2015

Behavior-based anomaly detection on big data

Hyunjoo Kim
Electronics & Telecommunications Research Institute (ETRI), Daejeon, Korea, Sungkyunkwan University

Jonghyun Kim
Electronics & Telecommunications Research Institute (ETRI), Daejeon, Korea

Ikkyun Kim
Electronics & Telecommunications Research Institute (ETRI), Daejeon, Korea

Tai-myung Chung
Sungkyunkwan University

DOI: 10.4225/75/57b69d1ed938e
This paper was originally presented at The Proceedings of [the] 13th Australian Information Security Management Conference, held from the 30 November – 2 December, 2015 (pp. 73-80), Edith Cowan University Joondalup Campus, Perth, Western Australia.

BEHAVIOR-BASED ANOMALY DETECTION ON BIG DATA

Hyunjoo Kim[1,2], Jonghyun Kim[1], Ikkyun Kim[1] , Tai-myung Chung[2]
[1]Cyber Security Research Laboratory,
Electronics & Telecommunications Research Institute (ETRI), Daejeon, Korea
[2]Computer Engineering Dept., Sungkyunkwan University, Suwon, Korea
e-mail: {hjookim, jhk, ikkim21}@etri.re.kr, tmchung@skku.edu

Abstract
Recently, cyber-targeted attacks such as APT (Advanced Persistent Threat) are rapidly growing as a social and national threat. It is an intelligent cyber-attack that infiltrates the target organization and enterprise clandestinely using various methods and causes considerable damage by making a final attack after long-term and through preparations. These attacks are threatening cyber worlds such as Internet by infecting and attacking the devices on this environment with the malicious code, and by destroying them or gaining their authorities. Detecting these attacks requires collecting and analysing data from various sources (network, host, security equipment, and devices) over the long haul. Therefore, we propose the method that can recognize the cyber-targeted attack and detect the abnormal behavior based on Big Data. The proposed approach analyses faster and precisely various logs and monitoring data using Big Data storage and processing technology. In particular, we evaluated that the suspicious behavior analysis using MapReduce is effective in analysing large-scale behavior monitoring and log data from various sources.

Keywords
Anomaly detection, Abnormal behavior, Behavior feature, MapReduce, Big Data, Cyber-targeted attack, APT (Advanced Persistent Threat)

INTRODUCTION

The cyber-targeted attack is rapidly growing as an intelligent, persistent social and national threat. It is aimed at stealing industrial secrets or military secrets from major government agencies or enterprises and customer information from various smart devices and PCs, paralyzing the industrial control system and consequently causing astronomical physical damages (Kim et al., 2013). Unlike the traditional attack targeting unspecified many systems, the cyber-targeted attack is an organizational attack with a clear target; nowadays, it is used as an attack method for cyber terror, hacktivism, and cyber warfare. APT (Advanced Persistent Threat) is an intelligent cyber-targeted attack designed to steal confidential information or seize control over major facilities after infiltrating the network of the target organization. It is difficult to detect and respond to this attack in advance because the attack is made over a long period of time, and various malicious codes and attack roots are used. The detection of this attack requires collecting the organization's large-scale data from various sources (network, host/server, security equipment, devices, etc.) over the long haul and analysing the behavior history. In addition, correlation analysis is important in understanding the meaning of each individual attack behavior, since the attack behaviors are attempted confidentially and continuously until a final attack is successful.

With the development of cloud computing base technology (Weng et al., 2013; Kim, Byun and Jeong, 2013), we now have enough computing power to process large-scale data (e.g., Big Data). Therefore, we can integrate and analyse the numerous attack data that was difficult to do in the past. As a result, security intelligence technology based on Big Data analysis emerged. This paper proposes the abnormal behavior detecting method using MapReduce for the recognition and response of the cyber-targeted attack. This method is the security intelligence technology based on Big Data and focuses on the analysis of host log information among the various log data - enormous amounts of security logs, network and host information, and application transactions. Because all behavior of malicious codes is logged basically on the host.

The rest of this paper is organized as follows: Section 2 introduces the security intelligence technology, Big Data and cloud computing technology, and malware detection as related work. Section 3 describes the proposed method to analyse the abnormal behavior using MapReduce. Section 4 presents the experiment result and environments. Finally, Section 5 concludes and gives an outline for future work.

