JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 1, pp. 181-191, Jan. 2018

오픈소스 ELK Stack 활용 정보보호 빅데이터 분석을 통한 보안관제 구현

현정훈 · 김형중*
고려대학교 빅데이터 응용 및 보안학과

Security Operation Implementation through Big Data Analysis by Using Open Source ELK Stack

Jeong-Hoon Hyun · Hyoung-Joong Kim*
Department of Big Data Application and Security, Korea University

[요 약]

IT발전과 함께 해킹 범죄는 지능화, 정교화 되고 있다. 침해대응에 있어 빅데이터 분석이란 정보보호 시스템에서 발생하는 정상로그 등 전체 로그를 수집, 저장, 분석 및 시각화하여 이상행위와 같은 특이점을 도출하는 것이다. 기존에 간과해왔던 데이터를 포함하는 전수 로그를 활용하여 사이버 침해의 초기단계에서부터 침해에 대한 이상 징후를 탐지 및 대응하고자 한다. 정보보호 시스템과 단말 및 서버 등에서 발생하는 비정형에 가까운 빅데이터를 분석하기 위해서 오픈소스 기술을 사용하였다. ELK Stack 오픈소스를 사용한다는 점은 해당 기관의 자체 인력으로 기업 환경에 최적화된 정보보호 관제 체계를 구축하는 것이다. 고가의 상용 데이터 통합 분석 솔루션에 의존할 필요가 없으며, 자체 인력으로 직접 정보보호 관제 체계를 구현함으로써 침해대응의 기술 노하우 축적이 가능하다.

[Abstract]

With the development of IT, hacking crimes are becoming intelligent and refined. In Emergency response, Big data analysis in information security is to derive problems such as abnormal behavior through collecting, storing, analyzing and visualizing whole log including normal log generated from various information protection system. By using the full log data, including data we have been overlooked, we seek to detect and respond to the abnormal signs of the cyber attack from the early stage of the cyber attack. We used open-source ELK Stack technology to analyze big data like unstructured data that occur in information protection system, terminal and server. By using this technology, we can make it possible to build an information security control system that is optimized for the business environment with its own staff and technology. It is not necessary to rely on high-cost data analysis solution, and it is possible to accumulate technologies to defend from cyber attacks by implementing protection control system directly with its own manpower.

색인어 : 침해대응, 비정형, 시각화, 오픈소스, 엘크스택
Key word : Emergency Response, Unstructured, Visualize, Open-source, ELK Stack

http://dx.doi.org/10.9728/dcs.2018.19.1.181

Received 21 November 2017 ; **Revised** 23 January 2018
Accepted 29 January 2018

***Corresponding Author;** Hyoung-Joong Kim
Tel: +82-02-3290-4895, 010-6251-6343
E-mail: dcs@naver.com

Ⅰ. 서 론

IT환경의 급격한 변화와 함께 잦은 사이버 침해사고가 발생하고 있다. 해킹기술 및 공격유형 역시 빠르게 진화하고 있기 때문이다. 해커들은 끊임없이 해킹을 하고자 한 목표에 대해서 일반적으로 드러나지 않은 취약점을 찾는다. 이렇게 침해 행위는 수주, 수개월 동안 꾸준히 취약한 부분을 찾아내기 위한 행위가 우선으로 진행되며, 일단 취약점이 노출될 경우 그 취약점이 조치되기 전 시스템을 파괴 하거나 아니면 조용히 필요한 모든 중요 정보를 탈취한다. 대부분의 기업들은 보안이 필요한 영역에 DDoS대응, 침입차단, 침입탐지 등 적절한 정보보호 시스템을 구축 운영하고 있다. 다만, 보안 정책에 따른 허용(Permit)/거부(Deny), 알려진 침해 패턴을 탐지하고 대응하는 사후 대응에 가까운 수준이다. 따라서 본문에서는 패턴위주의 사후대응 체계가 아닌 실시간으로 발생하는 로그를 바탕으로 행위위주 탐지 및 대응 체계를 구현했다.

정보보호 시스템에서는 매순간 많은 양의 로그 데이터가 발생한다. 그것도 각 시스템별로 다양하고 비정형에 가까운 형태로 발생하고 있으나, 시스템 운영자들은 이벤트성 데이터에만 주의를 기울이고 그 외의 데이터는 간과하고 있는 것이 현실이다.

일반적으로 대부분의 침해성 데이터는 정상적인 형태로 가장하여 접속하기 마련이다. 침해 패턴을 우회하는 접속은 악성코드를 내려 받게 만드는 CnC(Command and Control) 서버로의 접속을 유도하고 CnC 서버와 접속되는 순간 악성코드가 다운로드 되면서 정보탈취 등 일련의 해킹이 이루어진다. 이러한 과정에서 정보보호시스템이 대응하는 구간은 어디가 적절할지 운영자는 잘 알고 있다. 일반적으로 인터넷 인접구간은 DDoS차단, 침입차단, 침입탐지 등 필수 보안장비로 대응이 이루어지며, 내부구간은 망분리와 함께 접근통제 시스템과 같은 보안장비로 내부 중요시스템을 침해로부터 보호하고 있다. 그럼에도 불구하고 일부 악성코드의 경우는 빈번히 내·외부 정보보호 시스템을 모두 통과하여 사용자 단말에 다운로드 된 후 악성행위를 한다. 다행이 알려진 패턴이면 탐지 및 조치가 되지만 해당 패턴이 없다면 악성코드에 감염된 사실을 인식하지 못하고 중요정보는 탈취 당한다. 이러한 이유에서 빅데이터 분석을 통한 보안관제는 접속시도부터 패킷이 접속해서 이동하는 과정에서 발생하는 다양한 형태의 동작 로그 데이터를 분석 및 시각화로 이상행위를 탐지했다.

그림 1에서 DDoS 차단시스템의 경우는 주로 DDoS 공격 패턴에 대해 오탐(False-Positive)이 고려된 임계값 정책을 적용한다[1][2]. 임계값 정책이란 DDoS 공격을 차단을 목적으로 사전에 설정해 둔 트래픽 허용한도 이상의 트래픽이 발생 시 차단이 적용되는 정책이다. 이러한 정책을 적용한 경우 용량이 작은 트래픽을 발생시키는 DDoS 공격은 차단되지 않으며, 정책적으로 허용 된다. 이렇게 허용된 공격이 웹서버나 중요한 내부서버에 영향을 주는 경우도 간혹 발생한다.

