November-December 2019
ISSN: 0193-4120 Page No. 2447 - 2451

A Study on Data Analysis of Vector RDBMS Using Security Data

Cheolhee YOON[1], Jang Mook KANG[2], Jung joong KIM [3]
Police Science Institute, Asan Republic of Korea E-mail: bertter@police.ac.kr
Global Cyber University, Seoul, Republic of Korea. E-mail: honukang@gw.global.ac.kr
eGlobal Systems, Seoul, Republic of Korea.E-mail: john.kim@eglobalsys.co.kr

Article Info
Volume 81
Page Number: 2447 – 2451
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *12 December 2019*

Abstract

With the combination of ICT and manufacturing technology, Security Analystics has only recently been applied to industrial sites and public institution. Although Security data is increasing exponentially, it is difficult to grasp the hidden information in the process because of its huge amount. However, many analytical attempts have already been made using the Big Data Platform, and as a result, the analysis of production process data, which is the basis of Security and Safety, is progressing more and more efficiently. This paper also Sourced to study system model for quality defect analysis of Security Data Analysis. The proposed model is composed mainly of the data produced in the security data log for efficient analysis. The main equipment where log data is loaded is Web service segments such as Web servers and WAS, as well as firewalls, ips, DDos, and network equipment that control it in front of them. Artificial intelligence data analysis of log data is considered important, and schools and public institutions are also filtering out the vast amount of log data.

In this paper, the correlation analysis between events was applied in order to obtain the information required through artificial intelligence techniques and large-scale data analysis. In addition, the Commission proposed a secure log data analysis system that builds a virtualization-based infrastructure for analysis of massive amounts of collected secure log data and uses Vector DBMS to collect and store data, which is much faster than traditional methods.

Keywords*: Bigdata, Artificial Intelligence, Log Data, Vector DBMS*

1. INTRODUCTION

Network management at various sites, such as businesses, public institutions and schools, is exponentially increasing and complicating the amount of access data due to the procedures to manage, consolidate and enhance existing networks. It is the log that accounts for the majority of this data, and this log data can inform the various services of the information system in operation. Public institutions and school institutions are the main targets of malicious attacks such as phishing and spam attacks, and recently become IoT devices such as routers, smart televisions and refrigerators installed in ordinary homes as the waypoint for malicious attacks. Such cyberattacks are becoming increasingly sophisticated day by day, and if personal information collected by malicious hackers leads to financial accidents, the damage will be even greater. The main equipment where log data is loaded is Web service segments such as Web servers and WAS, as well as firewalls, ips, ddos, and network equipment that control it in front of them. Public institutions typically load logs from servers, security devices, switches, intrusion detection systems, and anti-ddos with an operating system, and collect log data from other desktops and laptops where internal users occur. [1]

Artificial intelligence data analysis of log data is considered important, and schools and public institutions are also filtering out the vast amount of log data. In this paper, the correlation analysis between events was applied in order to obtain the information required through artificial intelligence techniques and large-scale data analysis. In this paper, we proposed a security log data analysis system that builds a virtualization-based infrastructure for analysis of the massive amount of secure log data collected and that processes much faster than conventional methods using Vector Type DBMS for collecting and storing data

November-December 2019
ISSN: 0193-4120 Page No. 2447 - 2451

2. CASE STUDIES AI-BASED SECURITY LOG

A. *Preparing for AI-Based Security Log Analysis*

Web logs are logs that are stored on the server when users access the site on the Web, including access times, sites they connect to, IP addresses, browser identifiers, and operating systems. That is, it refers to the path used by the Web server user. Analyzing web usage in the direction of improving a website is called web log data analysis. The challenge for many organizations is the need to efficiently store, analyze, and manage rapidly growing logs with limited resources. A constant normalization process is required to store consistent log data in various formats. Normalization helps network administrators who control security systems to effectively analyze security log data. The usual logarithmic analysis tool makes the point of finding multiple log logic patterns for the same event that a person cannot easily find. Table 1 below shows the details of where the logs attempted for this analysis occurred. Logs collected through policies in detail as preventive activities for prevention of security agents should be collected and managed through the detection/response details, and even logs generated during policy changes should be collected and managed through the artificial intelligence data analysis methodology at the same time.[2][3]

B. *Vector Log Analysis*

Conventional logarithmic analyses were analyzed using the general RDBMS method and NoSQL method. Relational databases are generally very simple principles of tabulating the simple relationship between keys and values, and NoSql method provides a mechanism for storing and retrieving data using a less restrictive consistency model than traditional relational database[2] The NoSQL database is a highly optimized Key/Value storage method for simple SQL searches or additional operations. Although it has significant performance in relation to Latency and Throughput. This paper analyzed the automated log of artificial intelligence type through the vector-type RDBMS.