RELATED WORK

Security Intelligence Technology

Security intelligence technology is integrated security management technology that configures network and system security events to defend against a cyber-targeted attack. Research studies on internal behavior surveillance technology and product development are conducted in full scale using Big Data processing/analysis technology. With the Big Data processing/analysis technology expanded to various application areas, the technology is utilized as security event analysis technology. SIEM (Security Information & Event Management) leaders are introducing and utilizing Big Data analysis technology to apply intelligent security and respond to new attacks (Kim et al., 2013).

Related overseas projects include the CINDER (Cyber-INsiDER) and cyber genome programs that have been implemented by the US's DARPA since 2010. The CINDER program focuses on the analysis of internal staff's behavior, whereas the cyber genome project expresses the correlation and properties among application software, data flow, and users based on the formal analysis of abnormal behavior (DARPA, 2010). IBM QRader is a representative product that provides security intelligence technology. Having acquired Q1 Labs, which provides strong SIEM technology, IBM provides the function of collecting data from various sources and analysing the network and application behavior using QRader (IBM). Splunk provides predefined functions to support enterprise security monitoring and analysis and use case. For this, large-scale data and real-time correlation are analysed flexibly using the Splunk application (Splunk).

Big Data and Cloud Computing

Recently, Big Data technology based on distributed clustering (Sinha et al., 2013) and cloud computing technology has been applied to the various areas. Big Data storage and processing technology can be utilized in the security area. Currently, the researches related to security using Big Data technology are introduced. Lee et al. (2013) explained the technology that analyses Internet traffic using Hadoop MapReduce. To analyse the IP, TCP, HTTP, and Netflow data, MapReduce was implemented for each, with the I/O format defined. Another research (Choi et al., 2011) proposed a MapReduce-based security log analysis system that collects and analyses large-scale heterogeneous security logs (firewall, intrusion detection system, and web logs) in an integrated manner. Cloud computing technology is especially applied to the education system (e-learning), as it is an open and powerful environment. Weng et al. (2013) suggested a cloud-based learning center and a learning assistant on user devices, and Kim, Byun, and Jeong (2013) proposed cloud-computing based AEHS (Adaptive Education Hypermedia System) that enables a learner's preference.

Malware and Malicious Behavior Detection

Many researchers have been studying about malware or malicious behavior detection. Hwang et al. (2013) suggested the method classifying malicious web pages by an adaptive support vector machine. To classify malicious web pages, they defined the features to represent the essential characteristics of a web page and selected an adaptive support vector machine (aSVM) for learning training data. But this research was focused on only malicious web page. Ding et al. (2009) introduced the behavior-based dynamic heuristic analysis approach. The approach is similar to our approach using behavior based features, but our approach can make behavior model based on malicious and benign behavior by training a large amount of data on Big Data platform.

ABNORMAL BEHAVIOR DETECTION BASED ON BIG DATA

The APT attack intrudes upon the internal network and goes through a latent period to prepare a final attack (confidential information leak, system breakdown, etc.). The abnormal behavior detection method proposed in this paper can help to cope with the cyber-targeted attack in advance before the final attack is made. This section especially focuses on the analysis technique using Big Data processing and analysis technology (MapReduce) to detect the abnormal behavior.

Architecture of the proposed approach

In this paper, as we mentioned in the previous section, we analyse the large-scale accumulated data generated by the hosts and devices. That is because traces of malicious code or attacking tool which are frequently used for a cyber-targeted attack are mainly left in the hosts or devices.

Figure 1 shows the overall flow of our proposed approach. It is composed of two procedures: Behavior Rule Generation and Abnormal Behavior Detection. First to generate the behavior rules, our approach received the large-scale host data from the storage of Big Data platform (Hadoop) and extracted the behavior features. And then it generated the behavior rules using the machine learning technology, decision tree based on statistical data and stored them in the database. The second procedure is abnormal behavior detection which generated the feature description using MapReduce, compared and analysed it with the behavior rules, and detected the abnormal behavior.

