Security Operation Center Architecture for E-government based on Big Data Analysis

[1]Zaur Fataliyev, [2]Yadigar Imamverdiyev, [3]Hanseok KO
[1,2]ANAS Institute of Information Technology, Baku, Azerbaijan
[1,3]School of Electrical Engineering, Korea University, Seoul, Korea
[1]zaur@korea.ac.kr, [2]yadigar@lan.ab.az, [3]hsko@korea.ac.kr

***Abstract*— This paper proposes a new architecture for Security Operations Center for handling Big Data and usage of Big Data Analytics for network security of organization. Design flow of proposed architecture has been explained for each step. Proposed architecture is designed using open-source platforms which is designed to handle Big Data.**

Keywords - Big data, E-government, SOC, hadoop, Big data analysis

I. Introduction

Transparency and productivity has always been of prime importance for government organizations. Therefore, in order to provide transparent and clear communication in short service time between citizens and government, increase demand for public services, increase financial efficiency, decrease cost and eliminate corruption, it is important to build information system infrastructure. Existing advances in technology is able to address solutions to aforementioned issues due to several reasons.

Public sector accumulates huge amounts of data, whereas no commercial organization can accumulate as large data as public sector. Although 70% of accumulated data is unstructured [1]. Inflow of data is stored in form of texts and different documents such as plans, applications, reports, complaints, proposals, and so on. Accumulated data is growing fast, universal average of growth is about 30%. Due to continuous interaction of citizens and government organizations, as well as inter-organizational and intra-organizational interactions result in massive amount of data.

As a result, growing data requires more resources to process. It is getting more and more difficult to structuralize, process archived data and search and obtain required data. Inflow of unstructured data is processed before storage which slows down the processing cycle substantially and this in turn increases processing and storage expenses.

Thus, huge amount of data accumulated in public sector and a number arise while storage and analysis of such data. Without Big Data technology, it is becoming nearly impossible to process and analyze accumulated unstructured data with trivial approach based on relational database. As a result, in near future, government organizations will face the challenges of Big Data. That’s why it is crucial to tune existing infrastructure which can handle those challenges. In order to provide effective communication between government organizations and citizens, it is important to overcome such challenges created by Big Data.

One of promising application areas of Big Data technologies is sophistication of security center of e-government. This paper focuses on redesign of Security Operation Center (SOC) which can handle Big Data and provide better security for e-government.

II. Problem statement

A. Big Data and E-government

Amount of data inflow in public sector is increasing, and size of such databases is in terms of terabytes and even can reach to petabytes. Research by MGI experts shows that, European Union can potentially reduce administrative costs by 15-20% (i.e.150-300 billion euro) with application of Big Data, which can bring benefits to countries outside Europe as well [3].

IBM has prepared a report titled “Realizing the Promise of Big Data” based on Survey of 28 IT directors and 10 trends of applications of Big Data in public sector is increasing, and size of database is in terms of terabytes and even can reach to petabytes. Research by MGI experts shows that, European Union can potentially decrease reduce administrative costs by 15-20% with , application of Big Data [4].IBM prepared a report based on survey of 28 IT director named “Realizing the Promise of Big Data” which contains 10 trends of Big Data related to government projects. Research shows that public organizations are in the beginning stage of applications of Big Data.

Research by Gartner shows that Big Data investments are least development in Government sector (only 16%) [5]. Australian government has adopted “The Australian Public Service Big Data Strategy” in august of 2013. This strategy is a systematic public approach to application of Big Data and support ICT investments in order to expand existing services, propose new services and providing better political advices, at the same time protecting privacy of citizens.

Big Data has captured the interest of law enforcement organizations and national security agencies in order to accurately detect criminals and terrorists.

Companies and government organizations face the problem of Big Data such as legality and quality, and also uncertainty of essence of data. Another challenge is the quality of process and analysis of Big Data. Any wrong decisions can result in dreadful consequences for citizens. Not only citizens, but also social life, economy and politics might as well suffer from such wrongdoings.