Table 1. Type of security control

Type	Description
Security Protection Activity	Managing Firewall Configuration
	About Security Trend Statistics
	Security Trend Statistics Details
Security Protection Detection Response Activity	Firewall Allowed Log Analysis
	Firewall Denied Log Analysis
	Analyze IDS and IPS events
	Analysis of web firewall events
Operation management Security system	Firewall Policy Configuration Management
	Firewall Backup and Recycle Management

Table 2. Lists of analysis logs.

Type	Item
Firewall Allowed Log Analysis	Allow (ALLOW) Threshold Maximum TOP N
	SOURCE CENTER TOP N
	Destination Center TOP N
	Source IP TOP N
	Destination IP TOP N
	Abnormal threat port statistics status
	Destination Port Statistics Status
	Monthly event occurrence status comparison
Firewall Denied Log Analysis	Deny threshold maximum TOP N
	SOURCE CENTER TOP N
	Destination Center TOP N
	Source IP TOP N
	Destination IP TOP N
	Abnormal threat port statistics status
	Destination Port Statistics Status
	Monthly event occurrence status comparison

C. Vector Processing Test

The test system uses Intel Xeon E7-8867 v3, 2.50 GHz, 45MB L3 Cache. 65 core as HP ProLiant DL 580, Gen 9 equipment, and memory size is set at 1TB for large capacity testing. For storage where data is stored, SAS-type SSDs and HDDs were prepared and separate Crucial SSDs were used to improve performance.

3. AI-BASED SECURITY LOG ANALYSIS

We need a mechanism that can integrate, analyze, and automate the security log. The ability to analyze the results of collecting, classifying and integrating various security logs is essential. In order to refer to complex analysis results as well as simple search in search and analysis, it is necessary to establish a high-speed analysis system that can handle large capacity in real time. The following table5 distinguishes differences between the two methods involving file I/O and CPU processing only from the CPU processing perspective.

CPU Processing is the time that takes the CPU process to perform regardless of the speed and time required of File Input/Output. File I/O Processing includes the time spent processing the file.[4]

In this test, security logs take the form of SAM files and have a system for collecting and sorting operational security logs into about 60 columns. The original logs collected were about 20TB, and loading for data analysis took advantage of CPU parallel processing and cluster methods. The loading time

November-December 2019
ISSN: 0193-4120 Page No. 2447 - 2451

was approximately 1 hour per TB of data. The size of the data to be tested was chosen as a 1-day security log, which is about 1TB, or around 3 billion cases. Since this test is based on large capacity, the security log for the 20th day was calculated to be about 20TB in total. Typically, a 30-day secure log for monthly statistics is available, requiring a size of around 30TB on a file basis.

First, the security logs of security equipment (systems) in use in the actual security operation were collected, loaded, and classified to facilitate search and analysis. Second, data cleansing was carried out to fit the table configuration in order to reuse the configuration details of the sheriff's point of view from the database into various scenarios. Third, all the SQL phrases based on the security system work, scenario processing, and rule-based practices were newly constructed to better use.