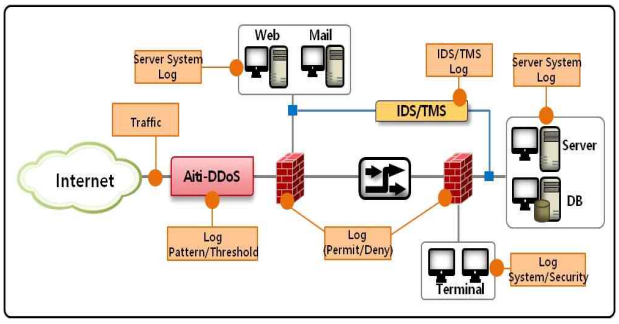


그림 1. 정보보호시스템 구간별 로그 데이터[3]
Fig. 1. Log Data by Information Security System Section [3]

침입차단 시스템인 방화벽도 비슷하다. 방화벽은 허용/거부 정책으로 운영되는 네트워크 보안시스템이다. 방화벽으로 분리되는 비 신뢰구역 즉 비무장지대(DMZ)에 위치하는 웹서버의 경우 모든 출발지에서 해당 웹서버로 80/tcp, 8080/tcp, 443/tcp에 대해 Any Open 정책이 적용된다. 물론, 대외적으로 Any Open 정책 적용 이전 모의해킹 등 보안성 검토 절차가 마무리된 이후 오픈하는 것이 정식 절차이다. 이렇게 정상적으로 오픈된 포트로 불특정 다수의 사용자가 접근 가능하게 되며, 웹서버 취약점을 발견할 경우 권한 탈취 등 침해행위가 이루어진다. 모든 침해행위의 출발점은 정상접속이다. 출발지에서 목적지까지 도착하는 접근 경로에 구축된 다양한 정보보호 시스템에서 발생하는 정상로그를 포함한 전체 로그 데이터를 실시간으로 수집하여 연관 분석을 통해 침해에 대한 사전징후를 찾는 것이 이 연구의 목적이다.

Ⅱ. 본 론

2-1 침해대응 관제 체계

사이버 침해 발생 주된 요인으로는 IT발전에 있다. 자금력을 바탕으로 AWS(Amazon Web Services)와 같은 클라우드 인프라를 이용한 조직적이고 지능화된 사이버 공격이 이루어지고 있다. 여기에 기업의 개방적인 비즈니스 추진으로 시스템적으로 다양한 요구사항의 증가와 여러 분야의 협업으로 인한 취약한 부분이 쉽게 노출된다. 또한 제도적인 면으로는 복잡한 규제를 지키면서 사업을 추진하는 것이 어렵다는 이유로 규정준수를 소홀히 함으로써 발생하는 취약점과 보안관련 담당부서나 기관에서 규정이 잘 준수되고 관리되는지 여부를 모니터링하는 어려움도 있다. 보안 취약점은 끊임없이 발생하고 있다. 취약점을 없애고 근본적으로 현재의 침해대응 체계를 개선하기 위해서는 패턴위주의 사후대응 체계를 탈피해야 한다[9].

오늘날 대부분 침해대응 체계는 개별 정보보호시스템에서 발생하는 이벤트 로그 모니터링을 바탕으로 이루어진다.

표 1. 현재 네트워크 및 정보보호 모니터링 체계
Table 1. Network and Information Security Monitoring System (As-Is)

Index		Monitoring System	Etc.
Network		Network monitoring system	Traffic, Devices
Info Security	DDoS	Dedicated management system	Effective control of cyber threats with the operation of information protection integrated monitoring system
	Firewall	Dedicated management system	
	IDS/IPS	Dedicated management system	
	Spam	Dedicated management system with limited monitoring function	
	Server	Using monitoring tools like server availability	
	Terminal	Antivirus and using separated terminal management solution	

내부에 보호해야할 시스템을 보유하고 있고 인터넷을 사용하는 기업이면 DDoS차단, 방화벽, 침입탐지 시스템은 기본적으로 갖추고 운영해야한다. 위 표 1을 보면 최소 3대 이상 관리시스템을 이용해야 하고 정보보호 시스템 운영인력도 가용성을 고려하여 배치해야 한다. 정보보호 시스템을 최소로 운영하더라도 APT(Advanced Persistent Threat)공격과 같은 기존 보안시스템으로 대응이 불가능하다면 추가로 도입, 구축 및 운영해야한다[4]. 이런 식으로 새로운 침해 패턴이 나타날 때마다 대응 시스템을 구축한다면 그것을 운영하는 인력의 추가 배치도 필요하다. 운영 인력 측면만을 고려한다면, 정보보호 통합관리 시스템 운영은 분명히 효과적이다[5]. 그러나 침해대응은 효과적으로 인력을 운영하는 것으로는 부족하다. 패턴에 의존하는 사후 대응체계 유지하는 것은 적절치 않다. 사전대응 체계로의 전환이 절실히 요구되는 시점이며, 정보보호 분야에서도 빅데이터 기술을 활용하여 정보보호 시스템에서 발생하는 전수 로그 데이터를 수집 및 상호 연관성 분석을 통하여 이상행위를 탐지해야 한다[6][7][8].

2012년에 Verizon사에서 웹상에 발표한 데이터 유출사고 조사보고서에는 보안 침해 사건의 91%가 단 며칠 만에 중요 데이터의 탈취로 이어졌고, 보안 침해 사건의 79%는 발견되기까지 수주 이상 소요되는 것으로 확인 되었다[9][10]. 그래서 오늘날 침해대응은 데이터 탈취나 시스템에 위해를 가하기 전에 탐지하고 대응하는 체계를 갖추어야 한다.

2-2 빅데이터 이용 침해대응 관제

공공기관, 정부 및 대다수 기업들은 심화되는 보안위협에 대응하기 위해 내·외부 보안장비 및 규제를 강화하고 있으나, 보안 담당자들의 노력에도 불구하고 효과적인 침해대응에 많은 어려움을 겪고 있는 실정이다.

빅데이터 기술의 등장으로 보안 전문가들은 다양한 통계기법과 고급 분석 기술을 이용하여 침해에 대한 사후 조치가 아닌 사전 대응을 가능하게 하는 예측 분석 시스템에 관심을 가지는 추세이다[6].

빅데이터 기술이 정보보호 침해대응 분야에 적용되기 위해서는 다양한 보안시스템에서 발생되는 각양각색의 대용량 로그를 시계열을 기준으로 연관분석이 가능해야 한다. 과거 축적된 데이터를 바탕으로 정상동작 패턴과 특이점을 가진 이상동작 패턴을 구별하여 데이터베이스로 축적해야한다. 인입되는 패킷이 목적지로 도달하는 과정에서 발생하는 로그로부터 정상의 범주를 벗어나는 특이점으로 구별해야 한다.

표 2. 빅데이터 침해대응 관제 요건[9]
Table 2. Requirement of Big Data Emergency Response Monitoring [9]

Infrastructure	Analysis Tool	Intelligence
▪ Flexibility through parallel processing (Scale-Out) ▪ Business support with cloud computing ▪ Build quickly and ensure scalability	▪ Provide visibility to ensure visualization ▪ Malware, Malicious code analyzer, etc. ▪ Utilizing accumulated statistical data ▪ Digital forensic analysis	▪ Prediction through historical data analysis ▪ Corelation analysis based on time series

표 2는 빅데이터를 이용한 침해대응 관제의 요건이며, 요건에 적합한 빅데이터 분석기술이 필요하다. 그래서 본 연구에서는 오픈소스 기술인 ELK Stack을 사용하였다. 분석기술로써 ELK Stack의 강점은 단순 반복적인 작업을 최소화할 수 있고 터미널 로그 데이터까지 분석에 반영하여 특이점을 보여줄 수 있다. 또한, 여러 시스템의 로그를 함께 보여줄 수 있어 연관분석이 가능하며, 검색엔진을 사용함으로써 시각화 반영이 빠르다. 이러한 강점을 이용하여 축적된 데이터와 실시간 데이터로 내재된 위협을 사전에 찾고자 한다.

