International Journal of Security and Its Applications
Vol. 10, No. 1 (2016), pp.13-20
http://dx.doi.org/10.14257/ijsia.2016.10.1.02

A Study on the Big Data Log Analysis for Security

Kyung-Sik Jeon[1], Se-Jeong Park[2], Sam-Hyun Chun[3] and Jong-Bae Kim[4*]

[1]*Department of IT Policy and Management, Graduate School of Soongsil University, Seoul, 156-743, Korea*
[2, 4*]*Graduate School of Software, Soongsil University, Seoul, 156-743, Korea*
[3]*Department of Law, Soongsil University, Seoul, 156-743, Korea*
[1]*ksjeon@inovacnc.co.kr*, [2]*sejung90@naver.com*, [3]*shchun@ssu.ac.kr*,
[4*]*kjb123@ssu.ac.kr*

Abstract

Recently, cyber-attack has become the serious national treat such as shut down industry control system, and an act of war. Therefore, the issue is suggested about the necessity of Enterprise Security Management (ESM) that is for integrated management of network system such as firewall, IPS, VPN, and etc. However, current ESM has the limit of blocking only cyber-attack from outside due to using the networking attack detection method that monitoring the traffic inflows from outside to inside. Therefore, this study suggests the new security log system using big data that enhances intelligence of security by analyzing the relationship between security and data events created from network, system, application service of main IT infrastructure. We assume to apply the distribution-based saving/processing technology through the security log system using big data which is suggested in the study. Moreover a ripple effect of enhanced customer service satisfaction due to the possibility of inflow and infection (spread) of malignant code in-house and real time monitoring.

Keywords: *Big Data, Log, Security, Cyber-attack, analysis*

1. Introduction

The A.P.T attack, which recently raises a significant social chaos by new malignant code, is one of the types of targeting attack in the past. It is difficult to detect and block because it achieves the goal by using whole possible methods after securing information of the target.

In the past, cyber-attack was mainly a random attack type. However, recently hackers choose systematical and long-term attack type by cooperating to find weakness of the target. Therefore the company damaged by APT attack has been increased such as Hyundai Capital, Auction, SK communications, and Nonghyup.

It is crucial to have ESM (Enterprise Security Management) system that manages integrated in-house network secure systems such as firewall, IPS, and VPN to prevent outflow of company's assets by the treats of intelligent A.P.T attack. Current ESM collects logs and saves it in database system (RDBMS, same as DB), and then shows present condition on a dash board after analyzing the saved data. At the end it alarms to a manager when there is a problem. Unfortunately, current secure system only blocks cyber-attack from outside because it uses network-based attack detective method which only monitors traffic inflow from outside to inside. It is the reason why it shows the weakness to the method of direct attack to in-house

* Corresponding author. Tel. : +82-10-9027-3148.
Email address: kjb123@ssu.ac.kr(Jong-Bae Kim).

ISSN: 1738-9976 IJSIA

user PC client. It only collects present log condition of systems and present event condition individually. It causes to show lack of application detection, recognizing unknown attacks by users and correlation analysis between attacks.

Therefore, we suggest the security log system using Big data that enhances intelligence of security, distinguishes the manager and the attacker who steals right of manager by analyzing data of application and security event rather than the form of attack [7, 8]. There is a total analyzing system showed in Table 1.