B. Paradigm Shift in Information Security

There are number of conditions that gives us grounds of premises to talk about the paradigm shift in information security. As complexity of software products and information systems increase, vulnerability loopholes in them increases as well. Malicious attacks by hackers using such vulnerabilities will be more dreadful than ever.

Counter actors are becoming more and more professional and organizing more sophisticated attacks. Asymmetric actors are able to invest adequately in malicious attacks. In contrast to trivial general damage oriented attacks, modern cyber-attacks are targeted to specific victims with sophisticated strategies. The main purpose of such attacks is to get access to valuable properties and keep it for as long as possible. Valuable properties are meant to be intellectual properties such as software codes, algorithms, customer database and other corporative secrets. Specific victim oriented attacks are mostly “Advanced Persistent Threats (APT)” [6]. Term appeared first time in US Military during such attacks by foreign intelligence / government supported threats.

APT is a special type of threat that is targeted to government structures, organizations and individuals. Such attacks utilize latest advancements in psychological, social engineering alongside with advanced technology. Each word in this complex term has precise reason to be in there.

- Advanced – source of the attack is well-supported organization, and has sufficient resource, technology, well-trained engineers to realize such attacks.
- Persistent – Attacker is patient, advances with specific steps and ready to utilize significant efforts in order to achieve its goal. If one approach fails, another approach is tried to get the work done. In contrast to trivial attacks, target of such attacks are chosen carefully and attacks may continue from a month to several years.
- Threat – source of attack is a threat to interests of target.

Currently, majority of information security systems are based on finding threats by correlating incoming signals and known threats. Such systems is not able to address solution to unknown threats and also APT’s. Big Data technology has potential to address solutions to aforementioned problem.

C. Big Data Analytics and Information Security

Big Data technologies have been applied in information security for a while now. Security Information and Event Management (SIEM) system has been introduced to market in 1996 and it has had widespread usage in mid-2000s. Its unquestionable fact that SIEM systems should be attributed to Big Data technologies: event collection, normalizing and correlation is essential and must-have unit of all modern Big Data systems.

Trivial SIEM systems has several limitations:

- Detection and investigation of threats is limited;
- Limited number of data sources is used;
- Data loss during data filtering and normalizing is unavoidable;
- Relational Database Management System (RDMS) is a bottleneck for data scale and processing speed;
- Forensic investigation of incidents;

There are several ways of integrating SIEM system with Big Data technology. Currently the second generation of SIEM systems are emerging, but their functionality to a large extend prolongs to Security Intelligence.

According to research by Gartner, 25% of companies will be using Big Data analytics for cyber security. It has been concluded that, Big Data analytics will provide more reliable ground of premise for Security Systems and will decrease false alarm rate significantly. Big Data analytics will not only provide better security systems, but also provide prospect of correlating events with business systems.

According to experts, Big Data has following promises in the area of information security.

- Better detection accuracy of cyber-attacks;
- Real-time correlation and anomaly detection;
- Deep forensic analysis;
- Visualization and analytics tools for Big Data;
- Faster and better decisions;

Aforementioned opportunities require comprehensive research and development in the area of Big Data Processing and Analysis.

III. SOC Architecture based on Big Data Technology

A. Trivial Approach

Trivial Security Operation Center (SOC) is made up of five distinct modules: event generators, event collectors, message database, analysis engines and reaction management software. These units are still valid for big data, although trivial design is not able to handle huge amount of data.

- **Event generators** are responsible for event generation.
- **Event collectors’** purpose is to gather information from different sensors and translate them into a standard format, in order to have a homogeneous base of messages.
- **Message database** is database of messages. Besides storing data, it also performs basic level of correlation such as identifying duplicates.
- **Analysis engines** are responsible for the analysis of events stored in Message database. They are to perform various operations in order to provide qualified alert messages.

- **Reaction management** software used to define the ensemble of reaction and reporting tools used to react against the offending events taking place on or within the supervised systems.

