Study on Effective Management of Cyber Incidents in Graph Database

Seulgi Lee, Hyeisun Cho, Byungik Kim, and Taejin Lee
Security R&D Team 1
Korea Internet & Security Agency
Seoul, Republic of Korea
{sglee, hscho, kbi1983, tjlee}@kisa.or.kr

Abstract—**Nowadays, cyber incidents are becoming increasingly intelligent, and they have escalated dramatically. For this reason, our research focuses on finding a solution to counter cyber incidents. We decided to build a multiple- and unified data warehouse, one of the many ways of controlling massive information and gathering meaningful intelligence to respond to cyber incidents. The major idea of this paper consists in correlating information based on the massive data set in a graph database. We concentrated on managing massive information in the cyber area and solving the problem when managing malicious information in a relational database. This project is also developing the system based on the architecture in a graph database. We expect the system to contribute to creating various intelligence types. This paper describes how to manage correlated information for building a data warehouse, which is meant to be a kind of infrastructure for responding to cyber-attacks effectively.**

Keywords- information management; cyber incidents; graph database; cyber threat intelligence

I. INTRODUCTION

Nowadays, cyber incidents are becoming increasingly intelligent, and they have escalated dramatically. Recently, many studies have been done on intelligence for cyber incidents. This is still very much a work in progress. Intelligence is generally accepted to be useful for tracking bad guys who conduct espionage in the cyber area [1]. In achieving this goal, there are many changes in the course of the process derived from past research. The purpose of this study is to describe some problems in the established analysis systems and solutions. The methodology proposed in this study is expected to be used for the management of massive correlated information in a graph database. The expectation is that the established systems will be able to provide intelligence for forecasting cyber incidents. We have developed a unified hub system to counter cyber incidents in a relational database. The system has structural characteristics and consists of two parts: a gathering subsystem and an analysis subsystem. The subsystems are suitable to deploy and operate independently, since the data structure with several gathering channels was designed to offer flexible extension; the same is true for the gathering subsystem. The rest of this paper is organized as follows: Section II introduces the previous studies on intelligence analysis system in a relational database; Section III presents ideas for management in a graph database as proposed in this paper; we conclude in Section IV.

II. RELATED WORK

A. *Data warehouse for cyber incidents*

This section describes the overall organization of the developed intelligence system. In a recent study, there was an attempt to react to potential cyber incidents in the future [2]. This study proposed a cyber defense operation framework which can classify attack groups and predict cyber incidents. It also suggested 5 type keys of attack group identification such as Email, attached file, malware, and OSINT (Open Source Intelligence). However, there are many more indicators showing the characteristics of cyber incidents; we have studied effective management using more diverse information. We have constructed a data warehouse for cyber security. The system can be divided into two main subsystems: gathering and analysis. Fig. 1 shows the system overview. The gathering system consists of gathering, scheduling, and management modules. The essential function in the gathering system is collecting information for cyber incidents, which is opened to the public. Following the collected information, the system groups derived pieces of information for managing their relationships, such as landing site, phishing email, and malware.

The analysis subsystem periodically pulls in information from the gathering subsystem into its database. As the system is pulling, resources and attributes are given unique IDs and stored. For instance, if information with the same value is already stored in a database, the system will not store that. It also extracts essential information from pulling data. The essential information originates from the intelligence report presented by leading antivirus vendors. Using this extracted information, we expect to be able to identify and cluster identical hackers or hacking groups. Through this past research, we could analyze the infrastructure used by most hackers to procure some zombie PCs for attacking victims.

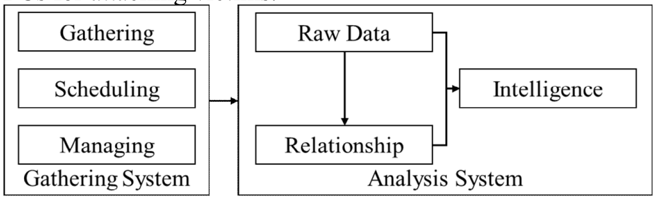


Figure 1. Data warehouse in a relational database

ISBN: 978-1-61208-509-8

Since there are various and numerous information feeds for cyber security in public, the system needs to manage the massive correlated information. In a DNS (Domain Name System) based blacklist, for example, a gathering subsystem usually collects around two million pieces of data per day. Because we have focused on tracing the relationship with cyber incidents, we have concentrated on a graph database that need not do a JOIN operation like a relational database. In the data warehouse, which uses a relational database, the clustering method of identical hacker was running with difficulty. Hackers experience difficulty with complex infrastructures which are used for cyber incidents. Therefore, hackers usually attack victims from a previously generated infrastructure. A relational database is barely suitable for analysis of these relationships. For this reason, there is still room for improvement in analysis performance. The approach we have used in this study seeks to improve analysis performance and enables effective management through a graph database.