정보보호 침해대응 관제 분야에서는 이상 징후의 사전파악 및 선제적 대응을 위해 빅데이터 분석 기술이 절실히 필요하다. 침해 행위의 단계를 따라가면서 공격자의 행동 패턴을 이해하고 적절히 대응하기 위해 어떤 데이터가 체인처럼 엮여 있는지 파악하고 신속한 조치를 할 수 있게 하는 침해대응 관제모델을 구축하고 운영해야 한다[9]. 침해대응 관제에서 가장 중요한 부분은 실시간 분석이 가능해야 한다[6]. 시스템 자원의 고성능화로는 이것을 구현하기에는 한계가 있으며, Scale-Up이 아닌 분산 컴퓨팅 방식인 Scale-Out 방식으로 빅데이터 분석에 접근해야한다. 분산 컴퓨팅 방식의 병렬처리로 각 정보보호 시스템에서 발생되는 로그를 실시간으로 분석할 수 있게 수집, 처리하고 저장해야 한다[15].

침해대응 빅데이터 분석은 비교 지표라고 하는 명확한 기준이 있어야 한다. 기준은 과거에 축적된 데이터에서 근거를 찾을 수 있다. 과거의 여러 침해사례에 대한 유형 분석, 악성코드의 행동 분석을 통해 최초에 시작점이 어디인지 어떤 행위로부터 일어나는 가를 파악해야 하고 이것을 데이터베이스로 축적함으로써 판단의 근거가 되는 지표를 만들 수 있다[11].

빅데이터를 이용 침해대응 관제는 정보보호 시스템을 포함한 종단간(End-to-End) 통합 모니터링 체계를 구축해야한다. 출발지 IP나 URL에서 목적지가 되는 내부 웹서버나 인터넷 단말까지 한눈에 모니터링 함으로써 인입 패킷의 초기단계부터 동작을 바탕으로 한 특이점이 발견하는 순간 집중 모니터링 및 사전대응이 가능하다[11].

2-3 오픈소스 기술을 이용한 정보보호 빅데이터 분석

오픈소스 이용의 장점은 첫째, 기업이 보유하고 있는 데이터를 바탕으로 최적화된 침해대응 관제를 직접 구현함으로써 기술 및 운영노하우를 축적할 수 있다. 둘째, 고가의 상용 정보보호 관제 솔루션 도입의 부담을 줄일 수 있다. 셋째 오픈소스 특성상 글로벌하게 활성화된 기술 커뮤니티에 참여하여 기술공유와 라이선스 제약 없이 사용이 자유롭다는 것이다. 이러한 이유로 본 연구에서는 오픈소스를 사용하였다.

사용한 오픈소스 기술은 ElasticSearch, LogStash, Kibana 및 Beats로 구성된 ELK Stack이다. 정보보호 빅데이터를 수집, 저장, 검색, 분석 및 시각화를 구현할 수 있다. 본 연구에서는 다양한 침해 유형을 탐색하고 패턴을 수집하여 이상행위에 대한 다양한 시각화 지표를 도출하는 것에 활용하였다. 또한, 해킹을 위해 의도적으로 접근하는 접속 초기단계부터 모니터링 하는 것에도 활용하였다.

여기서는 오픈소스 ELK Stack에 대해 간단히 설명하겠다. 자세한 내용은 참고문헌[13]을 참고하길 바란다.

표 3. ELK Stack(오픈소스) 활용 빅데이터 분석 로직
Table 3. Big Data Analysis Logic by Using ELK Stack (Open-Source)

Big Data	Store	Analysis	Visualization
Log Data	LogStash	ElasticSearch	Kibana
▪ Target system ▪ Store data ▪ How to store	▪ Log refining ▪ Filtering ▪ Indexing	▪ Realtime analysis ▪ Correlation analysis ▪ Calculate indicators	▪ Visualize ▪ Deriving singularity ▪ Proactive ▪ Prediction

1) ElasticSearch

ElasticSearch는 분산 환경을 지원하는 JAVA기반의 오픈소스 검색엔진 라이브러리인 Lucene[1]을 기반으로 구현된 분산 검색엔진으로 설치와 서버 확장이 편리한 기술이다. ElasticSearch는 검색 기능뿐만 아니라 검색한 데이터를 집계할 수 있는 기능이 있어 분석 엔진으로도 활용이 가능하다.

본 연구에서는 검색, 집계 및 빠른 분석 기능을 정보보호 빅데이터 분석에 활용하였다. ElasticSearch 검색기술은 깃허브(Github), 이베이(E-Bay), 가디언(Gideon), 위키피디아(Wikipedia), 넷플릭스(Netflix) 등 많은 기업이 구축·운영하고 있다. RESTful API[2]를 지원하기 때문에 다양한 환경에서 사용이 가능한 분산 검색엔진으로 Scale-out 확장성과 JSON 문서 기반으로 안정성과 관리의 편리성을 제공한다[13].

2) LogStash

LogStash는 정보보호 시스템으로부터 발생된 데이터를 읽어 분석이 가능한 포맷으로 변환한 뒤 ElasticSearch의 데이터 저장소로 전달한다. LogStash는 데이터를 전달하기 전에 반드시 분석이 용이한 형태로 데이터를 정제하는 코딩 과정을 거쳐야 한다. 데이터 정제는 JSON 형태로 입력(input), 정제(Filter) 및 출력(output)을 선언해 주어야 한다. 다양한 플러그인을 지원하기 때문에 여러 기술과 함께 사용이 가능한 동적 데이터 수집/처리 파이프라인이라고 한다[13].

