount of alerts. Moreover, the conventional methods are time consuming and also costly.

2. Background and System Overview

Data is one of the most valuable terms in today's world and security of this data is a major concern. Various types of measures like antivirus, firewall, intrusion detection system, unified threat manager etc are being used for providing security to the data. With the increasing complexity of the security threats, these traditional approaches could no longer provide proper security. So, analysis of the logs from these systems help companies focus on the most important information related to security in the data they have collected. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to identify patterns or signatures for providing threat perceptions by determining the traffic statistics. Unified threat manager (or UTM) is an evolution of the traditional firewall into an all-inclusive security product which is able to perform multiple security functions with one single appliance like;

- Network firewalling
- Network intrusion prevention
- Gateway anti-spam
- VPN, content filtering, load balancing
- Data leakage protection and
- On-appliance reporting

But, since the volume of logs generated is massive and are of both structured and unstructured data, it is easy to ignore some crucial information. Thus it becomes very difficult to process using traditional database and software techniques. Moreover, the conventional methods that are being used are not time efficient and also are very expensive. The proposed system efficiently uses Hadoop for analysis of this massive volume of logs. Hadoop [10] makes it possible to run applications on systems with thousands of nodes involving thousands of terabytes of data. The distributed file system in Hadoop facilitates rapid data transfer rates among the nodes and thereby allows the system to continue operating uninterrupted in case of a node failure. Even when a significant number of nodes become inoperative, this approach lowers the risk of catastrophic system failure. Hadoop was inspired by Google's MapReduce, which is a software framework where an application is broken down into numerous small parts. Any of these parts (also called fragments or blocks) can be run on any node in the cluster.

Hadoop kernel, the Hadoop distributed file system (HDFS) [24], MapReduce and a number of related units such as Apache Hive, HBase and Zookeeper are present in the current Apache Hadoop ecosystem.

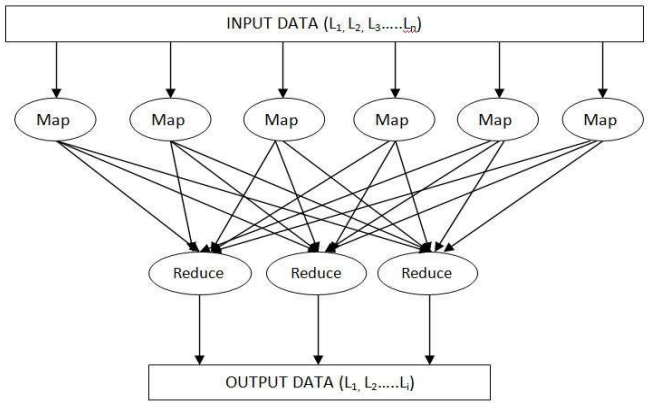


Figure 1: Map-Reduce architecture

Analysis of UTM logs using Hadoop (which is an open source framework running applications on large clusters built of commodity hardware) to present some threat perceptions by determining the traffic statistics and produce summarized alarm reports for discovering, predicting and stopping the security threats in the system in a relatively small amount of time (within an hour or two) is in the scope of this paper. This system could be installed in an extra computer system connected directly to the UTM [12] or firewall. This will facilitate in getting the logs directly at the time of some alerts, analyze those as-well-as historical data (logs) in order to search for attack signatures for predicting attacks and produce reports that could be sent to the administrator via email or SMS.

3. System Design

Security of data is one of the major concerns in the present time. With the increasing complexity of computer systems, databases and the networks around the globe, the attacks on these systems have also grown very complex. The use of the traditional antivirus, firewall etc individually or even unified threat management (UTM) a complex system which merges antivirus, firewalls, IDS, VPN etc cannot provide total protection from the new and complex types of attacks to the system and data. Every system has vulnerabilities. However, for organizations with UTM or multiple firewalls effectively monitoring and responding to the high volumes of alerts can be a very complex and time consuming process. As discussed earlier, there exist a lot of different approaches for that are being used for analyzing the logs. But, in the situations when the volume of logs is huge, the effectiveness of these approaches diminishes.

UTM generates a large volume of alerts, where many of them are either false positives or of low importance. Regular monitoring of UTM logs can reveal a wealth of information about threats to an organization. In the event of a compromise the logs play a critical role in evaluating the extent of the attack, and as evidence against the attacker. This makes it hard for the human to spot alerts which need more attention. Hence, the Traffic Statistic Determiner (TSD) can be used in order to effectively analyze these logs.

TSD is a log analysis system especially for big data analysis (huge log files) based on Hadoop. The system architecture of TSD is shown in figure 2, which provides a more effective and efficient analysis technique for analyzing large log files. The UTM logs (L) are composed of the different fields like <date> <time> <devname> <devid> <logid> <type> <eventtype> <level> <policyid> <sessionid> <srcip> <srcport> <dstip> <dstport> <dstintf> <service> <hostname> <profile> <status> <reqtype> <sentbyte> <rcvdbyte> <msg> <method> <cat> <catdesc> etc. So, in this technique, from each line of log only certain fields are considered for analysis whereas the others are filtered out. This reduces the volume of logs and thereby increases the speed of analysis.