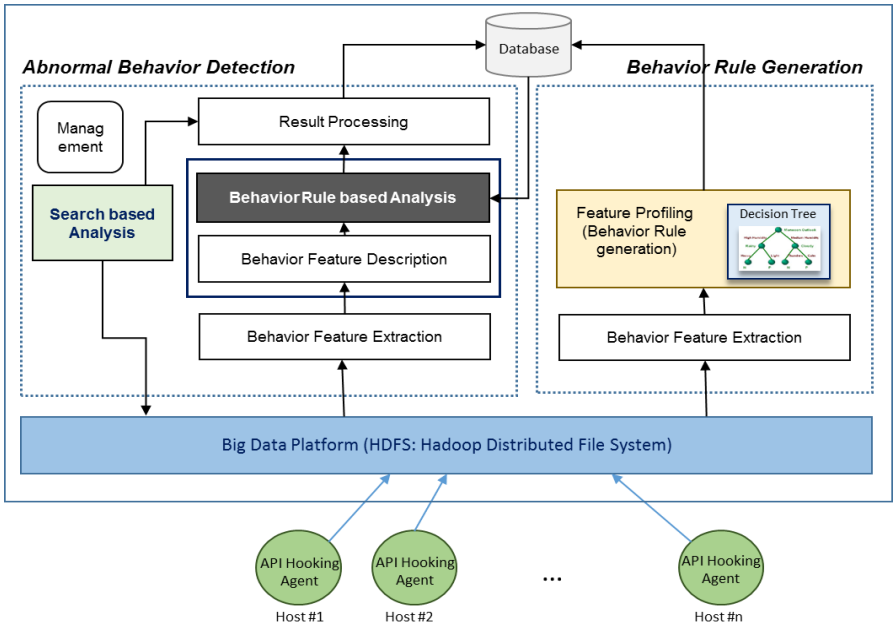

Figure 1: The overall flow of our approach

Table 1 summarizes the functions of our approach with their I/O data.

Table 1: Functions of the large-scale data analysis engine

Function name	I/O data	Description
Feature profiling (Behavior rule generation)	Feature information/ behavior rules	Defines the feature needed for large-scale accumulated data analysis and generates the behavior rules in advance
Feature extraction	Large-scale accumulated data/feature information	Extracts the behavior feature from the large-scale accumulated data
Pre-processing (Feature description)	Feature information/ feature information processed for analysis	Processes the information of the extracted feature into a form used for analysis
Search based analysis	Specified information/ analysis results	Search data related to specific info. from large-scale accumulated data
Behavior Rule-based analysis	Pre-processed feature information/ analysis results	Behavior Rule-based, large-scale accumulated data analysis
Result processing	Analysis results/ Result with data type used by other engines	Processes the analysis results into the data requested by other engines and stores them in database
Management	-	Overall management function for the analysis functions

Abnormal Behavior Features

All behavior of application including a malicious code (or malware) falls into four categories (process, registry, file and network). The behavior of malicious code is generated in the following five-step.

- **Step1. File Creation:** To hide malicious code, it copies its file in the system somewhere, downloads other malicious code through the network, or drops other malicious code. (File/Network related)
- **Step 2. Registry Registration:** It registers itself in registry to ensure persistence between reboots. (Registry related)
- **Step 3. Process Action:** It is executed as an independent process or thread injected in other process. Also it searches and kills other processes to remove the security programs such as anti-virus. (Process related)
- **Step 4. Network Action:** Through the network, it leaks system information, receives the attacker's command and other malicious code, or propagates it. (Network related)
- **Step 5. Goal of Attack:** Leakage of the confidential information, target systems destruction, and etc.

Therefore, to detect a suspicious process which has malicious action or similar action by malicious code, we need to define the feature to characterize the abnormal behavior according to this classification in advance. Abnormal Behavior Feature (ABF) has many single ABFs and complex ABFs with relatively high risk. Table 2 shows the major host-based features and related APIs. Currently, the Operating System of the targeted host is Windows 7.