○ LogStash에 적용된 로그 정제 Script(TCP Dump, 방화벽)

```
input {
        file {
                path => "J:/UBuntu/logs/tcpdump/*.txt"
                start_position => "beginning"
                type => "tcpdump"
                codec => plain { charset => "EUC-KR" }
        }
}
filter {
        if [type] == "tcpdump"
        {
                grok {
                        match => [ "message", "%{TCPDUMP_LOG}" ]
                        patterns_dir => "C:/logstash-5.2.2/patterns/"
                        add_tag => ["network", "tcp_connection_started"]
                }
                mutate {
                        add_field => {"timestamp1" => "2017-11-10 %{timestamp}"}
                }
                date {
                        match => ["timestamp1", "YYYY-MM-dd HH:mm:ss.SSSSSS"]
                        target => "@timestamp1"
                        timezone => "UTC"
                }
        }
}
output {
        if "_grokparsefailure" in [tags] {
                file {
                        path => "C:/logstash-5.2.2/log/fail-%{type}-%{+YYYY.MM.dd}.log"
                }
        }
        if [type] == "tcpdump"
        {
                elasticsearch {
                        hosts => ["10.21.10.22:9200"]
                        index => "fw-tcpdump-%{+YYYY.MM.dd}"
                }
        }
        stdout { }
}
```

```
input {
        file {
                path => "/usr/local/fwlog/*.txt"
                start_position => "beginning"
                type => "snmg"
        }
}
filter {
        if [type] == "snmg"
        {
                csv {
                        separator => ","
                        columns => ["priority", "date1", "date2", "rule id", "srcip", "srcport", "protocol", "destip", "destport", "sendbps", "rcvbps", "duration",
"zone", "reason"]
                        convert => {
                                                "date1" => "date"
                                                "date2" => "date"
                                                "sendbps" => "integer"
                                                "rcvbps" => "integer"
                        }
                }
                date {
                        match => ["date1", "yyyy-MM-dd HH:mm:ss"]
                        target => "date1"
                        timezone => "UTC"
                }
                date {
                        match => ["date2", "yyyy-MM-dd HH:mm:ss"]
                        target => "date2"
                        timezone => "UTC"
                }
}
output {
        if "_grokparsefailure" in [tags] {
                        file {
                                        path => "/usr/local/logstash/logs/fail-%{type}-%{+YYYY.MM.dd}.log"
                        }
                elasticsearch {
                        hosts => ["localhost:9200"]
                        index => "snmg-%{+YYYY.MM.dd}"
                }
                stdout {
                }
}
```

3) Kibana

Kibana는 시각화를 위한 오픈소스 기술이다. 시계열 분석 등 다양한 그래프를 이용해서 가시성을 보장하고 ELK Stack

1) Lucene : JAVA로 개발된 오픈소스 정보검색(IR : Information Retrieval) 라이브러리

2) RESTful API : Open API가 지원하는 XML, JSON, RSS 타입으로 그 차체가 별도 해석도구 없이 바로 사용 가능하며 이러한 기본 개념을 REST하다고 하며, 이것을 충실히 지키는 API

을 전체적으로 구성 및 기능을 통합 관리한다. 시계열 데이터의 경우 분석한 결과를 시각화하여 각 데이터의 시계열적인 연관성을 찾아준다. Geo location[3]을 이용하여 지리적으로 특정위치의 분포도를 지도상에 시각화함으로써 데이터의 상관관계를 보여준다. 로그상태에서는 볼 수 없었던 특이점을 찾아내거나, 과거와 실시간으로 발생하는 로그를 동시에 시각적으로 보여주기 때문에 누적된 과거 데이터로부터 얻은 추이를 바탕으로 실시간 데이터 분석으로 미래를 예측하게 해준다[13].

4) Beats

Beats는 로그 분석을 목적으로 PC 혹은 서버에 에이전트로 설치되어, End단에서 발생하는 시스템 로그를 비롯한 각종 이벤트성 데이터를 LogStash와 ElasticSearch로 전달하는 기능을 한다. 오프소스로 가볍게 개발된 데이터 수집 및 전송용 플랫폼이며, 클라이언트 에이전트이다[13].

2-4 사이버 침해 사전예측 관제 연구 모델

최근 사회적 이슈가 된 WannaCry 랜섬웨어와 같은 새로운 악성코드의 유포 등 지능화된 해킹 기술이 등장하고 있다. 중요 자산을 보호하기 위해 다양한 정보보호 시스템을 운영하고 있으나 여전히 데이터 유출과 같은 침해사고는 계속해서 발생한다. 알려진 패턴을 탐지 대응하는 체계는 이미 한계가 있다. 알려진 패턴이 아닌 실제 데이터가 보여주는 현상을 파악하고 발생 가능성이 있는 침해유형을 찾아야 한다.

IP환경에서는 정상 혹은 비정상 여부에 관계없이 접속의 시도부터 접속이 이루어지는 과정은 빠짐없이 관련 시스템에 로그로 남는다. 접속세션이 이루어지는 구간에 구축된 각종 정보보호시스템에서 생산되는 로그와 트래픽, 단말 등 관련 데이터를 이용하여 접속 시도 초기부터 종료까지 발생되는 데이터를 전수분석 해야 한다. 전수분석을 통해 접속에 대한 동작 패턴을 파악한 후 정상과 비정상 동작 패턴의 차이점을 인지해야 한다. 침해 패턴의 결과가 아닌 동작 패턴을 이용하여 침해유형을 탐지하는 것이 본 연구의 목표이다.

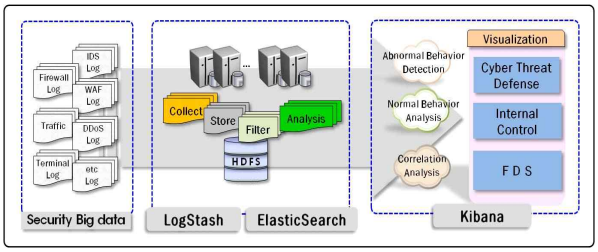


그림 2. 빅데이터 활용 정보보호 침해대응 관제 연구모델
Fig. 2. The Study Model of Information Security Emergency Response Monitoring by Using Big Data

3) Geo Location : 유무선 통신망에 연결된 휴대전화, 컴퓨터 등 기기의 지리적 위치정보

침해대응 관제 연구모델은 정보보호 시스템에서 발생되는 빅데이터를 활용, 분석을 통해 비정상 동작 패턴을 탐지하여 침해사고를 사전에 대비하고자 한다[11].

2-5 사이버 침해 사전예측 관제 연구 내용

침해대응 관제 체계를 구축하기 위한 연구 내용으로, 현재의 체계와 빅데이터를 이용한 체계의 차별점을 명확히 도출해야 한다. 이 연구를 위해 PoC(Proof of Concept) 인프라 구축을 통해 발생하는 데이터 규모, 분석 내용 등 면밀히 검토하여 침해대응의 방향을 구체적으로 설정해야 한다.

1) 침해대응 보안관제 체계 변화

표 4. 침해대응 보안관제 체계 비교
Table 4. Comparison of Emergency Response Security Monitoring Management

AS-IS	TO-BE
▪ Using the monitoring system provided by each security system ▪ Pattern based detection through emergency response systems ▪ Limit of response to bypassing attacks against existing pattern such as new and APT attacks ▪ Limit of detection abnormal signs first and response	▪ Big Data correlation analysis system against cyber threats ▪ Trace tracking and behavior-based detection of attackers through log analysis ▪ Build of prediction defense proactive response system

본 연구 모델의 전반적인 흐름은 전수 데이터를 수집, 저장, 분석하여 공격 가능성이 있는 특이점을 파악한 후, 초기 접속부터 결과까지 행동패턴을 수집한다. 분석을 통해 비정상 행위를 판별하여 해당 행동패턴을 데이터베이스화 한다. 향후 유사한 특이점이 발생할 경우, 침해의 가능성을 사전에 예측하고 대응 하는 것이다[11][14].

