빅데이터 처리를 위한 보안관제 시각화 구현과 평가

[1*]전상준, [2]윤성열, [3]김정호

Design and Evaluation Security Control Iconology for Big Data Processing

[1*]Jeon Sang June, [2]Yun Seong Yul, [3]Kim Jeong Ho

요 약

본 연구에서는 민간기업들이 전체적인 보안관제 인프라를 구축 할 수 있도록 오픈소스 빅데이터 솔루션을 이용하여 보안관제 체계를 구축하는 방법을 기술한다. 특히, 보안관제 시스템을 구축할 때 비용·개발시간을 단축 할 수 있는 하나의 방법으로 무료 오픈소스 빅데이터 분석 솔루션 중 하나인 Elastic Stack을 활용하여 인프라를 구축했으며, 산업에 많이 도입되는 제품인 Splunk와 비교실험을 진행했다. 또한 두 솔루션을 기능, 사용의 용이성, 서비스지원, 기술지원 등을 비교해석 한 결과, Elastic Stack이 사용자간 커뮤니티, 오픈 솔루션면에서 빅데이터의 보안관제가 유리함을 알 수 있었다. Elastic Stack을 활용해 보안 로그를 단계별로 수집-분석-시각화 하여 대시보드를 만들고 대용량 로그를 입력 후 보안관제 검색 속도를 측정하였다. 이를 통해 Elastic Stack이 Splunk를 대체 할 수 있는 빅데이터 분석 솔루션으로 기업들이 접근 가능성을 얻을 수 있다.

Abstract

This study describes how to build a security control system using an open source big data solution so that private companies can build an overall security control infrastructure. In particular, the infrastructure was built using the Elastic Stack, one of the free open source big data analysis solutions, as a way to shorten the cost and development time when building a security control system. A comparative experiment was conducted. In addition, as a result of comparing and analyzing the functions, convenience, service and technical support of the two solution, it was found that the Elastic Stack has advantages in the security control of Big Data in terms of community and open solution. Using the Elastic Stack, security logs were collected, analyzed, and visualized step by step to create a dashboard, input large logs, and measure the search speed. Through this, we discovered the possibility of the Elastic Stack as a big data analysis solution that could replace Splunk.

Keywords: *Elastic Stack, Big Data, Security Control, Security Monitoring, Splunk, SearchEngine*

[1*] 교신저자 *한밭대학교 컴퓨터공학과 박사과정(sjjeon@kigam.re.kr)*

[2] *한국인터넷진흥원 연구원(yso920616@naver.com)*

[3] *한밭대학교 컴퓨터공학과 교수(jhkim@hanbat.ac.kr)*

Received: Dec. 05, 2020, Revised: Dec. 24, 2020, Accepted: Dec. 29, 2020

I. 서론

1.1 연구배경

G-PRIVACY 2019 에서 한국인터넷진흥원(KISA)의 발표자료에 의하면 사이버 침해사고는 대부분 취약한 솔루션을 사용하는 기업의 웹사이트 및 인터넷 기반 신 서비스가 해킹에 취약해 정보유출 되는 경우가 80%를 차지한다고 발표했다[1]. 이제는 정부, 공공기관 뿐만 아니라 일반 기업체들도 정보보호에 대한 중요성의 인식이 퍼지며 서비스 구축 시 다양한 국가 지원사업 및 자체 예산을 통해 네트워크 장비, 서버, 보안 장비 등을 포함한 전산자원을 필수로 구축하고 있지만, 문제는 침해사고 발생 및 전산 장애 등을 실시간으로 한눈에 볼 수 있는 시스템의 구축은 일반 기업에서는 비용 및 개발인력부족 등의 이유로 도입을 망설이고 있는 실정이다. 따라서 민간기업들이 보안관제 인프라를 구축 할 수 있도록 오픈소스 빅데이터 솔루션을 이용하여 보안관제 체계를 구축하는 방법을 제안하고 Elastic Stack 을 이용한 보안관제용 빅데이타 솔루선을 해석하였다.[2,3]