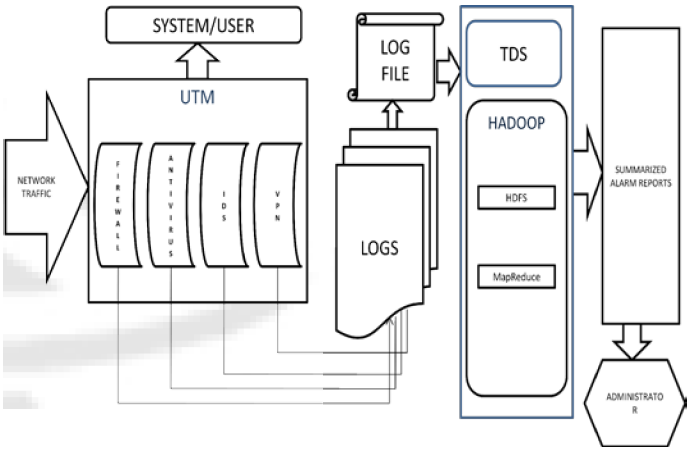


Figure 2: TSD system architecture

If L_n be a line of log, then
$L_n = (L_{date}, L_{time}, L_{devname}, L_{devid}, L_{logid}, L_{type}, L_{subtype}, L_{eventtype}, L_{level}, L_{vd}, L_{policyid}, L_{identidx}, L_{sessionid}, L_{srcip}, L_{srcport}, L_{srcintf}, L_{dstip}, L_{dstport}, L_{dstintf}, L_{service}, L_{hostname}, L_{profiletype}, L_{profile}, L_{status}, L_{reqtype}, L_{sentbyte}, L_{rcvdbyte}, L_{msg}, L_{method}, L_{cat}, L_{catdesc})$.

After filtering out some of the less important fields, a reduced log (L_R) is obtained,
$L_R = (L_{level}, L_{srcip}, L_{srcport}, L_{dstip}, L_{dstport}, L_{status}, L_{msg},)$.

For L_R,

Table 1: Results after filtering out some less important fields

Level	srcip	Srcport	Dstip	Status	dstport	Msg
Lvl 1	*S Ip1*	*Sp i*	*D ip1*	*Stat a*	*Dstp q*	*xyz*
Lvl 3	*S Ip2*	*Sp j*	*D ip2*	*Stat c*	*Dstp r*	pqr
Lvl 2	*S Ip3*	*Sp i*	*Dip5*	*Stat a*	*Dstp s*	cde
Lvl 1	*S Ip4*	*Sp k*	*Dip3*	*Stat f*	*Dstp t*	xy

These reduced logs as shown in Table.1 are then analyzed taking any L_{status} as a filter which will result in a further reduced log file as shown in Table.2.

Table 2: Results after taking L_{status} for filtering

Level	srcip	Srcport	Dstip	Status	dstport	Msg
Lvl 1	*S Ip1*	*Sp i*	*D ip1*	*Stat a*	*Dstp q*	*xyz*
Lvl 2	*S Ip3*	*Sp i*	*Dip5*	*Stat a*	*Dstp s*	cde

Finally, the records are displayed taking I_{dstip} as keyword for

listing out all the logs from different sources to that ip along with the number of times if redundant. Thus, TSD produces the analytical report that can be easily examined by the administrator to percept threats.

4. Related Works

Analyzing logs with the existing system is a very time consuming approach and is inefficient (the analysis which is not completed in time is not useful or invalid). There exist a lot of different approaches for this purpose. The data mining concepts like Frequent Itemset Mining and data clustering [3], Building IDS Log Analysis System on Novel Grid Computing Architecture [1], Analysis Console for Intrusion database, Clustering Event Logs Using Iterative Partitioning [5], Analyzing cluster log files using Logsurfer [16], Stateful inspection of logs [7], association mining of the logs, event correlation etc are being used for effectively analyze the logs. However, these existing methods prove to be inefficient and time consuming while it comes to big data. In contrast to the existing methods cited above this system proves to be better for the analysis of huge UTM log files.

5. Implementation and Results

The output of the proposed system is a reduced and simplified record file. The final output was verified and an instance of it is shown in following figure

```
dstip=117.121.249.254
level=warning srcip=10.100.9.3 srcport=51695 dstport=80 status=blocked msg=URL was blocked because it contained banned word(s)
level=warning srcip=10.100.9.3 srcport=58422 dstport=80 status=blocked msg=URL was blocked because it contained banned word(s)
level=warning srcip=10.100.9.3 srcport=51663 dstport=80 status=blocked msg=URL belongs to a denied category in policy
```

Figure 3: TSD output

6. Conclusion

To summarize, security of data is a major concern in the present time. With the increasing complexity of computer system, databases and the networks around the globe, the attacks on these systems have also grown very complex. The use of the traditional antivirus, firewall etc individually or even unified threat management (UTM) a complex system which merges antivirus, firewalls, IDS, VPN etc cannot provide total protection from the new and complex types of attacks to the system and data. Every system has vulnerabilities. Regular monitoring of UTM logs can reveal a wealth of information about threats to an organization. In the event of compromise the logs can be used in evaluating the extent of the attack, and as evidence against the attacker. UTM generates large amounts of alerts, with most of them being either of low importance or false positives. This makes it hard for the human to spot alerts which need more attention. Hence, hadoop is being used to effectively analyze these logs for a better threat perception.