B. Requirement for Improving the Existing System

1) Performance: In one of the proposed ways, the major analysis method creates a venn diagram-based stored relationship. Moreover, the sets in a diagram were extracted by recursive JOIN operations. Therefore, the system tends to produce a result slowly when much information is calculated. Although the system architecture considered huge amounts of information, the system did not work well.

2) Hard Grouping: The relationships between resources and attributes are grouped in a hard manner. The system computes groups by relationship. Moreover, these groups are originated by connected information derived by the initial resource. When the system analyzes and creates groups, the component parts of groups may be overlapped. Therefore, the system has difficulty supporting long-distance information from initial data.

3) Uncomfortable Visualization: The intelligence is caused by connected information. Or, put differently, the information is extracted from raw data collected by the gathering system. GUI, which has its own roles for control with users, has to present stored and created intelligence effectively. Unfortunately, it is hard to imprint users with intelligence.

III. Management in Graph Database

A. Proposed Scheme

We propose a scheme in graph database for building the management system that responds to cyber incidents. In earlier times, we thought that the information simply migrates to a graph database from a relational database. To apply information to a graph database, however, a hybrid architecture should be considered for the management of classified data like ordinary NoSQL. In this manner, we decided to divide data into two-tier information. Being a traditional database, the relational database has its own advantages with the effect of storing structured and patterned data. Such could commit massive data. This allows us to make a data warehouse from the raw data stored. Note, however, that a relational database makes managing the relationship between essential information and extracted gathering data difficult. That being the case, we structuralize architecture in a graph database whose structure needs to store the worked data. Effective management leads us to determine atypical information that should be stored in a graph database [3].

Fig. 2 shows the hybrid architecture that we structuralized. The node as entity in a relational database is a defined resources and attributes. Moreover, the relationship in a graph database is the relationship between resources and attributes just as it is.

Since the information is composed of fragmented information, there is a need to do preprocessing for managing the information collected by each channel in a graph database. For instance, the one with various channels in the gathering system, which could collect massive URLs for landing malware, stores raw data in a relational database as shown in Fig. 3. The developed system processes the information secured throughout the collected information in refined form and conducts relationship analysis in order to check for cyber-attacks, which seemed to be irrelevant superficially. The system then disassembles the information and analyzes cyber-attacks.

B. Stage for Building the Management System

1) Migrating Existing Information: If we want to manage data in a graph database, we should first mitigate the data pulled in from the gathering system from a relational database. If we do that, we would have to begin with the resource. We tried migrating information using a binay protocol called Bolt, which is served from Neo4j [4].

2) Reconstruction of Relationship: In the existing structure, we have to establish a relationship directly. For this reason, we could not use migration tools, which were used by most graph database communities. Therefore, the new system has to recreate the relationship. Furthermore, because infringement information (attack resources and attributes) is disassembled, the method of control relationship needs to be improved. TABLE I shows the components in a graph database as defined on grounds of APT (Advanced Persistent Threat) reports. As stated above, resources and attributes are stored as nodes on the system; relationship is configured in the same manner. This work makes a fundamental prototype in a graph database.

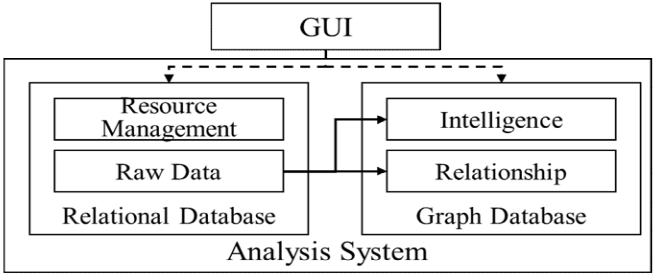


Figure 2. Hybrid architecture using a graph database

TABLE I. COMPONENTS OF A GRAPH DATABASE

Node		**Relationship**
Class	***Name***	
Resource	IP	C&C
	Domain	Create malware
	Hash	Defacing
Attribute	Account	Distribution
	Email	Download
	Filename	Filename
	FilePath	File Path
	Location	File String
	Process	Location
	Registry	Mapping
	String	Process
	Timestamp	Registrant
	URL	Registry
	URLPath	Landing

IV. CONCLUSION

With regard to utilization, we could create various intelligence types forecasting future cyber-crime from this system [5]. The importance of utilizing intelligence has been demonstrated by leading antivirus vendors. Before we enter into discussions on the detailed utilization of the system, we would like to stress that it is important to focus on the cyber incident report because the analysis system's output is nothing other than the intelligence that was carried out in the report. In this study, an efficient, accurate scheme was proposed to solve performance, which was derived by analyzing information in a relational database.