Table2: Major host-based features

Type	Major host-based features	API	Code	Risk level
File	File deletion in the system folder	CreateFile /ReadFile WriteFile /CopyFile GetSystemDirectory GetWindowsDirectory …	F1	H
	File renaming in the system folder		F2	H
	File creation in the system folder		F3	H
	Executable file creation in the temporary folder		F4	H
	File creation in the temporary folder		F5	M
	Executable file creation		F6	H
	File creation		F7	M
Registry	Registry deletion	RegCreateKey RegOpenKeyExA RegSetValueExA RegQueryValueExA CreateServiceA OpenServiceA StartServiceA…	R1	H
	Service deletion		R2	H
	Adding automatic execution		R3	H
	Registry registration		R4	H
	Service registration		R5	H
	Adding a BHO item		R6	M
Process	Other process creation	CreateProcess FindProcess TerminateProcess CreateThread CreateRemoteThread WriteProcessMemory ShellExecute…	P1	H
	Other process termination		P2	H
	Other process search		P3	H
	DLL code injection		P4	H
	Thread creation		P5	M
Network	Port opening	WSAStartup WSASend Connect / Listen Send /Recv /Accept/ Gethostbyname InternetGetConnectedState…	N1	M
	Port binding		N2	M
	Network connection		N3	M
	Network disconnection		N4	M
	Data sending		N5	M
	Data receiving		N6	M
Complex	File creation through the network	URLDownloadToFileW or Connect, Send, NtCreateFile	C1	H
	DLL injection with createRemoteThread	NtOpenProcess,	C2	H

		VirtualAlloc, WriteProcessMemory, CreateRemoteThread		
	IAT hooking	LdrLoadDll, Strcmp, VirtualProtect	C3	H

Features are extracted from the large-scale accumulated data (related to the host) saved in HDFS for each process's behavior and are presented in a form needed for analysis using the pre-processing function, as shown in Figure 2.

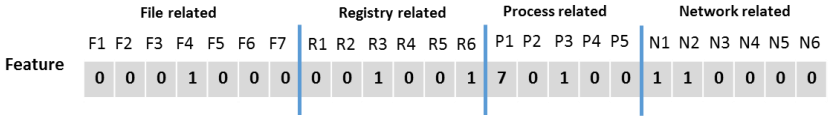

Figure 2: Feature description

The described data is created for each process. At this time, the process is analysed by combining the parent process and child process into a single process to identify the subject and flow of overall behavior. This is because most malicious codes perform abnormal behavior by creating a new process.

Abnormal Behavior Analysis Method using MapReduce

MapReduce improves the processing and analysis speed by handling the big data accumulated over the long haul in a distributed manner (Dean et al., 2008). In this paper, MapReduce is used to extract and describe features from the large-scale data saved in HDFS and to perform analysis. Figure 3 shows the host-analysis MapReduce handling process that analyses large-scale accumulated data to detect host-based abnormal behavior.