Figure 1 shows the trivial SOC Design. Most of modern SOCs are based on this architecture. This architecture fails to handle Big Data because it was not designed for it. In following sections, we will look at how SOC should be designed to handle Big Data [6].

B. SOC for Big Data

With massive amount of data inflow, trivial SOC is not able to handle Big Data. In order to handle Big Data inflow, following four steps should be optimized at each level.

1) Data collection

Data comes from different sources, and number of such sources are increasing everyday. Data is collected from heterogeneous sources i.e. cloud, virtual and real devices.

2) Data Integration

Collected data parsed to derive intelligence from cryptic log messages. Parsing and storage of data in forms that minimizes the access time latency. One main issue here is that trivial relation based database is not effective for storage purpose. New generation of graph based databases is more suitable for the purpose, which can handle linked data. Data is to be stored in form of linked data.

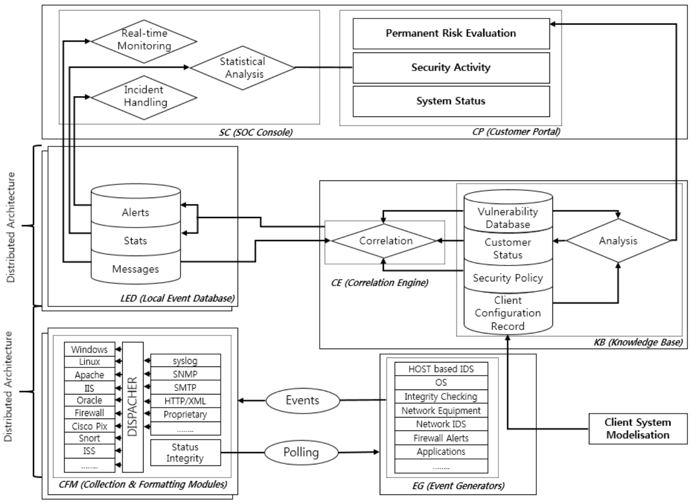

Figure 1: Trivial design of Security Operation Center (SOC)

3) Data Analytics

Analyzing data collectively which involves combining logs from multiple sources and correlating events together to create real-time alerts. Other than correlating events, unsupervised machine learning algorithms should be developed to detect or alarm about possible threat.

4) Threat Response

Appropriate response to detected threats. Timely response to threat is important because attack may spread into system if not handled in time. Incident handles respond to suspected threats, recover from any damage and extract features of threat for database retraining and future history.

With massive amount of data flow, following issues are priority in Security of organization.

5) Real-time correlation and anomaly detection

Correlation and anomaly detection is arguably the most important part of SOC. Vulnerability database should be rich and updated in regular bases with detected attacks, possible attacks and so on.

6) High-speed processing

In order to process large-scale data in adequate time, processing speed should be high. Distributed architecture for such as Hadoop is suitable for Big Data Processing.

7) Flexible Processing

Data coming from different sources, in different forms, and different times. Processing should be flexible to handle any kind of data inflow.

8) Forensics for deep visibility of network activity

It is important for advanced analytics, such as analyzing how applications are being used, how threat is moving across network, previous history of such attacks and etc.

Figure 2 shows what moment Security platforms should provide solutions to, and what Big Data is able to provide. Advancements of Machine Learning, Data Science, Artificial Intelligence and other fields may be applied to extract meaningful data from Big Data.

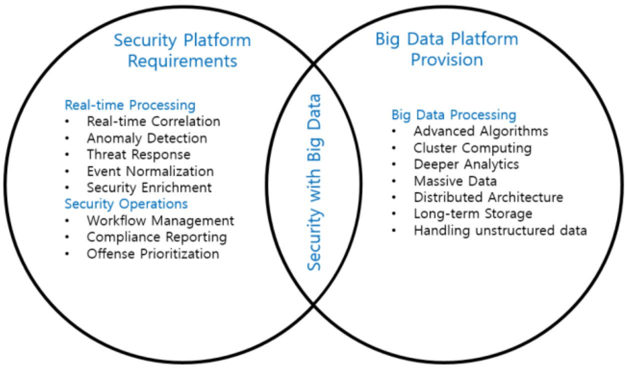

Figure 2 Security Platform Requirements and Big Data Platform Provisions

C. Security Analytics with Big Data

Security analytics using Big Data is in development stage, several companies are working on such platforms.