1.2 연구목적

본 논문에서는 빅데이터 분석시스템 도입 시 비용 · 개발시간을 단축 할 수 있는 하나의 방법으로 무료 오픈소스 솔루션 중 하나인 Elastic Stack 에 대해 기술하고, 빅데이터 분석 솔루션중 산업계에서 주로 사용하는 제품인 Splunk 와 검색 성능 비교실험을 진행하였다. 이를 통해 Elastic Stack 이 Splunk 를 대체 할 수 있는 솔루션인지 실험하였다. 오픈소스 코드 솔루션 Elastic Stack 을 활용해 로그 분석시스템을 구축했으며, 대용량 보안 로그 분석은 유료 솔루션과 비슷한 성능을 발휘한다는 것을 확인하였다. 또한, 단순한 문자열 분석뿐만 아니라 시각화 분석과 대시보드를 통해 실시간 보안이벤트 처리가 가능한 SIEM(Security Information and Event Management) 솔루션으로의 발전 가능성도 확인할 수 있었다.

II. 관련연구

2.1 보안관제 빅데이터 솔루션 활용 이유

보안관제에서 빅데이터 솔루션을 활용하는건 단순히 대용량 로그를 분석하는걸 의미 하지 않으며, 침해사고를 사전에 탐지하기 위해서는 운영 중인 모든 정보시스템에서 발생하는 로그를 연관성 있게 분석하여 간결하게 나타내야 한다 또한 효과적인 보안관제를 위해서 모든 정보시스템의 로그를 수집 및 저장하고, 분석하여 시각화할 수 있는 도구가 필요하다.

기존의 보안관제 솔루션은 RDBMS(Relational Database Management) 기반으로 구축되는 경우가 많았다. 하지만 RDBMS 는 대량의 데이터를 처리할 경우에 성능에 문제가 발생할 수 있고, 비정형 데이터에 대해 처리에 대해 한계가 있다. 빅데이터 솔루션은 기존의 키(Key)와 값(Value)들의 관계를 테이블화 하여 index 를 수행하는 RDBMS 와는 반대 구조인 inverted index 구조로 데이터를 토큰화 시켜 저장한다. 때문에 <표 1>과 같이 비정형 데이터에 대해서도 Full Text 검색 속도가 RDBMS 에 비해 매우 빠르다.[4,5]

Table 1. Comparison with DBMS and Big Data Engine

Type	RDBMS	Big Data Engine
Data Storage Method	Normalization	DeNormalization
Full Text Searching Speed	Slow	Fast
Semantic Search	Impossible	Possible
Join	Possible	Impossible
Update / Delete	Fast	Slow

2.2 관련 도구들

2.2.1 Splunk

Splunk 는 빅데이터 분석솔루션으로 현재 산업계에서 가장 많이 사용되는 솔루션이다. 시스템에서 발생하는 모든 데이터를 실시간으로 저장, 분류하고 한 곳에서 검색, 분석, 시각화하여 수집된 데이터에 대하여 가시성과 대응 능력을 극대화하는 대량 분산 처리 모델의 단일 플랫폼 소프트웨어이다. 통신, 정보보안, 금융, 의료, 교육, 비영리단체, 공공부분, 온라인서비스 등 다양한 분야에서 빅데이터 분석을 위해 활용 중이다.

Splunk 는 강력한 UI 를 지원하고 사용자가 원하는 UI 로 변경 가능하며 데이터 즉시 분석이 가능하다. Fortune 100 대 기업 중 92 명이 이용하고 있으며 그 외에 9,000 여 개 이상의 기업, 서비스공급자와 정부가 Splunk 솔루션을 이용하고 있다. 국내에서도 많은 기업과 공공기관에서 빅데이터 분석을 위해 Splunk 를 이용하고 있다. 2020 년에 Splunk 는 7 년 연속 '가트너 매직 쿼드런트' 의 SIEM(Security Information & Event Management) 부분에서 리더로 선정될 정도로 보안업계에서도 많이 활용되고 있다.

2.2.2 Hadoop

Hadoop 은 아파치 재단의 프로젝트로 빅데이터 처리 플랫폼으로 많이 알려져 있으며 Hadoop 은 아파치 재단의 다양한 서브 프로젝트들과 융합되어 하둡 에코 시스템이라는 명칭으로 불린다.