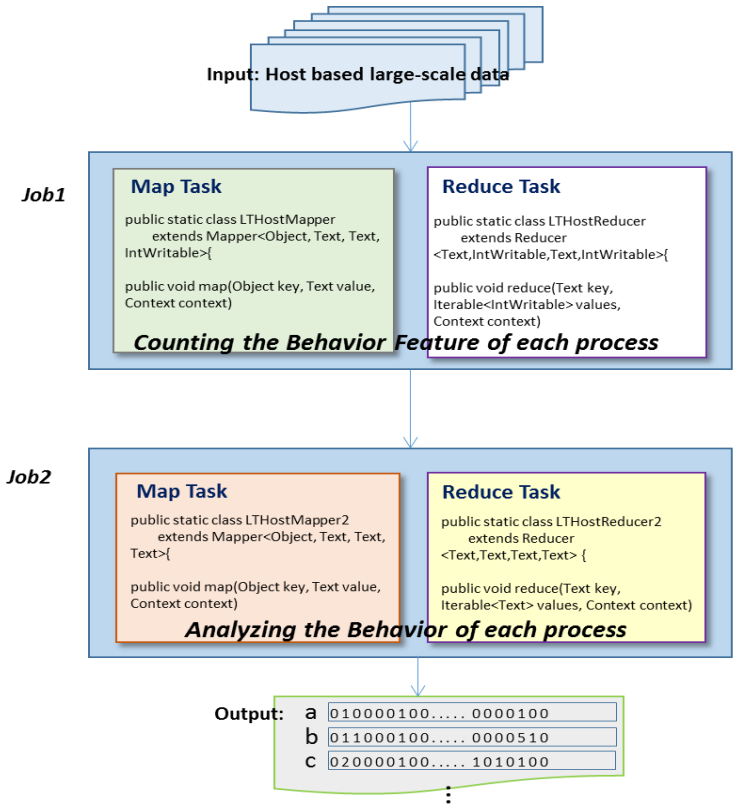

Figure 3: MapReduce to detect abnormal behavior in the host

The host-analysis MapReduce job for the abnormal behavior detection is composed of two jobs as shown in Figure 3. The job1 of MapReduce extracts process's features from large-scale data and counts the behavior features of each process. And the job2 makes the process's feature description with the output of MapReduce job1 and returns the result to the behavior rule based analysis module.

Firstly, job1 receives the host-based large-scale accumulated data and reads them sequentially. The mapper of job1 extracts the pair of process identifier (PID) and the behavior event defined as the feature. For example, in a pair (b, F1), b is PID and F1 is behaviour feature code. The reducer of job1 counts all pairs from the mapper according to the (PID, feature code). Therefore the result type of job1 consists of three tuples (PID, feature code, count). The interim result value generated by the job1 is entered as input value for job2.

Job2 makes the feature description to detect abnormal behavior of process (in figure 3, a, b, c is process identifier). The mapper of job2 creates one-dimensional array for each process as shown in Figure 2. Then it maps the result of job1 the array. Finally the reducer of job2 merges the arrays and store them in a file.

Through the result of host-analysis MapReduce job, we can identify the type and frequency of abnormal behavior performed by the process during a certain period of time (analysis interval).

In the behavior rule based analysis module, this result is compared with the abnormal behavior detection rule. The behavior detection rule was generated from the Decision Tree. It supports a tree-like graph for modeling decisions in the machine learning in feature profiling module. The decision tree has been trained and generated from the frequency of behavior (API call) of the process in our previous study (Moon et al., 2014) by the C4.5 algorithm of WEKA data mining tool. About the 3133 malwares and 1049 benign data from the previous study were used to train the decision tree model. This decision tree has 2.0% and 5.5% false positive rate and false negative rate. If the result of MapReduce meets the condition of the detection rule, that process is detected as an abnormal process.

EXPERIMENT RESULTS

Experiment Environment

The proposed method was implemented and tested on the Big Data platform composed of 12 nodes as shown in Table 3. First, 7 systems (Intel Xeon E5620 2.4GHz 8-core CPU, 20GB RAM, 2.5TB HDD) were clustered to build a Hadoop platform, and 2 systems (2.5GHz 6-core CPU, 64GB RAM, 500GB HDD) were used as MySQL cluster to save the data in real time. A total of 6 systems (excluding Hadoop master system) provide physical space (15TB) to save the accumulated big data. A total of 7 systems used as HDFS were running on CentOS 6.4, with Hadoop 1.2.1, HBase 0.94.11, and Zookeeper 3.4.5; MySQL Cluster 7.3 version was installed as in-memory database. Hadoop had a block size of 64MB and a replication of three.

Table 3: Big Data platform specification of our approach

	System	Specification	Quantity(EA)	Capacity(GB)
Real-Time	MySQLCluster (IBMx3550M4)	6core(2.5GHz) 64GB RAM 500GB HDD	2	128
	Esper (IBMx3550M4)	6core(2.5GHz) 40GB RAM 500GB HDD	1	40
	Storm (APPRO1624)	8core(2.4GHz) 20GB RAM 500GB HDD	2	40
Large-scale	HDFS (APPRO1624)	8core(2.4GHz) 20GB RAM 2.5TB HDD	7	1750