4 Steps of SOC using Big Data are the following:

1) Data input

Data sources for security analytics can be categorized into two, namely, active and passive data sources. Active data source are data generators in real-time.

Passive data sources: Computer-based data, Mobile-based data, Physical data of user, Human Resource data, Travel data, SIEM data, Data from external sources and etc.

Active data sources: Credential data, One-time passwords, Digital Certificates, Knowledge-based questions, Biometric identification data, Social media data and etc.

2) Data Formatting

Data comes from different sources. Before processing these data, Data is to be formatted properly and stored in database. Tools such as Email2DB (formatting emails), Log Parser (Formatting Logs), JSON, ParseRat and many others exist to format data and store in database for easy processing. Custom tools can be developed as well to format data and store according to organizations protocols.

3) Data Enrichment

As data comes from multiple sources, certain portion of the data is not complete. Data may include outliers, missing values, errors and etc. This kind of data is to be refined before further step. Tools for different purposes such as Maximo, MDO clarify, Sparesfinder and others has been developed. Other than available tools, custom tools may be developed to remove outliers or recover missing data.

4) Big Data Processing

Collected data have always been analyzed for security. But traditional design is not able to handle big data for two main reasons

- Size of Big Data is very large – traditional systems are designed to handle certain amount of data due to amount of inflow and storage problem. Data collected in organizations were usually deleted after few months
- Big Data is unstructured – trivial systems are rigidly bound to predefined schemas

A case study Zion Bandcorporation found out that traditional SIEM system is not able to handle Big Data. Searching among month’s load of data took 20 minutes to up to an hour while new Hadoop system running queries with Hive gave same result in approximately one minute [1, 3]. That’s why new technologies have been developing (such as Hadoop, Hive, NoSQL, WINE, etc.) to handle big data. Figure 3 is proposed overall SOC architecture which can handle Big Data.

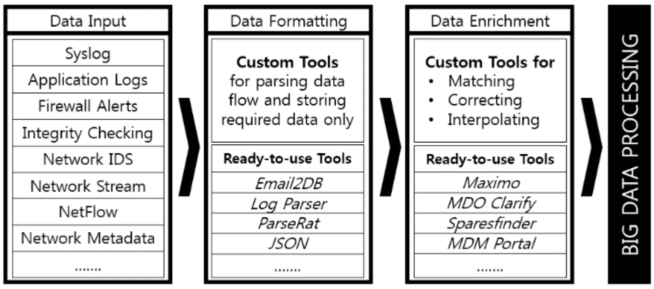

Figure 3: SOC Architecture integrating Security Analytics using Big Data

Processing Big Data is important part of the architecture. Trivial SOC design is not able to handle the huge flow of data, that’s why new architecture is proposed based on technology for Big Data. The overview of Big Data Processing is given in Figure 4.