Hadoop 은 HDFS(Hadoop Distributed File System) 등을 이용하여 대용량 데이터를 분산저장하고, 저장된 데이터를 빠르게 처리 할 수 있으며 저사양 서버를 이용한 스토리지 구성도 가능하여 가격대비 뛰어난 효율을 보인다. 허나 Hadoop 은 다양한 프로젝트 들과 융합되어 시너지 높은 모델이 될 수 있지만 다양한 프로젝트 간의 호환성과 보안관제에서 중요한 시각화에 대한 약점이 있어 보안관제 솔루션으로는 적합하지 않다.[6]

2.2.3 Spark

Spark 는 2009 년 미국 버클리 대학에서 개발한 오픈소스 스프트웨어로 메모리를 활용하여 빅데이터를 저장하고 처리하기 때문에 Hadoop 의 처리 성능에 비해 약 30 배 이상 차이난다. 하지만 빅데이터를 분석하기 위하여 원천 데이터를 RDD 로 변경하여 메모리로 데이터를 처리하기 때문에 구축 비용이 매우 비싸기 때문에 서브 분석용도로만 사용하고 있다.

III. Elastic Stack 을 적용한 솔루션

3.1 Elastic Stack

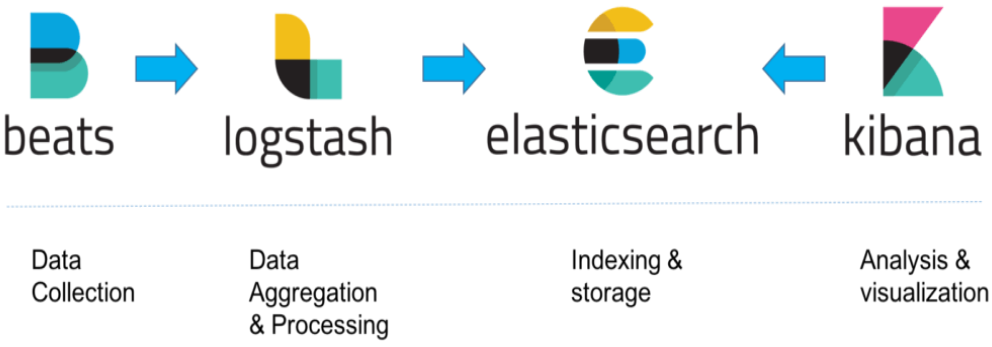

Figure 1. Configuration of Elastic Stack

Elastic Stack 은 <그림 1>처럼 Elasticsearch+Logstash+Kibana+filebeats 로 구성되어 있다. Elastic Stack 중 Elasticsearch 는 Apache Lucene 를 바탕으로 개발된 검색엔진 솔루션이며. Logstash 는 beats 등을 이용하여 수집한 각종 로그를 JSON 형태로 만들어 Elasticsearch 로 전송하는 역할을 하고 Kibana 는 Elasticsearch 에 저장된 Data 를 사용자에게 그래프, 테이블 등 시각화 형태로 보여주는 솔루션이다.

3.2 Elasticsearch

Elasticsearch 는 셰이 배논(Shay Banon)에 의해 Lucene 을 기반으로 만들어진 분산 검색 엔진이다. Lucene 은 더그 커팅 (Doug Cutting)이 개발했고 손쉽게 검색 기능을 추가할 수 있게 도와주는 자바 형태의 검색 라이브러리이다. 인덱스 저장, 관리 그리고 쿼리 문을 수행하여 결과값을 도출해 주고, JSON 기반의 정형, 반정형, 비정형 형태의 데이터를 검색하고, 분석하는데 사용되는 오픈소스이며, 분산 및 병렬처리, 실시간 검색 그리고 멀티테넌시를 지원하고 다양한 플러그인을 사용할 수 있는 특징이 있다.

1) 분산(Distributed) 및 확장성

Elasticsearch 는 규모가 수평적으로 늘어나도록 하게 설계되어 있으므로 더 많은 용량이 필요하면 그저 노드를 추가하고 클러스터가 인식할 수 있게 하여 추가적인 하드웨어로 이용할 수 있도록 해주면 된다.

2) 고가용성(High availability)

Elasticsearch 는 동작 중에 죽은 노드를 감지하고 삭제하며 사용자의 데이터가 안전하고 접근가능 하도록 유지한다. 즉, 동작 중에 일부 노드에 문제가 생기더라도 문제없이 서비스를 제공한다.