Performance Evaluation

In order to examine the influence of the various input file size, we executed host behavior analysis jobs with from 15MB to 28GB process behavior monitoring files on the 7-node and 1-node Hadoop environment. Figure 4 shows the performance of host analysis job1, job2 and total analysis job as the input file increases 15MB to 28GB on the 7-node and 1-node Hadoop environment. As shown in figure 4(a), the completion time of job2 is measured almost constant value, consequently, we do not consider the performance of a job2. When the file size is small (less than 300MB), there were a little differences of performance between 7-node and 1-node. However in figure 4(b), as the file size increases more than 300MB, analysis job on the 7-node achieves the improved throughput of 132 Mbps for 28GB. On the contrary, analysis job on the 1-node shows a slightly increased

throughput of 18.7Mbps for 3.3GB and a nearly constant throughput of 22Mbps for more than 3.3GB. In other words, the parallel/distributed computing such as MapReduce is efficient for processing and analysing large data sets. Especially because MapReduce is the most effective in selection-then-grouping-by-aggregation processing, we could improve the performance of processing and analysing data as applying MapReduce to our approach.

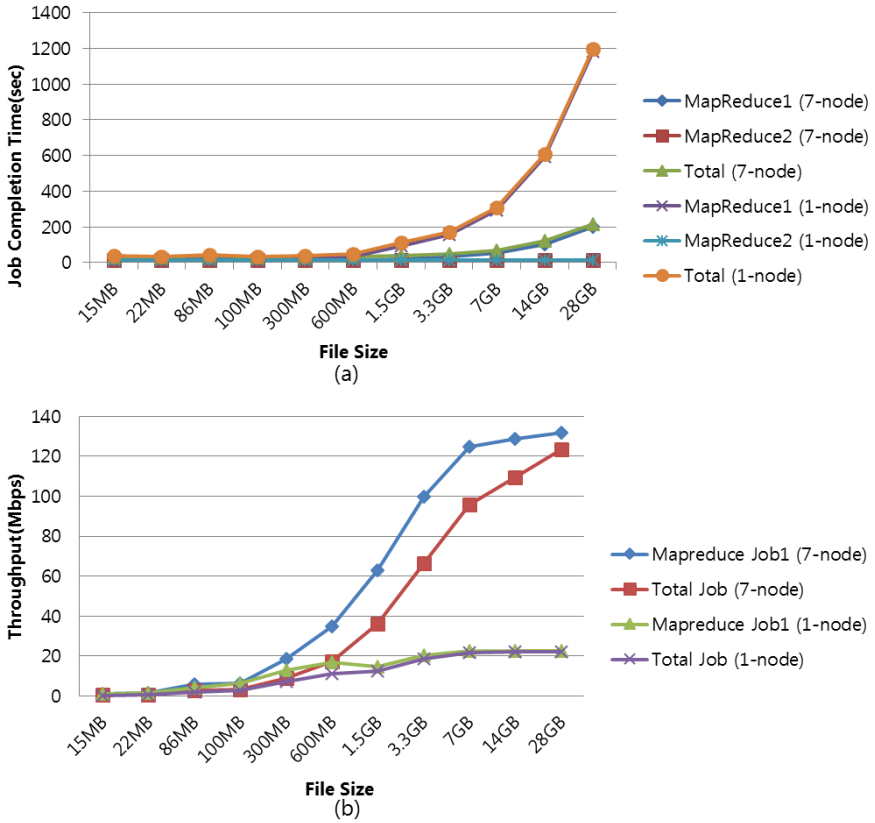

Figure 4: Performance on the 7-node and 1-node: (a) job completion time and (b) throughput

Figure 5 shows how host-analysis job throughput is enhanced when we increase the file size on the 7-node and 1-node test environments. As the file size increases more than 300M, analysis job on the 7-node achieves up to 5.8x increased throughput for 28GB.

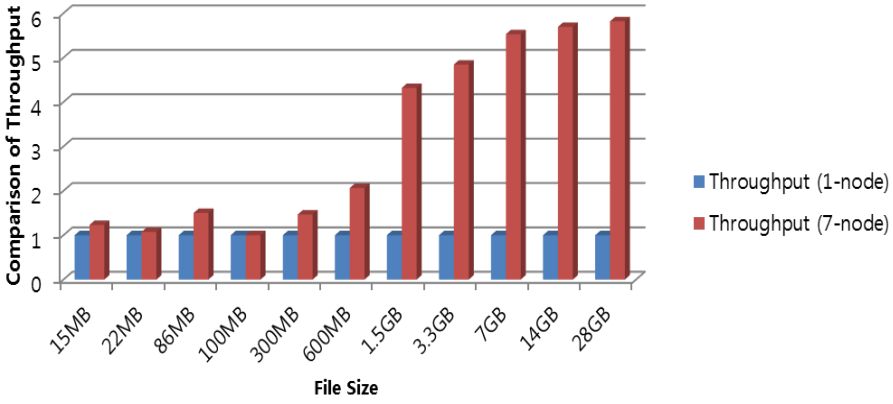

Figure 5: Comparison of Throughput of host-analysis job on the 1-node vs the 7-node