Table 1. Total Analyzing System

<table>
<tr><th>ESM log collection</th><th>Collection type</th><th>Details</th><th>Implication</th></tr>
<tr><td>Firewall / VPN / IDS / VMS / IPS / Web Firewall</td><td>Firewall</td><td>Institution code, departure IP/Port, destination IP/Port, protocol, Action, name of attack, times, etc.</td><td rowspan="8">1. Detection of IP, Port centered network class, detection and analyzing mainly known attacks
2. Link the centers of IP/Port when there is the correlation between network and terminal security system centered, analyzing the range with in short time(maximum one day)
3.Analysing mainly harmful IP, detecting and acting known attacks</td></tr>
<tr><td>Forgery log collection</td><td rowspan="3">IDS / IPS / Web Firewall / TMS</td><td rowspan="3">Institution code, departure IP/Port, destination IP/Port, protocol, Action, name of attack (URL), CVE-ID, times, etc.</td></tr>
<tr><td>Forgery detection / Obstacle detection</td></tr>
<tr><td>TMS log collection</td></tr>
<tr><td rowspan="4">TMS / total detection log of signature based</td><td>VPN</td><td>Institution code, departure IP/Port, destination IP/Port, protocol, Action, name of attack, times, etc.</td></tr>
<tr><td>VMS</td><td>Institution code, departure IP, virus name, infection times, etc.</td></tr>
<tr><td>WMS</td><td>Institution code, name of order (forgery/Falsify/obstacle/URL) etc.</td></tr>
<tr><td>Others</td><td>Possible to extend by types</td></tr>
</table>

In the research, I would like to collect and analyze rapidly the large amount of structured/unstructured data by distribution-based saving/processing technology through security log system using Big data. Especially suggesting real-time analysis performance by distribution-based multi-searching, presenting flexible dashboard for visible analysis, and applying correlation analysis technique of all elements. I assume that the integrated security control will be possible by securing availability and expandability if applying security log system using Big data.

2. Related Works

ESM (Enterprise Security Management) is for enhancing security and effectiveness of security management by integrating security control, management, and maintaining a consistent policy of organization. It often refers to a remote security management. Previous ESM related research is not only about firewall, IDS, VPN, and diverse security products but also the electronic invasion by connecting network equipment in real time. The studies use IT unit system log integrated, error process monitoring, and connected analysis of IT system and security system for security control. However, the studies are having difficulty to prevent recent APT attack due to a problem of performance and expenses of database (DB) which stores a large amount of data [1].

SIEM includes function of ESM, it is the system model which is able to analyze unknown security threat correlation by using the connected application as a total analysis. Which means it makes possible to control integrated security by collecting a large amount of structured/unstructured data [3, 4, 6]. There are characteristics of SIEM in Table 2 [2].

Table 2. Characteristic of SIEM

Category	Current SIEM	Future of SIEM
Collection	1. Possible to obtain and collect logs and event data automatically from diverse server 2. Impossible to process huge amount of security information	Provide the architecture of distributed data
Data	1. Create the Overall storage for security data 2. The data included in limited log collection	1. Collect network traffic 2. Restore session for detecting and investigating about infiltration and attack method of an attacker 3. Offer visualization of threat after collecting threat information from an external source automatically
Investigation	1. Integrate the log data to create the integrated storage of main security data. 2. Spend a lot of time to low-level of usefulness and event detection	Offer the internal user interface to complement the method of detection of security analyst
Analyzing	1. Offer the high level of control report 2. Do not raise the control class of security threat.	As a result of programs focused in security, show the evidence of compliance
Prediction	1. Offer warning alarm by grasping the current accident 2. Detection of attack rely on recognizing the way of attack previously or attack authentication	1. Construct integrated platform to collect security data from diverse environment 2. Perceive attacker when suspicious activity is captured 3. It is reduced that the number of accident analyst needed for distinguish and investigate threats

To make integrated security control possible, new generation security information analysis technique should enhance intelligence of security by analyzing the correlation between security events and data created from network of IT based important facilities, system, and applications. It is expected that "SIEM(Security Information & Event management)" system will be settled as the crucial technique of intelligent security by integration with Big data analysis technique, as the new generation security information analysis technique [2, 4, 5].

3. Architecture of Security Log System Using Big Data

3.1. Intelligent Information Analyzing Platform

Intelligent information analysis platform is composed of collecting, saving, analyzing, and optimized element of data. Each function constructs different formats. They collect data stably from diverse data source, save the data equally by multiple parallel structures, offer the system structure that is able to analyze intelligently based on high-speed searching. There is the intelligent information analysis platform in Table 3.