We have to establish a system for responding to cyber incidents because cyber incidents have escalated. Graph database is used for managing massive information like social network services. We can collect huge amounts of information about cyber incidents and construct the relationship between them. It is also a matter of increasing utilization using this system and making various intelligence types. This matter is a subject of further study. Therefore, future work should develop the intelligence analysis system and include verification utilization.

ACKNOWLEDGMENT

This work was supported by a grant from the Institute for Information & Communications Technology Promotion (IITP) funded by the Korean government's MSIP (no. B0101-16-0300, Development of Cyber Blackbox and Integrated Security Analysis Technology for Proactive and Reactive Cyber Incident Response).

REFERENCES

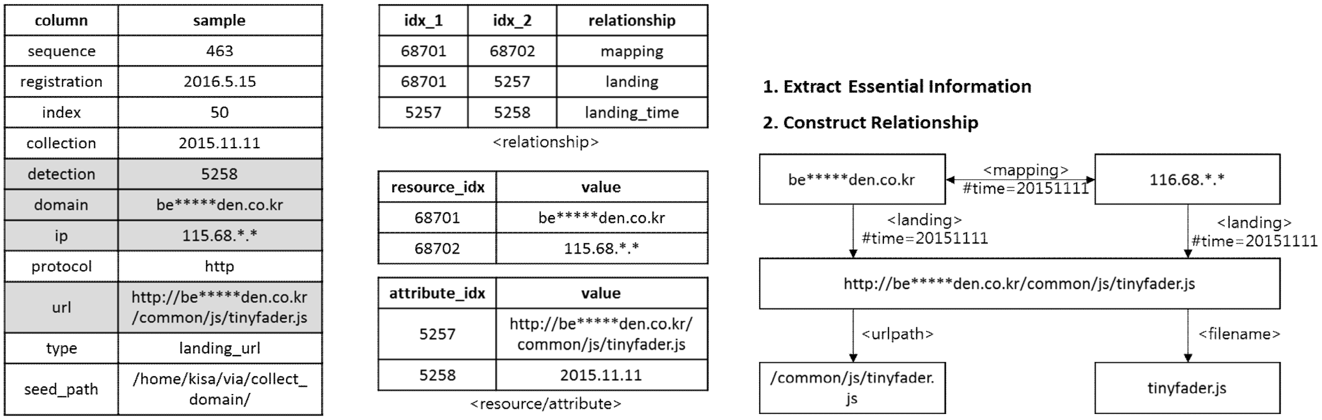
[1] G. Wangen, "The Role of Malware in Reported Cyber Espionage: A review of the Impact and Mechanism," Information, 6(2015), pp. 183-211, 2015.

[2] W. Kim, C. Park, S. Lee, and J Lim, "Methods for Classification and Attack Groups based on Framework of Cyber Defense Operations," KTCP, Vol.20 No.6(2014), pp. 317-328, 2014.

[3] C. Vicknair et al., "A comparison of a graph database and a relational database: a data provenance perspective," In proceedings of the 48th Annual South-east Regional Conference, ACM SE '10, pages 42:2-42:6, 2010.

[4] https://neo4j.com/developer/language-guides/ [accessed September 2016]

[5] H. Cho, S. Lee, B. Kim, Y. Shin, and T. Lee, "The study of prediction of same attack group by comparing similarity of domain, " Information and Communication Technology Convergence (ICTC), 2015 International Conference, pp.1220–1222, October 2015.



column	sample
sequence	463
registration	2016.5.15
index	50
collection	2015.11.11
detection	5258
domain	be*****den.co.kr
ip	115.68.*.*
protocol	http
url	http://be*****den.co.kr /common/js/tinyfader.js
type	landing_url
seed_path	/home/kisa/via/collect_ domain/

(a) Raw data(Landing sites)

idx_1	idx_2	relationship
68701	68702	mapping
68701	5257	landing
5257	5258	landing_time

<relationship>

resource_idx	value
68701	be*****den.co.kr
68702	115.68.*.*

attribute_idx	value
5257	http://be*****den.co.kr/ common/js/tinyfader.js
5258	2015.11.11

<resource/attribute>

(b) Management

(c) Refined raw data in a graph database

Figure 3. Example of stored raw data in a relational and graph database