A. Case Test

Four test categories were divided into Test 1, Test 2, Test 3, and Test 4. Test1 and Test2 have the same number of searches in the secure log, but they include simple and complex searches depending on the search type. The 10 billion simple searches were divided into groups A, B and C, and 10 billion complex searches will be Group D. The following test is a simple search for 10 billion data, based on the time spent on CPU processing, excluding the time spent on File I/O. This was done separately because File I/O differed from several hundred MB per second to several GBs. There were seven types of SQL used in search and analysis, including aggregation, conditionality, aggregate/conditions, sector search, conditions, aggregation, subquery, Aggregation group, alignment.[6-15]

B. Security Log Analysis Comparison System Performance Platform - Denmark

In order to evaluate the data storage and speed performance during the log collection phase of the proposed system, the time of search and analysis was measured by applying the data to the stem based on Actian Vector in the test environment[5]

Table 3. Test Type

Test Group	Search Type
Test 1 (A,B,C group)	10 billion cases, simple data search
Test2 (D group)	10 billion cases complex search
Test 3	32 billion cases massive search
Test 4	55 billion cases massive search

Table 4. Details of SQL type and Syntax

SQL type	SQL Syntax (samples)
count	SELECT COUNT(*) FROM LOG_TABLE
where	SELECT * FROM LOG_TABLE WHERE NVL(SRC_IP,0) = 'x .x .x .x' AND NVL(DST_IP,0) = 'y.249.0.233' AND NVL(AGENT_IP,0) ='.x. x. x. x‘
count / where	SELECT COUNT(*) FROM LOG_TABLE WHERE AGENT_IP = 'x .x .x .x' AND CUSTOMER =Alpha
like / where	SELECT * FROM LOG_TABLE WHERE NVL(DST_IP,0) LIKE '211%' AND NVL(AGENT_IP,0) ='x.9.100.251' AND NVL(CUSTOMER,0)= Beta

C. 10 Billion Data Search

Use the Select count (*) from Log_table to complete the data. Loading has been carried out. At this time, it was shown that the first run time was 0.15 seconds and the second run time was 0.14 seconds measured. On the other hand, a composite run time is 3.17 seconds and 3.26 seconds is recorded. In 30 billion cases and 50 billion cases test, it showed high performance. The average time performed by all syntax is shown in the table7 below. Overall, loading speed was very good than normal RDBMS and NoSQL in the same environment.

Table 5. Details of SQL type and Test type

SQL Type [CPU Processing (In-Memory Technique)]	10billion cases Data Simple Search	10 billion cases Data complex Search	32 billion cases, Massive Search	55 billion cases, Massive Search
count	0.15 sec	-	0.22 sec	0.29 sec
where	3.77 sec	-	3.73 sec	5.00 sec
count / where	3.27 sec	-	4.27 sec	7.51 sec
like / where	3.45 sec	-	3.71 sec	5.57 sec
count / like / where	9.48 sec	-	27.65 sec	48.90 sec

November-December 2019
ISSN: 0193-4120 Page No. 2447 - 2451

4. CONCLUSION

High performance loading was maintained across all tables, which would dramatically increase performance if an artificial intelligence vector-type log analysis system was used throughout the environment. In other words, using Vector Processing technology to collect, process, store, and visualize log data ensures high performance compared to traditional Hadoop-based security log analysis systems. We have already seen very good performance at query processing time and log data processing Lee speed. The results of this study can be useful for improving the performance of existing security log analysis systems [17-18].

Log data, which is the track record of history, actions performed and records in all information systems, continues to occur in various places where the network exists. These log data can be used as important information that can be used in the future in the event of an emergency through analysis of the system. High-capacity Vector Processing-based security log analysis is required to help logs address unexpected security threats and contribute to the permanence of those systems. In traditional logarithmic analysis systems, the collection and storage of log data and the performance problems of the typical RDBMS in extracting meaningful data for analysis have been the disadvantages of increasing time-consuming by performing special methods of reducing data capacity.

In this paper, the vector-type search engine was used to solve these problems, and log analysis show high performance. We expect an AI data analysis system with much faster processing with vector type log analysis system.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP). (No.2018-0-00705, Algorithm Design and Software Modeling for Judge Fake News based on Artificial Intelligence)

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03038738, Algorithm Design & Software Architecture Modeling to Judge Fake News based on Artificial Intelligence)