Table 3. Total Analyzing System

<table>
<tr><th>Collection object</th><th rowspan="2">Integrate the individual data processing resource</th><th>Intelligent information analyzing platform</th><th>Working area</th></tr>
<tr><td rowspan="2">- Network / security equipment
- Application
- Server
- User
- Other log</td><td>- Distributed file system
- Distributed DBMS
- Search engine
- Analyzing engine
=> Real time data process down
- Secure ability to process</td><td rowspan="2">- Total management
- Obstacle reaction
- Security control
- Inspection
- Report</td></tr>
<tr><td>Total monitoring of united management
- System manager
- Network manager
- Inspection
- Security manager
- Executive manager</td></tr>
</table>

3.2. Suggesting the Algorithm of Collecting Massive Data

Develop data collecting structure in considering all data collecting technique, massive data transmission, management stability, and high availability for collecting and engaging data. All of the information created in the security equipment is saved in real-time in collector through data transmitter such as sources, format data, structured/unstructured original log, and original log. The current data collecting process shows vulnerability in unusual traffic analysis while processing Web log in main homepages, WEB/WAS/DB management log, Net-Flow statistic information, detective event of DNS sinkhole, local network communication record, and detective event of Web symptom. Suggested integrated collector can analyze unusual traffics, detect web symptom, block harmful web site through DNS sinkhole, detect web hacking early by using Web Shell. It is also suggested to use two methods, agent/agentless for collecting information, and add flexibility to selection of collection methods by considering real-time and stability. Data transmitter automatically disperses error and load of data, and prevents loss of data by using automatic load distribution, detect error/repeat, and log forwarding technique. There is the algorithm used for data collecting through integrated data collector, which is suggested in the study, in Figure 1.

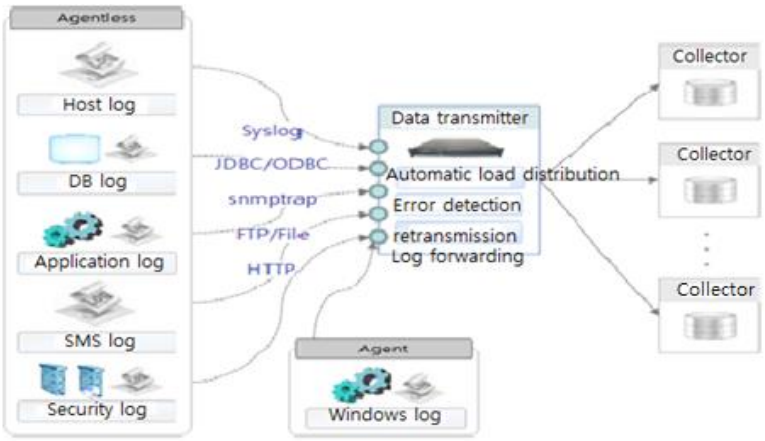

Figure 1. Data Collection Algorithm

A large amount of information is transmitted from the system log adaptor which is constructed by interlocked adaptor Syslog/SNMP, system performance adaptor with UNIX/Windows performance, UNIX Syslog, and Windows WMI by applying massive processing technology of UDP packet and data forwarding technology. To transmit the collected information through adaptor to transmission queue, it is done by verifying data consistency and adjusting data transmission volume by performance of equipment. When data overflow occur, the data cannot be transmitted and repeated to reserved collector. If error like network severance occurs during the process, save the data temporarily and repeat it through the data engaging adaptor with SSL certification, SSL decoder, SSL capture, and LOG filter functions. Complete the collection by transmitting the original log without modifying the collector when the data arrives at HA transmission module through transmission queue.