3) 멀티 태넌시(Multi-tenancy)

클러스터는 여러 개의 인덱스을 저장하고 관리할 수 있으며, 독립된 하나의 쿼리 혹은 그룹 쿼리로 여러 인덱스의 데이터를 검색할 수 있다.

4) 전문 검색(Full text search)

Elasticsearch 는 강력한 전문검색을 지원한다.

5) 문서 중심(Document Oriented)

Elasticsearch 는 복잡한 현실 세계의 요소들을 구조화된 JSON 문서 형식으로 저장한다. 모든 필드는 기본적으로 인덱싱되며, 모든 인덱스는 단일 쿼리로 빠르게 사용할 수 있다.

Elasticsearch 는 <그림 2>처럼 Cluster, Node, Shard, Replica, Gateway 로 구성된다.

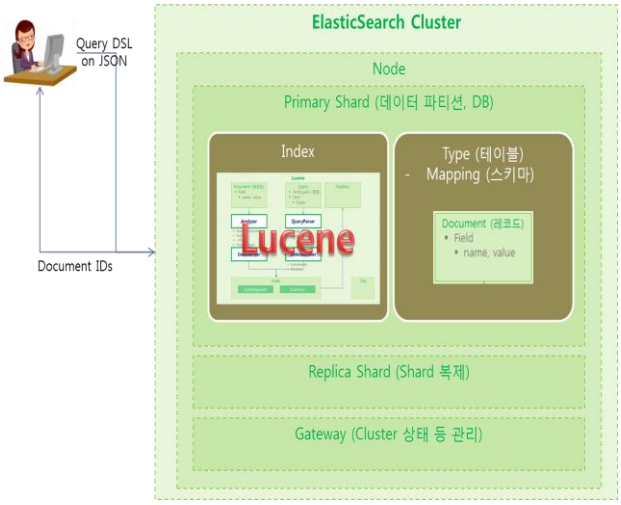

Figure 2. Conceptual Configuration of Elasticsearch

3.3 Logstash

Elasticsearch 는 뛰어난 검색엔진이지만 사용하려면 입력할 데이터를 JSON 형태로 가공해야 한다. Logstash 는 JRuby 로 만들어졌다. 로그 수집 및 가공을 위해 만들어졌으며 다양한 방식으로 데이터를 입력받아 Elasticsearch 로 전달하는 역할을 한다.

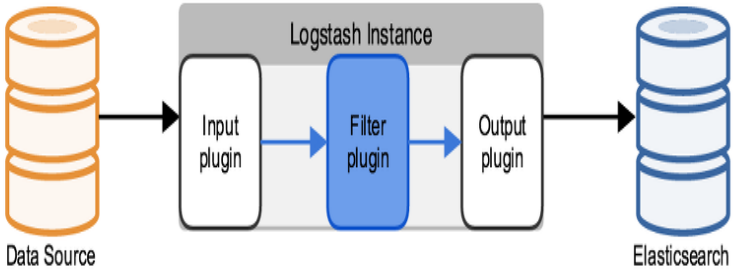

Figure 3. The Process of Logstash

Logstash 의 작동 과정은 <그림 3>처럼 크게 Input, Filter, Output 단계로 이뤄진다. Input 은 로그의 위치를 정의하고 데이터를 읽어온다. Filter 는 읽어온 데이터를 가공한다. Filter 를 이용하는 방법은 여러 가지가 있다. 정규표현식을 이용해 로그를 자르는 방법과 grok 을 이용해 미리 정의된 정규표현식을 사용하는 방법 CSV 를 이용해 데이터를 엑셀의 자르기 기능과 같이 일정한 패턴(콤마, 점) 등으로 자르는 방법과 XML 을 이용한 방법 등 다양하다. Output 에서는 필터링 된 보안 로그를 Elasticsearch 로 전달하는 역할을 한다.[7]

3.4 Kibana

<그림 4>는 Kibana 실행 화면이다. Kibana 는 Logstash 를 통해 Elasticsearch 에 모인 로그 데이터를 쉽게 검색하고, 다양한 시각화 분석을 할 수 있게 도와주는 솔루션이다.