REFERENCES

1. Wei-Yu Chen and Jazz Wang, "Building a Cloud Computing Analysis System for Intrusion Detection System", CLOUD SLAM, April 2009.
2. Cha Ji Hun, Lee Seung Ha, Kim Yang Woo, “Implementation of Security Log Collection System based on NoSQL”, Implementation of Security Log Collection System based on NoSQL
3. Michael Dirolf, Kristina Chodorow, "MongoDB The Definitive Guide", O'Reilly Media, 2011
4. Peter Boncz, Marcin Zukowski, and Niels Nes.MonetDB/X100: Hyper-pipelining query execution. In CIDR, 2005.
5. M. Zukowski, S. H_eman, N. Nes, and P. Boncz. Cooperative scans: Dynamic bandwidth sharing in a dbms. In Proceedings of VLDB, pages 723-734. VLDB Endowment, 2007.
6. M. Nam, and S. Lee, “Bigdata as a Solution to Shrinking the Shadow Economy”, The E-Business Studies, Vol. 17. No. 5. pp. 107-116, 2016. DOI: https://doi.org/10.20462/TeBS.2016.10.17.5.107.
7. S.H. Kim, S. Chang, and S.W. Lee, “Consumer Trend Platform Development for Combination Analysis of Structured and Unstructured Big Data”, Journal of Digital Convergence, Vol. 15. No. 6. pp. 133-143, 2017. DOI: https://doi.org/10.14400/JDC.2017.15.6.133.
8. Y. Kang, S. Kim, J. Kim, and S. Lee, “Examining the Impact of Weather Factors on Yield Industry Vitalization on Bigdata Foundation Technique”, Journal of the Korea Entertainment Industry Association, Vol. 11. No. 4. pp. 329-340, 2017. DOI: https://doi.org/10.21184/jkeia.2017.06.11.4.329.
9. S. Kim, H. Hwang, J. Lee, J. Choi, J. Kang, and S. Lee, “Design of Prevention Method Against Infectious Diseases based on Mobile Bigdata and Rule to Select Subjects Using Artificial Intelligence Concept”, International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 174-178, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18603.
10. I. Jung, H. Sun, J. Kang, C.H. Lee, and S. Lee, “Bigdata Analysis Model for MRO Business Using Artificial Intelligence System Concept”, International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 134-138, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18593.
11. S. Kim, S. Park, J. Kang, and S. Lee, “The Model of Bigdata Analysis for MICE Using IoT (Beacon) and Artificial Intelligence Service (Recommendation, Interest, and Movement)”, International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 314-318, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.21192.
12. S.H. Kim, J.K. Choi, J.S. Kim, A.R. Jang, J.H. Lee, K.J. Cha, and S.W. Lee, “Animal Infectious Diseases Prevention through Bigdata and Deep Learning”, Journal of Intelligence and Information Systems, Vol. 24. No. 4. pp. 137-154, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.4.137.
13. S. Lee, and I. Jung, “Development of a Platform Using Big Data-Based Artificial Intelligence to Predict New Demand of Shipbuilding”, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 19. No. 1. pp. 171-178, 2019. DOI: https://doi.org/10.7236/JIIBC.2019.19.1.171.
14. H. Hwang, S. Lee, S. Kim, and S. Lee, “Building an Analytical Platform of Bigdata for Quality Inspection in the Dairy Industry: A Machine Learning Approach”,

Journal of Intelligence and Information Systems, Vol. 24. No. 1. pp. 125-140, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.1.125.
15. Y. Shon, J. Park, J. Kang, and S. Lee, "Design of Link Evaluation Method to Improve Reliability based on Linked Open Bigdata and Natural Language Processing", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 168-173, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18601.
16. T. Minami and K. Baba, "A Study on Finding Potential Group of Patrons from Library's Loan Records", International Journal of Advanced Smart Convergence, Vol. 2, No. 2, pp. 23-26, 2013. DOI: https://doi.org/10.7236/IJASC2013.2.2.6
17. Chukurna, O., Nitsenko, V., Kralia, V., Sahachko, Y., Morkunas, M., & Volkov, A. (2019). Modelling and managing the effect of transferring the dynamics of exchange rates on prices of machine-building enterprises in Ukraine. Polish Journal of Management Studies, 19 (1), 117-129.
18. Hussain, H.I., Kamarudin, F., Thaker, H.M.T. & Salem, M.A. (2019) Artificial Neural Network to Model Managerial Timing Decision: Non-Linear Evidence of Deviation from Target Leverage, International Journal of Computational Intelligence Systems, 12 (2), 1282-1294.