CONCLUSION AND FUTURE WORK

The cyber-targeted attack which has a certain latent period after infiltrating the system secretly to achieve the final attack objective attempts phased attack using various methods over a long period of time to hide its malicious behavior. Therefore, the attack cannot be detected without analysing large-scale accumulated data.

This paper has discussed the abnormal behavior detection method using MapReduce based on Big Data. We defined the behavior features to detect the cyber-targeted attack and analysed abnormal behavior using MapReduce based on the distributed computing platform, Hadoop. From the experiments on Hadoop platform with 1 and 7 nodes, we have shown that the host-analysis MapReduce job proposed in this paper is effective in analysing and processing large-scale accumulated data. In the future work, we will extend the Hadoop platform by adding 19 more systems. The extended platform will have the 55TB and 192GB total physical space for the accumulated data and the real-time data. In this environment, we plan to experiment the methods to analyse large-scale network data as well as host data and verify the analysis algorithms with abnormal behavior detection rate and generate the behaviour rule with the sequence as well as the frequency of malicious code's behavior.

ACKNOWLEDGMENTS

This work was partly supported by the IT R&D program of MSIP/KEIT [No.B0101-15-1293, Cyber targeted attack recognition and trace-back technology based-on long-term historic analysis of multi-source data]

REFERENCES

Choi, D., Moon, G., Kim, Y., & Noh, B. (2011). An analysis of large-scale security log using MapReduce. *Journal of the Korean Institute of Information Technology*, 9(8), 125-132.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM - 50th anniversary*, 51(1), 107-113.

Ding, J., Jin, J., Bouvry, P., Hu, Y., & Guan, H. (2009). Behavior-based proactive detection of unknown malicious codes. *International Conference on Internet Monitoring and Protection (ICIMP2009)*, 72-77. doi:10.1109/ICIMP.2009.20

Hwang, Y. S., Kwon, J. B., Moon, J. C., & Cho, S. J. (2013). Classifying malicious web pages by using an adaptive support vector machine. *Journal of Information Processing System*, 9(3) 395-404.

Kim, J. H., Lim, S. H., Kim, I. K., Cho, H. S., & Noh, B. K. (2013). Technical trends of cyber security with big data. *Electronic Communication Trend Analysis*, 28(3), 19-29.

Kim, J., Byun, J., & Jeong, H. (2013). Cloud AEHS: Advanced learning system using user preferences. *Journal of Convergence*, 4(3), 31-36.

Lee, Y., & Lee, Y. (2013). Toward scalable Internet traffic measurement and analysis with Hadoop. *ACM SIGCOMM Computer Communication Review*, 43(1), 6-13.

Moon, D., Lee, H., & Kim, I. (2014). Host based feature description method for detecting APT attack. *Journal of The Korea Institute of Information Security & Cryptology*, 24(5), 839-850.

Sinha, A., & Lobiyal, D. K. (2013). Performance evaluation of data aggregation for cluster-based wireless sensor network. *Human-centric Computing and Information Sciences*, 3(13). doi:10.1186/2192-1962-3-13.

Weng, M. M., Shilh, T. K., & Hung, J. C. (2013). A personal tutoring mechanism based on the cloud environment. *Journal of Convergence*, 4(3), 37-44.

DARPA (2010). R&D Support of DARPA Cyber Genome Program. General Dynamics, March 2010. Retrieved from http://publicintelligence.net/hbgary-general-dynamics-darpa-cyber-genome-program-proposal/

IBM (2015), Retrieved Sep. 15, 2015, from http://www-03.ibm.com/software/products/en/qradar-siem

Splunk. (2015). Retrieved Sep. 15, 2015, from http://www.splunk.com/