2) 침해대응 분석 메카니즘

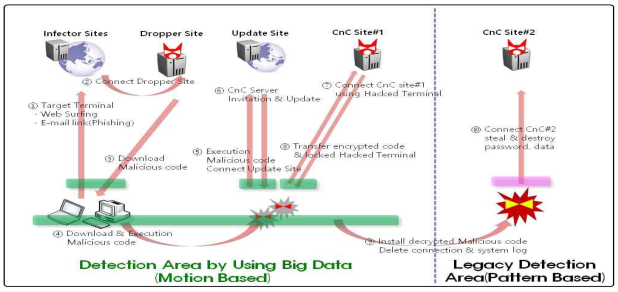


그림 3. 악성코드 라이프 사이클과 빅데이터 활용[11]
Fig. 3. Malicious Code Life Cycle and Use of Big Data [11]

그림 3에서 보여주는 라이프 사이클 상에서 실질적인 정보탈취와 같은 해킹이 이루어지는 단계는 ⑩단계까지 성공을 해야만 한다. 단계적으로 남겨지는 로그 데이터를 분석해 ① 단계부터 ⑩단계까지의 접속과정에서 CnC 서버 접속, 악성코드 다운로드와 같은 로그는 반드시 존재하고 이러한 특이

점을 찾기 위해서 전수 데이터를 분석하는 것이다. 악성행위가 ⑩단계까지 도달하기 전에 찾는다면 이것이 사전예측으로 이어지고 사전대응은 가능하다[7][8].

3) 침해대응 분석을 위한 빅데이터 규모

표 5는 연구대상인 빅데이터 규모이나, PoC 환경에서 전수분석은 시스템적으로 어려워 인터넷 방화벽, 웹방화벽 로그는 표 5의 로그를 사용하였고, 단말 및 서버 로그는 PoC 환경에서 발생하는 로그를 활용하였다.

표 5. 연구모델 빅데이터 분석 규모
Table 5. The Size of Big Data Study Model

Index	Data Size
Firewall	▪ 2 Gbytes/day (Permit/Deny Log), 4 Firewalls
Traffic	▪ Usage of All Internet PC : 700 Gbytes/day
PC number	▪ 1,100 PCs, 2.2Tbytes/day
Web	▪ 100 URLs

4) 침해대응 빅데이터 분석 아키텍처

그림 4 PoC구성을 통해 저장된 로그 데이터를 정제, 분석 및 시각화하여 정상과 이상 행위를 구별하였다. 또한 방화벽 및 단말의 경우 실시간 저장 및 분석이 가능하다.

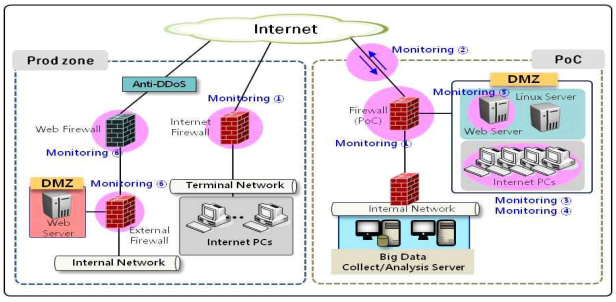


그림 4. 정보보호시스템 빅데이터 분석 PoC
Fig. 4. Information Security System Big Data Analysis PoC

본 연구에서 정보보호 빅데이터 분석 아키텍처는 누적된 로그 데이터를 활용하여 접속패턴을 데이터베이스화 하고 새로운 접속 패턴이 발생할 경우 축적된 데이터베이스와 유사도와 분석을 통해 침해 징후를 예측하는 것이다.

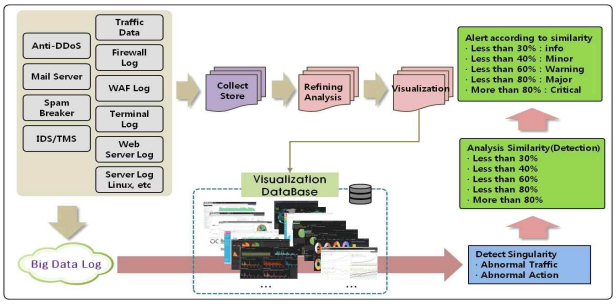


그림 5. 침해대응 빅데이터 분석 아키텍처
Fig. 5. Big Data Analysis Architecture for Emergency Response

2-6 사이버 침해 대응 관제 구현

ELK Stack을 활용하여 방화벽, 트래픽, 단말, 서버 로그를 1개월간 수집, 정제 및 분석한 후 시각화하여 침해대응 관제를 구현하였다. 이 결과로부터 도출할 수 있는 연구 성과로 산출된 시각화 지표를 이용하여 이상행위를 근거로 하여 사이버 침해징후를 사전예측 하였다.

1) 방화벽 로그 데이터 활용 보안관제

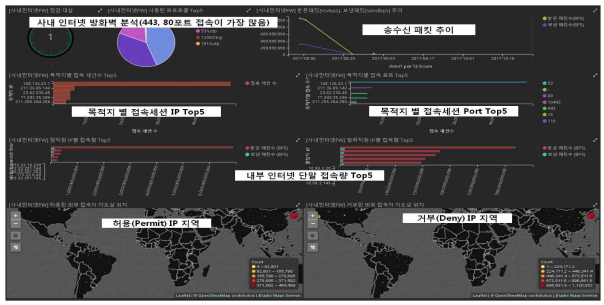


그림 6. 방화벽 로그 분석 시각화
Fig. 6. Visualization of Firewall Log Analysis

80/tcp, 443/tcp포트로 웹서버 접속이 가장 많음을 보여주며, 일별 접속 추이로 이상 접속을 확인하였다. 출발지 IP Top5로 다수의 접속을 발생시키는 IP를 파악하여 목적지 IP와 포트와 비교 정상접속 여부를 확인하였다. 또한, Geo Location 정보를 활용 국가별 허용/거부 IP를 탐지하여 유해 IP 데이터베이스와 비교 해외 CnC 서버 여부를 판별했다.

2) 방화벽 TCP Dump 정보를 이용한 트래픽 모니터링

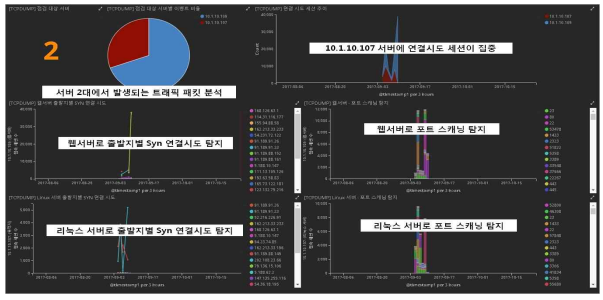


그림 7. 트래픽 모니터링 시각화
Fig. 7. Visualization of Traffic Monitoring

목적지 웹서버로 접근 내역을 보면 80/tcp, 443/tcp 외 많은 포트들로 접속한 내역이 탐지되었다. 그래프 상에서 튀는 구간과 동일 시간대에 유사하게 접속 포트의 튀는 구간을 탐지함으로써 해킹을 위한 초기단계로 취약한 포트가 오픈되어 있는지 파악하기 위한 포트 스캐닝 공격이 들어온 것을 확인하였다[12].