3.3. Data Saving

Collector is constructed by distribution-based log servers. There is the saving method of collector server. The data coming through the collecting system is made for clients to find the information initially through receiving and normalizing process, compared with normalized data, and obtain index value interacted with Index DB. The data of security log, system log, and application log is received and it is normalized through normalizing

engine, normalizing file, and data tagging. Use distributed architecture to save the massive security log file. The dispersed architecture is processed in parallel processing to store massive data, and runs saving and real-time indexing work by distribution-based multi indexer. So the Tera byte(TB) data per day can be processed by the distributed architecture, and each collector shows 200,000 EPS process performance. Especially that each collector automatically checks integrity when saving data, and saved data in compressed and encoded folder. The collectors automatically backup and restore by constructing data backup/hot spare collector to protect the original data automatically from possible defect of multi system. Theoretically, this management structure can store infinite data, and have expendability and stability. It also makes faster result than a serial process way by arranging collectors in parallel form causing proportion of number of collector and processing performance. The technique can make significantly huge effect to process the massive security log Big data when data size is small. There is the process of collection in Figure 2.

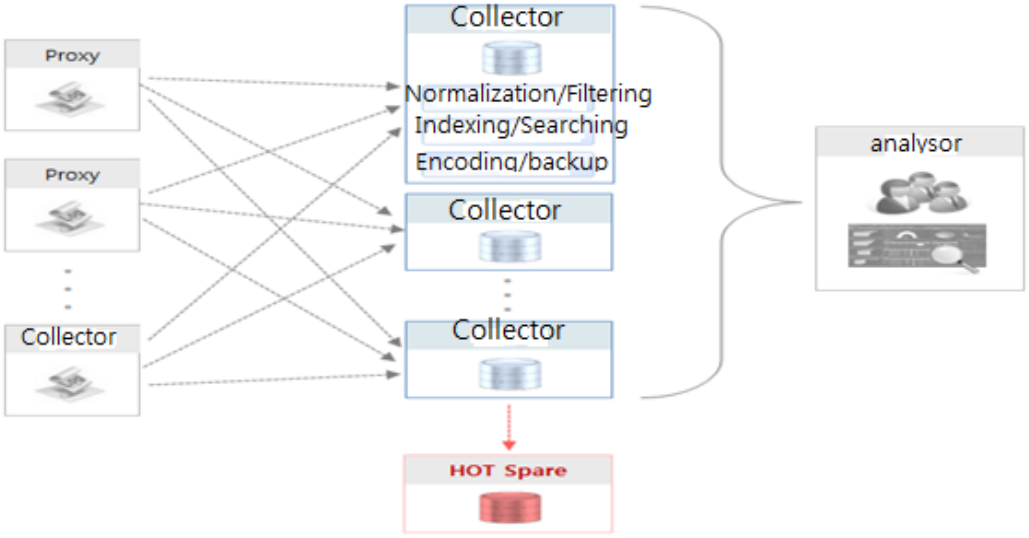

Figure 2. Data Collection Process

3.4. Data Analyzing System

The speed of massive data cannot be guaranteed by checking in real-time. However, it can be operated by finding indexing data by entering keywords or conditions of the indexing data saved in collector. The searched massive security log data analysis makes multi-scanning easier by data drilldown that analyzing data in subdividing problems into small pieces. Also the data from security equipment is guaranteed real-time analysis performance by two types of distribution-based multi-scanning. One of them is to detect rapid changes of data based on baseline and threshold value. Another one is using trending analysis that is to predict data based on statistics.

Analyzing the correlation of all the events and intuitionally display it in diagram form by real-time monitoring in equipment/log type develop diverse dashboard of movement of users for visible analysis of data. Turn an alarm to show threat in visualized form when error is found in real-time monitoring process. Maximum of 2 billion cases of single scanning is ran, and scanning in a minute under the condition of simple scanning condition of 200G~400G per day.

Figure 3 below shows analyzing system in a diagram form.