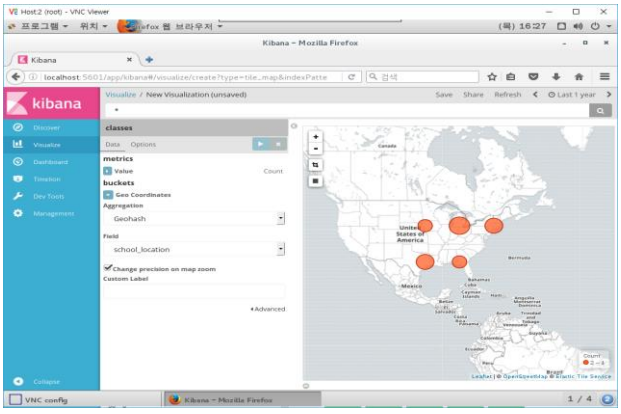

Figure 4. Screen Capture of Kibana

Kibana 의 주요 메뉴는 Discover, Visualize, Dashboard 로 나뉘어 있다.

1) Discover : IP, URL 등의 키워드를 사용할 수 있으며, 조회한 키워드를 저장했다가 나중에 다시 불러올 수도 있다. 또한, JSON 형태의 Elasticsearch 명령어를 직접 입력할 수도 있다.

2) Visualize : Elasticsearch 에 수집된 결과를 시각화 분석을 할 수 있다. 막대 그래프, Area chart, 테이블 등 여러 종류의 시각화 도구를 지원하고 있다.

3) Dashboard : Visualize 를 통해 시각화한 객체를 모아 하나의 Dashboard 에 배치하여 한눈에 확인할 수 있다.

IV. 빅데이터 솔루션 비교와 해석

4.1 Elastic Stack 와 Splunk 의 비교

1) 기능 : Elastic Stack 은 Splunk 모두 사용자 편의성에 맞춰 커스터마이징 가능하며 검색쿼리, 데이터 시각화(표, 차트, 대시보드 등)를 표현할 때 거의 동일한 기능이 가능하다.

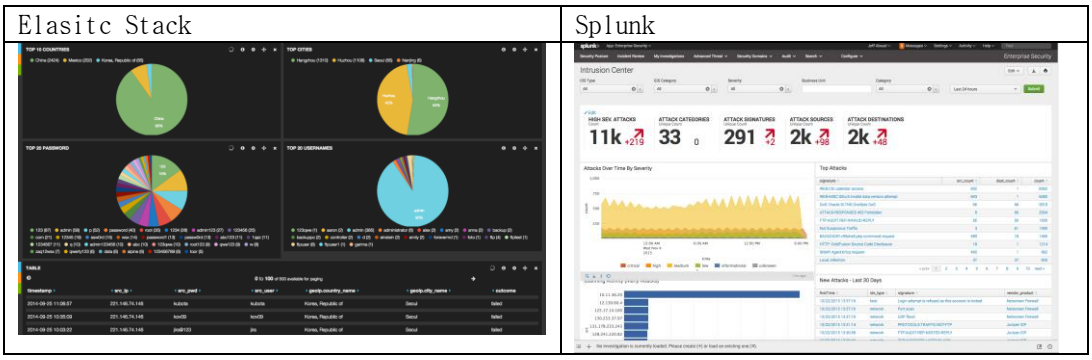

Figure 5. Comparison of visual Dashboard

2) 사용의 용이성 : 주관적인 판단 이지만 Splunk 는 완성형 솔루션인만큼 데이터 수집 및 분석과 분석데이터를 시각화 하는 기능이 Elasticsearch 에 비해 접근하기 쉬운 편이다. 하지만 Elasticsearch 는 AWS 등을 활용하는 클라우드 환경에서 서비스를 배포할 경우 강점을 보이고 있다.

3) 서비스 지원 : 빅데이터 솔루션을 활용하여 보안관제센터 구축시에 발생하는 문제를 해결할 수 있는 커뮤니티는 필요하다. 개발사가 아닌 이를 이용하는 사용자간에도 자유롭게 의견을 나눌 수 있는 커뮤니티 또한 중요한 요소이다. Elasticsearch 는 오픈소스 솔루션인 만큼 페이스북 한국그룹에 7400 여명의 회원들이 소통하고 있으며 공식 사이트에서 커뮤니티를 운영하고 있어, Splunk 에 비해 사용자 참여가 많은 편이다.