3) 인터넷 단말 이상행위 모니터링

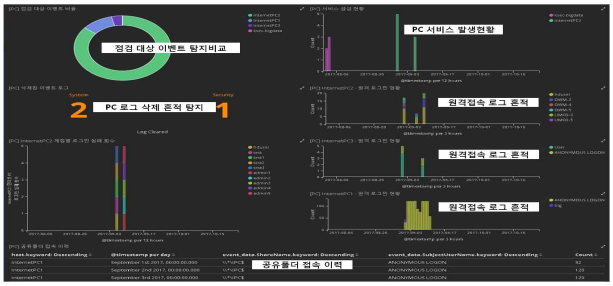


그림 8. 인터넷단말 로그 모니터링 시각화
Fig. 8. Visualization of Internet PC Log Monitoring

PC2에서 의심스런 서비스(프로세스)가 두 차례 3회 생성되었음을 탐지되었다. 악성코드는 서비스를 생성하고 스케줄에 따라 악성 행위가 이루어지도록 프로그래밍 된다. 다수의 서비스가 평소와 다르게 생성되면 감염을 의심할 수 있다.

해커들은 악성행위의 흔적을 감추기 위해 단말의 이벤트를 삭제한다. 이벤트 로그의 삭제행위가 3회 발생한 단말을 탐지했으며, 악성코드 감염을 의심할 수 있다. 또한, 원격접속 포트와 같이 취약점을 가진 포트를 중심으로 악성행위를 한다. 인터넷 단말에서 매일 익명(Anonymous)와 임의 계정으로 원격접속 로그인 시도와 공유폴더에 접근 시도를 탐지했다. 취약한 포트와 윈도우의 공유폴더의 취약점을 이용한 악성행위 시도로 볼 수 있다. 특정 포트로의 접속을 모니터링하는 것은 악성행위 탐지에 매우 중요하다. 내부 규정을 위반하고 업무 편의를 위해 원격접속 포트를 오픈한 단말을 탐지할 수 있다.

4) WannaCry 랜섬웨어 행위 분석

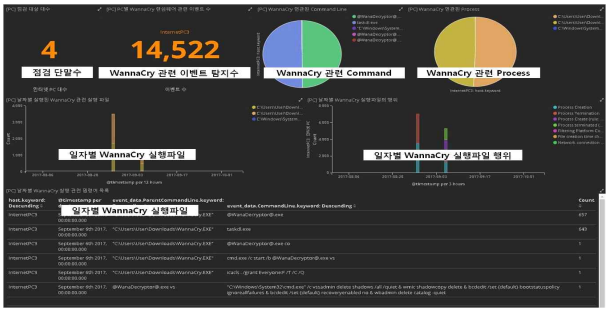


그림 9. 랜섬웨어(WannaCry) 행위분석 시각화
Fig. 9. Visualization of Ransomware Action Analysis

짧은 기간 동안 14,500개 이상 WannaCry 관련 이벤트가 탐지되었다. WannaCry가 최초 감염시 @WanaDecryptor@.exe와 taskdl.exe가 생성 및 가장 많이 실행되었고, @WanaDecryptor@.exe, taskdl.exe, taskkill.exe 프로세스가 일자별로 수백 번 이상 반복되어 실행됨을 확인하였다. WannaCry가 프로세스 생성, 종료 및 윈도우즈의 필터링 플랫폼에서의 차단, 네트워크 연결 등 행위를 탐지하였다. WannaCry 행동패턴을 보면 초기에는 WannaCry가 설치되어 있는 폴더 내 권한을 모든 계정이 접근할 수 있도록 everyone 권한을 부여함을 보여주고, 네트워크 통신을 가능하게 하는 토르 프로토콜 실행용으로 taskhsvc.exe를 실행하고 모든 행위를 관장하는 WannaDecryptor.exe를 657회, 암호화된 원본 파일을 삭제하는 taskdl.exe를 643회 실행하는 행위가 탐지되었다.

5) 웹서버 가용성 및 침해연관분석 모니터링

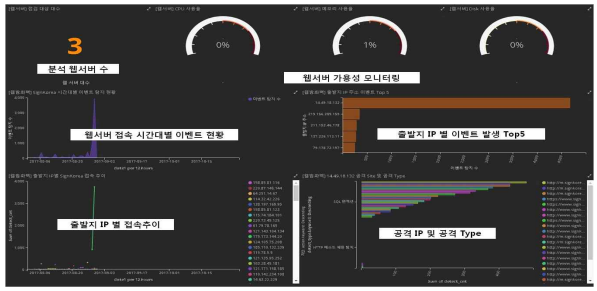


그림 10. 웹서버 가용성 분석 시각화
Fig. 10. Visualization of Web Service Availability Analysis

웹서버로 접속 시 방화벽에서 주고받은 패킷 추이를 확인한 결과 8/29 ~ 31일 동안 외부로 보낸 패킷이 수신한 패킷보다 평상시 추이와 비교해 상당히 많음을 확인하였다. 8/29일 최초로 해당 IP에서 80/tcp 포트로 평소보다 현격히 많은 접속을 탐지하였고, 동시에 방화벽 로그에서는 해당 IP에서 다수의 접속시도와 함께 Timeout이 발생, 접속 실패한 것을 확인하였다. 같은 시점에 웹방화벽 로그에서 특정IP (14.49.18.132)에서의 접근이 비정상적으로 많음을 탐지하였고 출발지 IP인 14.49.18.132에서의 접근 시도는 SQL Injection 공격시도임을 탐지하였다.

6) 유해 IP 탐지를 통한 보안관제 모니터링

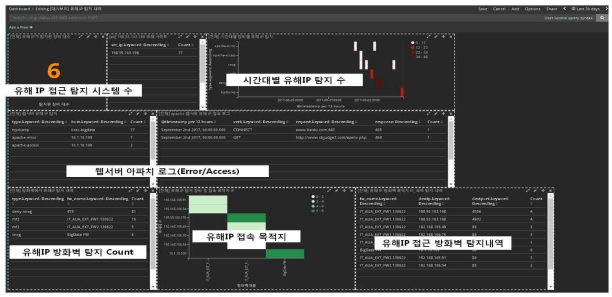


그림 11. 유해 IP를 통한 사이버 침해 모니터링 시각화
Fig. 11. Visualization of Cyber Threats through Harmful IP

유해 IP로 분류된 IP(198.55.103.198)에서 접근한 흔적이 방화벽과 웹서버에 로그에서 탐지되었다. PoC에 구축된 7대의 단말 및 서버로 접근이 확인되었고, Apache 접근, 에러, 인터넷 방화벽 허용/거부, PoC 방화벽 허용 로그 및 패킷 덤프에서 탐지되었다. 내부 시슬메을 경유하여 중국 사이트인 Baidu로 443/tcp 포트로 접속하려는 시도가 탐지되었고, CnC로 추정되는 서버로 접근 시도가 탐지되었다.