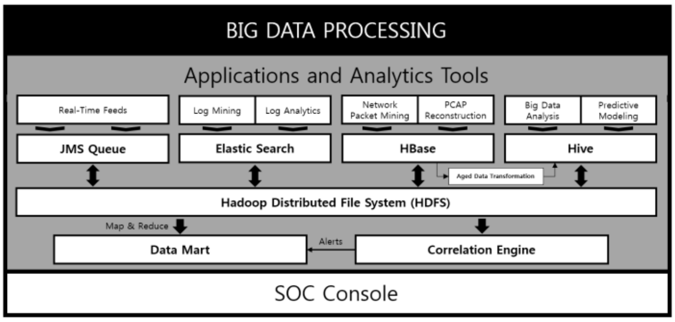

Figure 4: Big Data Processing

- **JMS Queue** is responsible for handling to process each message only once and one at a time.
- **Elastic Search** is a real-time search and analytics engine. It is able to search any kind of document. Log data from different devices come in different forms, and elastic search (Developed in Java) can handle any kind of log file [8].
- **HBase** is a non-relational, distributed database. It has favorable features such as compression, in-memory operation, Bloom filters and runs on top of HDFS [9].
- **Hive** is a data warehouse that runs on top of Hadoop for providing data summarization, query and analysis [10].
- **HDFS** (Hadoop Distributed File System) is responsible for handling of distributed file system [11].
- **Data Mart** is responsible for getting data out to the user from the data warehouse. Input to Data Mart is data MapRedused from HDFS and alerts from Correlation Engine.
- **SOC Console** is user interface for experts. It is used for monitoring network activity, responding to new, previously unknown threats and etc.

IV. RESULT

We have proposed a design for architecture of a Security Operation Center that can handle Big Data Processing. This architecture takes input from different sources such as application logs, real-time feeds, Network IDSs and others. Input data is formatted in the next step and stored in non-relational Database in form of linked data. Such database makes it convenient to access related data in low-latency. Because retrieved data has errors, such as missing data, outliers, data is enriched in the next step. Data Formatting and Data Enrichment can be built using ready-to-use tools developed by various companies or custom tools may be developed in order to meet organizational requirements. And after completing previous steps, big data is processed in Big Data Processing unit. This unit designed to handle processing of large data and its build using only open-source platforms. Figure 4 is the architecture of Big Data Processing unit, where all kinds of inputs are being handled and processed.

V. Conclusion

As usage of information technology becomes widespread, data to handle is increasing drastically. E-government is without any doubt, one of the large infrastructures which should handle processing of large-scale data and also generate data itself. Security is an important issue for e-government because it has data of citizens. Leakage of such information will violate privacy and it might even result in fatal consequences. It is time to utilize Big Data analytics for security purposes of e-government. Trivial Security Operation Center is not designed to handle large amount of data and it lacks properties such as flexibility, predictability and intelligent decision making. SOC architecture proposed in this paper is able to handle large amount of data, and integrate advanced Machine Learning algorithms for Network Security.

References

[1] Hitachi Data Systems. (February 2013). "How to Efficiently Manage All of Your Unstructured Data". Available: http://www.hds.com/assets/pdf/hitachi-white-paper-how-to-efficiently-manage-unstructured-data.pdf

[2] K. G. Coffman and A. M. Odlyzko (July 6, 2001), "Grows of the Internet" (Preliminary version), AT&T Labs – Research. Available: http://www.dtc.umn.edu/~odlyzko/doc/oft.internet.growth.pdf

[3] OECD, "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "big data"", Supporting Investment in Knowledge Capital, Growth and Innovation, pp 319-356, Oct. 2013

[4] Kevin C. Desouza, "Realizing the Promise of Big Data", Internet: http://www.businessofgovernment.org/report/realizing-promise-big-data

[5] Ian Bertram, "Small initiative in achieving Big Data", Internet: http://blogs.gartner.com/ian-bertram/

[6] Colin Tankard, "Persistent threats and how to monitor and deter them", Network Security, vol. 2011, pp. 16-19, August 2011.

[7] Renaud Bidou, "Security Operation Center Concepts and Implementation", [PDF] Available: http://iv2-technologies.com/SOCConceptAndImplementation.pdf

[8] Radu Gheorghe, Metthew Lee Hinman, "Elastic Search in Action" Version 11, Manning Publications 2014, pp. 1-22.

[9] Nich Dimiduk, Amandeep Khurana, "HBase in Action", Manning Publications 2013, pp. 3-51.

[10] Thejas M Nair, "Getting Started – Apache Hive", https://cwiki.apache.org/confluence/display/Hive/GettingStarted

[11] Dhruba Borthaku, "The Hadoop distributed file system: Architecture and design", The Apache Software Foundation, 2007.