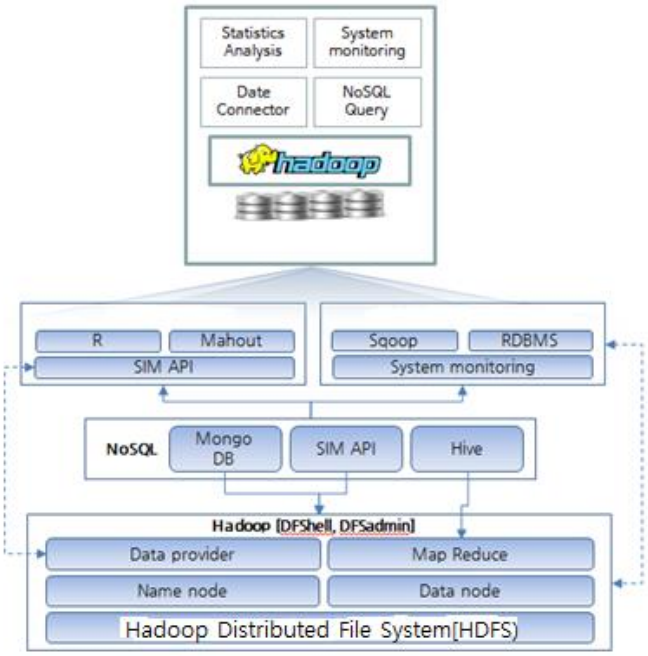

Figure 3. Analyzing System

Creating diverse statistic data by batch using the electronic governmental appointed Big data standard platform 'Hadoop' based on statistical analysis system. Analyzing Big data using Hadoop distributed file system, can plan lower enter cost, fast & flexible development, compatibility, reliability, and stability due to open source that is connected with statistical analysis. It is also considered to interlink with integrated security control system for individual data connection development by supporting NoSQL(Not only SQL) for data analysis & expending flexibility of statistic function, and data mining algorithm for Big data based on statistical analysis.

3.4.1. Log Structure for Data Analysis: The data collecting structure should be built in consideration of all data collecting technique, massive data transmission, management stability, and high availability for data collection and engagement. Therefore all of the source, format data, structured/unstructured original log, and original log from security equipment are collected in real-time to be stored in collector through data transmitter.

The collected log data structure is divided into two formats by level of normalization. First of all, the data is not saved in fixed field as unstructured type. For example, data can be analyzed in text structure such as text documents, images, motion, and audio data. Second of all, it is the structured data which is saved in the fixed field. For example, relative database and spread sheet which are made from modifying unstructured data.

The data is the original data without modifying and transmitted to collector which is collected from each terminal adaptor. IP address, security log, data collecting date and time of adaptor are saved in collector. After that it becomes a normalized data through manufacturing process of visualizing original data by normalizing process of the integrated log server.

Collect all of the trace of data modifying such as user information, data reading, and printing document one by one, and printing the information of the data. It becomes easier to access the data with an error message and is able to show the information of data to users by unifying structure.

3.4.2. Algorithm of Log Analysis: We suggest the method using PCRE (Perl Compatible Regular Expressions) technique, which is a library that supports special Separator technique and regular expression, for unstructured data normalizing technique.

The algorithm is to analyze logs that are divided into log collection server, log parser, and log transition. Log collection server passes collected data through data filter to the log parser by using collecting equipment such as FireWall Event, IDS Event, Traffic Event, and WebFw Event. Log parser works on parsing transformation of filtered data by collected equipment, and standardizes data to the regular data. The transformed data is saved in a file or memory and the log is transmitted. At the same time, complete normalization by checking original data, regular expressed & log transformation, delete, and selecting field value to save XML file. There is a process of normalizing of unstructured data in Figure 4.