유해 IP로 공지된 내역을 바탕으로 내부 서버로 접속을 탐지했고, 내부 접근 서버의 침해여부를 신속히 파악할 수 있다.

2-7 기존 상용솔루션과 비교

지금까지 오픈소스 ELK Stack을 활용한 침해대응 보안 관제를 구현했다. PoC 환경에서 구현을 했지만 기존 상용솔루션과는 어떤 차이가 있는지 스펙 및 구현에 대한 특징 위주로 비교했다.

표 6. ELK Stack과 상용 보안관제 솔루션 비교
Table 6. Comparison of ELK Stack with Commercial Security Monitoring Solutions

Index	ELK Stack	ESM	Splunk
Hardware	▪ 5 ~10 PCs or Servers	▪ 5 ~ 10 Servers	▪ 5 ~ 10 Servers
Software	▪ ELK Stack	▪ Spider ▪ Database (SQL, Oracle, ..)	▪ Splunk ▪ Database (SQL, Oracle, …)
Cost	▪ 50 Million won ▪ No license fee	▪ 0.3~0.5 Billion won	▪ 0.3~0.5 Billion won
Etc.	▪ No restriction on use	▪ There is license restriction	▪ There is license restriction

표 6에서 하드웨어 및 소프트웨어 구성과 비용은 표 5에 보여준 데이터량을 근거로 산정한 결과이며 다소 구성방법에 따라서는 오차가 있을 수 있다.

2-8 연구 성과(사전 예측)

본 연구 성과로 시각화 지표를 통한 사전예측 결과를 1) 단말 악성코드, 2) 서버침해, 3) 유해IP 접속 수, 4) 사이버 침해시도 징후 탐지 이렇게 네 가지 항목으로 도출하였다. 각 항목마다 보여 지는 시각화 지표에서 탐지항목에 해당되는 건수가 많을수록 이상행위의 가능성이 높다고 할 수 있다. 대응 조치로 이상 트래픽을 유발시킨 출발지 IP에 대해 집중 모니터링 하거나 심각도에 따라 차단이 이뤄져야 한다. 또한, 시각화 지표(탐지항목의 종류와 탐지건수)가 다양하게 축적될수록 침해예측의 정확도는 향상될 수 있다.

1) 단말 악성코드 탐지예측 결과

Detect	8/2~3	8/30	9/1	9/4	9/6	9/9	9/11	...
Traffic					2	1	1	
Harmful IP			6	1	34		151	
Create Pro.	5	5			3			
Delete log						2		
Remote con			2		2	11		
Severity	minor	minor	major	minor	critical	critical	critical	...

그림 12. 단말 악성코드 탐지 예측결과
Fig. 12. Prediction Result of Terminal Malicious Code Detection

단말 악성코드 탐지 예측 모델은 다섯 가지 시각화 지표(순간 트래픽 발생 건수, 알려진 유해 IP 매칭건수, 동 시간 프로세스 기동건수, 로그삭제 건수, 원격 접속요청 건수)를 탐지항목으로 분류하여 일자별 탐지 데이터를 확인하였다. 단말 악성코드 모니터링 결과, 인입 트래픽에서 최초로 이상행위가 탐지되었다. 발생 시점(9/6, 9/9, 9/11)에서 사전예측이 가능하고 유해IP 및 로그삭제, 원격접속 지표에 추가적으로 탐지됨으로써 Critical로 경고 수준을 높이고, 즉시 사전 대응이 가능하다.

2) 서버 침해 탐지 예측 결과

Detect	8/26	8/27	8/30	9/1	9/4	9/6	9/9	...
Traffic						2	1	
Src Top						125	4061	
WAF			3944					
Hist con		29					28	
Web error	374			6	1	34	9962	
Severity	major	minor	critical	minor	info	major	critical	...

그림 13. 서버 사이버 침해 탐지 예측결과
Fig. 13. Prediction Result of Server Cyber Threats Detection

서버 침해 탐지 예측 모델은 다섯 가지 시각화 지표(순간 트래픽 발생건수, 다량 검출된 주요 출발지, 웹방화벽 탐지건수, 서버접속 로그, 웹서버 이벤트 로그)를 탐지항목으로 분류하여 일자별 탐지 데이터를 확인하였다. 서버 탐지 모니터링 결과 9/6일 과 9/9일 특정 IP에서 내부 중요서버로 집중되는 다량의 이상 트래픽이 탐지 되었다. 동시간대 서버 접속 로그 증가됨이 확인되었고, 웹방화벽의 탐지패턴을 함께 분석한 결과 추가 침해 징후를 탐지하였다. 이러한 탐지 결과를 활용하여 유사 트래픽 행위와 출발지 IP를 근거로 사전 대응을 할 수 있다.

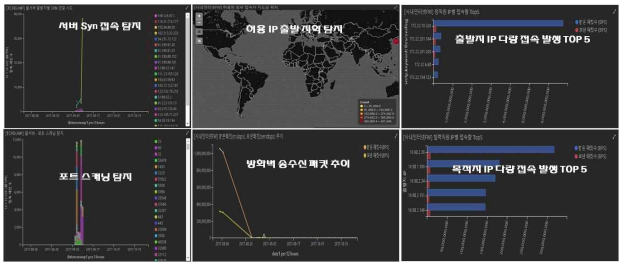
3) 유해 IP 접속 수에 따른 침해 예측 결과

Detect	8/2~3	8/30	9/1	9/4	9/6	9/9	9/11	...
Deny log								
Permit log								
Harmful IP			6	1	34		151	
Destination					2	1	1	
Severity	info	info	major	minor	critical	minor	critical	...

그림 14. 유해 IP 탐지를 통한 사이버 침해 예측결과
Fig. 14. Prediction Result of Cyber Threats through Harmful IP Detection

유해 IP 접속 수에 따른 탐지 예측 모델은 네 가지 시각화 지표(거부로그, 허용로그, 유해 IP 접속건수, 목적지로 유입 건수)를 탐지항목으로 분류하여 일자별 탐지 데이터를 확인하였다. 금융보안원과 KISA를 통해 공지되어 축적된 유해IP (최근 3년)를 바탕으로 외부에서 접근하는 IP를 비교 탐지한 결과 유해 IP 리스트에 매칭 되지 않아 인터넷 인입구간 방화벽에서 차단되지 않고 접속한 내부서버를 탐지하였다. 해당 탐지에 대한 조치로 우선적으로 인터넷 인입구간 방화벽에서 차단을 하였고 실시간으로 탐지된 접근 서버에 대한 신속한 침해여부의 점검을 수행하였다.

4) 사이버 침해 시도 징후 탐지 예측결과



Detect	8/2~3	8/30	9/5	9/6	9/9	9/10	9/11	...
Src Top5								
Dst Top5								
Global IP								
Syn conn				4288	3574	38167		
Port Scan			154	4060	1130	648		
Severity	info	info	minor	critical	critical	critical	info	...