4) 가격 및 지원 : 오픈소스 프로젝트인 Elastic Stack 은 기본적인 구축에는 서버 비용 및 구축인건비 정도의 비용이 들어가며, 자체인력이 구축할 경우 서버 비용만 들어간다. 인공지능 및 클라우드 서비스를 이용할 경우 이에따른 제반 비용이 있지만 이는 Splunk 도 동일하며 Splunk 는 로그 용량별로 가격을 책정한다.

5) 기술지원 : Splunk 는 개발시에 자주 사용되는 언어용 SDK 를 제작하여 배포할 뿐만 아니라 200 개 이상의 RESTful API 를 제공하며 문서화 또한 훌륭하다. Elasticsearch 또한 RESTful API 를 제공하며 개발언어에 대하여 SDK 를 제공한다. 하지만 Splunk 는 API 를 자체적으로 개발하여 확장하기에는 완성형 솔루션이라 어려운 편이지만 Elasticsearch 는 오픈소스 솔루션이므로 다양한 맞춤형 APP 을 개발할 수 있다.

4.2 검색 성능 비교

빅데이터 기반 로그 솔루션 2 개의 보안관제에서 검색 성능을 비교하였다. 빅데이터라고 하기엔 데이터가 적지만 최대 1 억건(40GB) 정도의 방화벽 로그데이터 Elasticsearch, Splunk 각각 index 작업을 거친 후 특정 IP 를 Full Text 검색 하였다. <표 2>와 같이 검색 결과가 미세하지만 Splunk 의 검색속도가 우세하였다.[8,9,10]

Table 2. Comparison of Searching Performance

ES Respons Time(ms)	Splunk Response Time(ms)	Volume(GB)
20	20	5
38	40	10
52	48	15
77	75	20
98	92	25
112	108	30
139	128	35
157	146	40