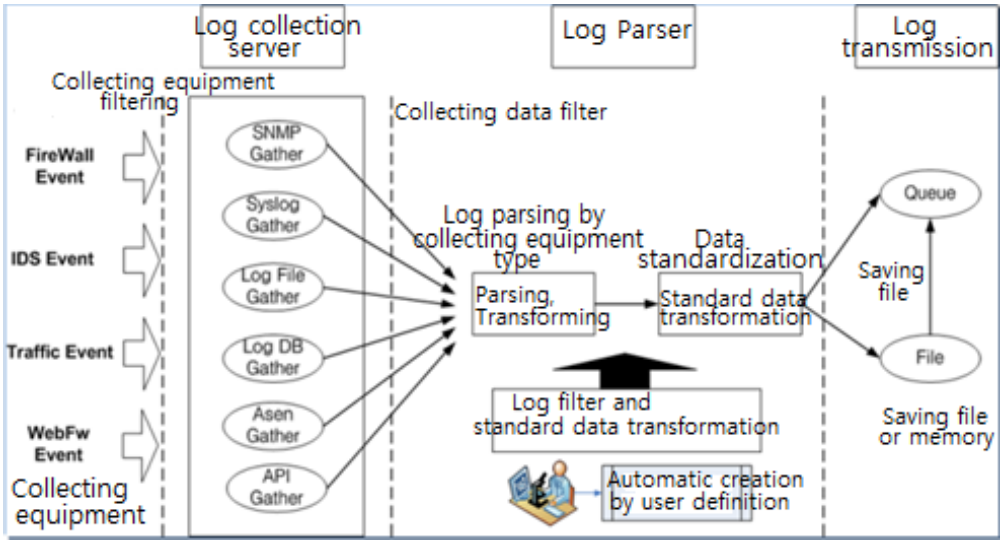

Figure 4. The Atypical Data Normalizing Technology

4. Conclusion

I applied collecting, saving, processing, and analyzing techniques based on intelligent information analysis platform for system construction of security log analysis using Big data. Saving massive data secures availability and expandability of collected security logs by using an Agent/Agentless way. Also, constructing the system becomes possible to analyze which was impossible in the past by supporting fast searching and showing visualized method which is now possible to analyze. Moreover, it is expected to enhance customer service satisfaction by inflow of harmful code in-house, and real-time monitoring becomes possible.

In this research, there is a limitation of selecting a part of security field among diverse big data methodology. The extra study of analysis techniques of Big data analysis area can now be applied to diverse fields such as manufacturing, service, and finance as well as security.

References

[1] L. Seung Ha, K. Seung Won, K. KiHong, P. Sechung, " Design of Big Data ETL Model for Aggregating of Security Log/Event ", KICS, **(2014)**.06
[2] M. Nicolett and K.M. Kavanagh, "Magic Quadrant for Security Information and Event Management," Gartner Group, **(2012)**.05
[3] K. M. Kavanagh, M. Nicolett, O. Rochford, "magic quadrant for security information and event management", Gartner Group, **(2014)**.06
[4] M. Nicolett and J. Feiman, "SIEM Enables Enterprise Security Intelligence," Gartner Group, **(2011)**.01

[5] M. Nicolett and K.M. Kavanagh, "Critical Capabilities for Security Information and Event Management," Gartner Group, **(2012)**.05
[6] N. MacDonald, "Information Security Is Becoming a Big Data Analytics Problem," Gartner Group, **(2012)**.05
[7] J-s Yun, H-s Kang, I-y Moon,"Analysis study of movement patterns using BigData analysis technology", Journal of Information and Communication Convergence Engineering, vol.28, no.5, **(2014)**.
[8] DG Park, SK Kim, "A Design and Implementation of Mobile Application Usage Pattern Analysis system", Journal of Information and Communication Convergence Engineering, vol.18, no.9, **(2014)**.

Authors

Kyung-Sik Jeon received his bachelor`s degree in Management from Seoul Cyber University in Korea, (2013). And he is studying his master's degree in software engineering in Soongsil University, Seoul. Now he is the CEO of INOVA C&C Co., Ltd. since 2014. His research interests focus on Digital Forensics.

Se-Jeong Park received her bachelor's degree in Software Engineering in Kongju National University (2014). And she is studying her master's degree of software engineering in the Graduate School of Software, Soongsil University, Seoul. Her current research interests include Database and Open Source Software.

Sam-Hyoun Chun received his doctor's degree of Law in Goethe University Frankfurt (1992). Now he is a professor in the School of Law, Soongsil University, Seoul, Korea.

Jong-Bae Kim received his bachelor's degree of Business Administration in University of Seoul, Seoul (1995) and master's degree (2002), doctor's degree of Computer Science in Soongsil University, Seoul (2006). Now he is a professor in the Graduate School of Software, Soongsil University, Seoul, Korea. His research interests focus on Software Engineering, and Open Source Software.