7. Future Scope

In the present scenario the input is taken as a .txt file from UTM, but this can be extended for different other systems logs and input formats. The TSD algorithm could be optimized for better performance and the output report needs to be integrated with more elements, such as attack verifications, suggestion approaches and so on.

References

[1] Wei-Yu Chen, Wen-Chieh Kuo, Yao-Tsung Wang *"Building IDS Log Analysis System on Novel Grid Computing Architecture"*, National Center for High-Performance Computing, Taiwan{ waue ,rock ,jazz}@nchc.org.tw
[2] AviKak (kak@purdue.edu), Avinash K *"Lecture23: Port and Vulnerability Scanning, Packet Sniffing, Intrusion Detection, and Penetration Testing"* Lecture Notes on"Computer and Network Security" April2, 2013, Purdue University.
[3] Risto Vaarandi and Krlis Podiš, *"Network IDS Alert Classification with Frequent Itemset Mining and Data Clustering"* 2010.
[4] Ziming Zheng, Zhiling Lan Byung H.Park*"System Log Preprocessing to Improve Failure Prediction"* Illinois Institute of Technology {zzheng11,lan}@iit.edu , Al Geist Oak Ridge National Laboratory{parkbh,gst}@ornl.gov
[5] Adetokunbo Makanju, A.NurZincir-Heywood, EvangelosE.Milios, "Clustering Event Logs Using Iterative Partitioning" 2010.
[6] Anukool Lakhina *"Diagnosing Network-Wide Traffic Anomalies"* Dept.of Computer Science, Boston University anukool@cs.bu.edu
[7] Cristiano Lincoln Mattos (lincoln@cesar.org.br), Evandro Curvelo Hora(evandro@cesar.org.br), Fabio Silva (fabio@cesar.org.br), Marco Antonio Carnut (kiko@cesar.org.br)*"Improving stateful inspection log analysis"*
[8] http://www.secureworks.com/resources/arti cles/ other_articles/firewall-primer/
[9] https://www.usenix.org/legacy/event/osdi04 /tech/full_papers/dean
[10] http://www.cloudera.com/content/cloudera /en/ why-cloudera/hadoop-and-big-data.html
[11] http://strata.oreilly.com/2011/01/what-is-hadoop .html
[12] http://en.wikipedia.org/wiki/Unified_threat_ management
[13] http://www.cs.northwestern.edu/~ychen/classes /msit458-s09/UTMMSSCloud_KurtisEMinder
[14] http://cuddletech.com/blog/?p=795
[15] http://www.ibm.com/developerworks/ library/ os-log-process-hadoop/
[16] James E.Prewett *"Analyzing cluster log files using Logsurfer"*
[17] D. Andersson, M. Fong, and A. Valdes, *"Heterogeneous Sensor Correlation: A Case Study of Live Traffic Analysis*" Jun. 2002, Third Ann. IEEE Information Assurance Workshop.
[18] M. Attig and J. Lockwood, *"A Framework for Rule Processing in Reconfigurable Network Systems"*, Apr. 2005, 13th Annual IEEE Symposium.

[19] F. Chang, J. Dean and S. Ghemawat, "*Bigtable: A Distributed Storage System for Structured Data*", OSDI 2006.
[20] D. E. Denning, "*An Intrusion Detection Model*" Feb. 1987 IEEE Transaction on Software Engine.
[21] J. Dean and S. Ghemawat, "*MapReduce: Simplified Data Processing on Large Clusters*" OSDI 2004, Dec. 2004.
[22] EC2, "*Amazon Elastic Compute Cloud*" http://www.amazon.com/gp/browse.html?node=201590011, 2008
[23] S. Ghemawat, H. Gobioff and S. Leung, "*The Google File System"* SOSP 2003, Dec. 2003.
[24] D. Borthakur, "*The Hadoop Distributed File System*" http://lucene.apache.org/hadoop, 2008.

Author Profile

Neelutpol Gogoi received his B.Tech (Computer Science and Engineering) from North Eastern Hill University, Meghalaya, India and undergoing M.Tech (Computer Science and Engineering) in Hindustan University, Chennai, Tamil Nadu, India. His area of interest is data-mining, Computer Security and Forensics.

Ms. M.N. Sushmitha received her B.E (Computer Science and Engineering) in 2003 and M.Tech (Computer Science and Engineering) in 2005. Currently, she is working as Assistant Professor in the department of Computer Science, Hindustan Institute of Technology and Science, Chennai, India.