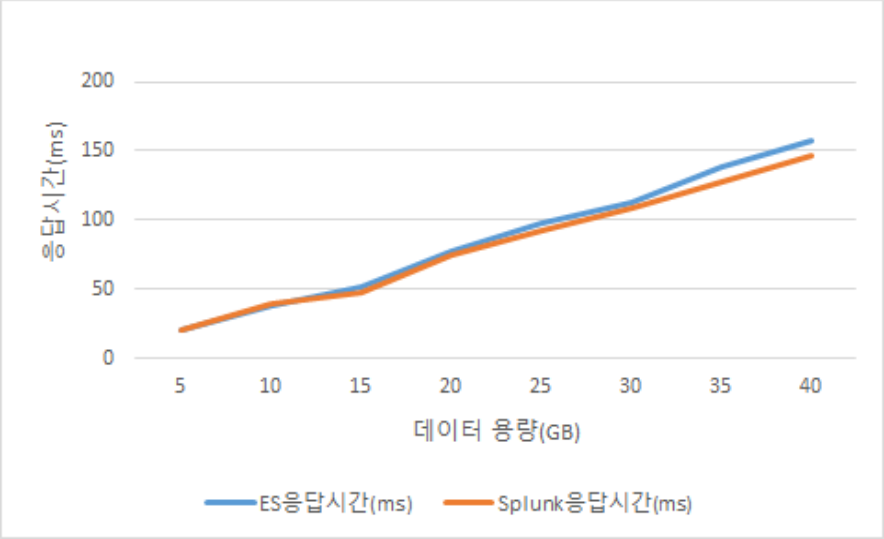

Figure 6. Comparison of visual Dashboard

V. 결론

최근 기업에서는 지능화되고 지속화하는 사이버공격에 효과적으로 대응하기 위하여 다양한 보안솔루션들을 도입 · 운영하고 있으며 이와 함께 오픈소스를 활용하거나 커뮤니티 관점에서도 보안관제 정책을 수립하고 있다. 기존의 상용화된 보안솔루션 제품들과 빅데이터 분석 솔루션들은 도입 시 비용을 지불하여야 하며 추후 증설 및 확장 개발 등이 필요할 때 제한적이고 솔루션 업체에 의존할 수 밖에 없다. 본 논문은 오픈소스의 활용, 사용의 용이성, 가격 및 지원, 커뮤니티 등을 활용하여 보안관제센터를 구축할 수 있다는 점에서 의미가 있다. 본 논문은 Elastic Stack 과 Splunk 의 두 솔루션의 기능, 사용의 용이성, 서비스 지원, 기술 지원 등을 비교 · 해석한 결과 Elastic Stack 이 사용자간 커뮤니티, 오픈 솔루션 면에서 다양한 보안솔루션들에서 수집되는 빅데이터 보안관제가 유리함을 알 수 있었다. 따라서 Elastic Stack 을 활용해 보안로그를 단계별로 수집-분석-시각화하여 대시보드를 만들고 대용량 로그를 입력 후 보안관제 검색속도를 측정하였다. 그 결과 보안관제 구축 시 오픈소스활용, 보안 검색 성능과 커뮤니티 등의 요소가 정보, 공공기관, 대기업뿐만 아닌 일반 중소기업들도 적합한 솔루션으로 선택하고 활용하여 구축할 수 있다.

후속 연구를 통해 머신러닝, 딥러닝 등을 활용한 보안 이벤트 분석을 통하여 사이버 침해사고 발생 징후에 대하여 사전 대처 등이 필요하다. 향후 사고 발생시 자동으로 대응할 수 있는 방안 등이 제시될 필요성이 있다.

VI. 참고문헌

[1] KISA, “2019 Personal Information Status Check Issues and Plans”, G-PRIVACY 2019. Apr. 2019
[2] Dailysecu, “Splunk selected as a leader in SIEM for 7 consecutive years,”(2020, Mar.), Available: https://www.dailysecu.com/news/articleView.html?idxno=107058”.Mar. 2020
[3] Yoo Ki-soon, Lim Sul-hwa, and Kim Hak-beom "Technology Trend and Development Direction of Integrated Log Management System", Journal of the Korea Information Security Society Vol. 23, No. 6 2013, p. 95
[4] Hanbitmedia, “Network security system construction and security control”,2016, pp. 38-42
[5] Infothebooks, “Security Control Practice Guide for Nurturing Next-Generation Information Security Talents”, 2017, pp. 45-48.
[6] Wikibooks, “ Start Hadoop programming”, 2014, pp. 320 – 328.
[7] Sang-Yong Lee, "Security log analysis system using log stash based on Apache Elasticsearch," Master's thesis, Daejeon University, 2016.
[8] Grartner, “Elastic vs Splunk”, Availble: https://www.gartner.com/reviews/market/security-information-event-management/compare/elasticsearch-vs-splunk, Dec. 2019.
[9] Hyun Jung-hoon, “Implementation of security control through analysis of information protection big data using open source ELK stack”, 2017.
[10] Hanbitmedia “Implementation and analysis of data visualization”, 2016, pp. 82-85.

저자 소개

전상준(*Jeon Sang June*)

2002 년 8 월 충남대학교 컴퓨터공학과 학사
2003 년 9 월 ~ 현재 한국지질자원연구원 지식정보실장
2015 년 2 월 한밭대학교 대학원 컴퓨터공학과 석사
2018 년 3 월 한밭대학교 대학원 컴퓨터공학과 박사과정

관심분야 : 정보보안, 네트워크

윤성열(*Yun Seong Yul*)

2018 년 2 월 한밭대학교 학사 졸업
2020 년 8 월 한밭대학교 석사 졸업
2017 년 2 월 ~ 2019 년 3 월 이글루시큐리티 침해대응본부
2019 년 3 월 ~ 현재 한국인터넷진흥원 주임연구원

관심분야 : 정보보안, 통합로그관리/분석

김정호(Kim Jeong Ho)

1994. 2. 단국대학교 대학원 공학박사
1983.2 경북대학교 공학석사
1983.3-1996.2 한국전자통신연구원
1996-.3- 현재 한밭대학교 컴퓨터공학과 교수
1989.8 정보처리기술사
1990.12 정보통신기술사

관심분야 : 네트워크와 데이터통신, 정보보호, 사물인터넷