그림 15. 사이버 침해 시도징후 탐지 예측결과
Fig. 15. Prediction Result of Cyber Threats Attempt Signs Detection

사이버 침해징후 탐지 예측 모델은 다섯 가지 시각화 지표(다량의 접속 출발지 IP, 다량의 접속 목적지 IP, 해외 IP, SYN only, 포트 스캐닝)를 탐지항목으로 분류하여 일자별 탐지 데이터를 확인하였다. 외부로부터 사이버 침해 초기단계로 취약점을 찾는 행위인 포트 스캐닝과 SYN Only와 같은 공격 패턴 탐지했다. 특정 IP에서 출발하는 이 같은 이상행위를 감지했으므로 즉시 해당 IP는 방화벽에서 차단 조치로 사전대응을 하였다.

Ⅲ. 결 론

위 여섯 개 부분으로 보안관제 모니터링의 결과를 바탕으로 침해징후 시각화 지표를 데이터베이스화 한 후 유입되는 패킷이 통과하는 구간별 시각화 지표를 적용하여 네 가지 항목으로 침해에 대한 사전예측을 하였다. 유입되는 패킷이 거치는 구간에 설치된 정보보호 시스템 로그를 분석 및 시각화하여 평소와 다른 이상징후를 탐지할 수 있었다. 현재 시스템에서 간과하고 있었던 각종 정보보호시스템에서 발생하는 정상 로그를 포함하여 전수 수집 및 분석함으로써 로그 자체로 특이점을 인지할 수 없었던 정보가 시각적으로 도출됨으로 사이버 침해 사전대응 관제가 가능함을 확인하였다.

빅데이터 분석기술을 이용한 상관분석 시스템이 많이 출시되고 있지만 주로 외산장비에 고비용의 투자가 필요하기 때문에 쉽게 도입을 결정할 수 없다. 도입을 한다고 하더라도 해당 기업에서 발생하는 데이터를 효과적으로 이용함에 있어 주도적인 침해대응 기술 확보가 어려우며, 제조사에 의존할 수밖에 없다[11].

ELK Stack과 같은 오픈소스 기술의 이용으로 고가의 상용 솔루션을 도입할 필요 없이 해당 기업에 최적화된 보안 관제를 구축함으로써 침해대응 기술 노하우까지 습득할 수 있다.

본 연구에서 보안관제에 빅데이터의 활용은 사이버 침해 사전대응이 가능하게 한다는 사실을 PoC를 통해 확인하였고, 기대효과는 아래 다섯 가지로 요약할 수 있다.

① 방화벽 허용/거부 로그, 트래픽 데이터 등 발생하는 데이터의 실시간 전수분석을 통해 내부 접근하는 이상행위 탐지 및 신속한 대응체계 구축 가능하다.
② 외부 공격에 가장 취약한 인터넷 단말은 백신만으로는 대응에 한계가 있으므로 단말에서 발생하는 행위 로그 분석과 프로세스 동작 등 동작기반 이상행위 탐지를 통해 APT와 같이 알려지지 않은 악성코드에 대응할 수 있는 기반을 마련할 수 있다.
③ 전 세계적으로 이슈가 된 WannaCry 랜섬웨어의 행위 분석을 통해 동작 패턴을 이해하고 향후 유사한 악성코드 발생 시 바이러스 백신에서 조치가 이루어지기 전 신속히 탐지 및 확산을 막을 수 있는 토대를 갖출 수 있다.

④ 공개된 서버(웹, 트레이딩) 서버의 실시간 가용성 및 정보보호시스템(방화벽, 웹방화벽, 침입탐지 등) 로그 데이터를 활용 시계열 상관분석으로 평상시와 다른 특이점 발생 시 집중 모니터링으로 외부로부터 유해행위 여부를 신속히 탐지하여 사전대응이 가능하다.

⑤ 오픈소스 기술은 공개된 기술커뮤니티를 통한 상호 다양한 기술공유가 가능하고 침해대응 보안관제에 적용할 경우 자체 데이터 및 운영인력을 활용하여 최적화된 사전대응 관제체계를 구현할 수 있다. 또한, 고가의 외산 침해대응 시스템 도입으로 낭비되는 비용을 절감할 수 있고, 이기종 시스템 접속에 대한 라이선스 문제도 없다.

참고문헌

[1] H. O. Koo, S. H. Baek, and C. S. Oh, “Effective traffic analysis in DDoS attack”, *Journal of the Korea Contents Society*, Vol. 2, No. 1, pp. 268-272, May 2004

[2] T. Y. Shim, I. J. Choi, J. I. Lee, B. K. Hong, and C. S. Oh, “Methodology for DDoS Detection Using Pattern Matching in Distributed Environment”, *Journal of the Korea Institute of Information Technology*, Vol. 11, No. 8, pp. 101-110, Jul. 2013

[3] M. Kaeo, *Designing Network Security*, 2nd ed. Cisco Press, pp. 343-353, Mar. 2004

[4] D. S. Moon, H. S. Lee, and I. K. Kim, “Host based Feature Description Method for Detecting APT Attack”, *Journal of the Korea Institute of Information Security & Cryptology*, Vol. 24, No. 5, pp. 839-850, Oct. 2014

[5] I. S. Jeon, K. H. Han, D. W. Kim, and J. Y. Choi, “Using the SIEM Software vulnerability detection model proposed”, *Journal of the Korea Institute of Information Security & Cryptology*, Vol. 24, No. 4, pp. 961-974, Aug. 2015

[6] S. J. Lee and D. H. Lee, “Real time predictive analytic system design and implementation using Big Data-log” *Journal of the Korea Institute of Information Security & Cryptology*, Vol. 25, No. 6, pp. 1399-1410, Dec. 2015

[7] D. K. Kim, S. B. Pyo, and C. H. Kim, “Study on APT Attack response Techniques Based on Big Data Analysis”, *The Society of Convergence Knowledge Transactions*, Vol. 4, No. 1, pp. 29-34, Jan. 2016

[8] J. S. Hong, Y. H. Lim, W. H. Park, and K. H. Kook, “Improved Security Monitoring and Control Using Analysis of Cyber Attack in Small Businesses”, *Journal of Society for e-Business Studies*, Vol. 19, No. 4, pp. 195-204, Nov. 2014

[9] Verizon. 2012 *Data Breach Investigations Report* [Internet]. Available: http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2012-ebk_en_xg.pdf

[10] H. J. Kim, J. H. Hyun, H. J. Lee, P. J. Park, and A. L. Lee, *The 4th Industrial Revolution Futures technology and Security of Company*, 1st ed. InfoTheBooks, pp. 61-73, Jan. 2017

[11] D. J. Jeon and D. G. Park, “Analysis Model for Prediction of Cyber Threats by Utilizing Big Data Technology”, *Journal of Korea Institute of Information Technology* Vol. 12, No. 5, pp. 81-100, May 31. 2014

[12] S. J. Moon, “Server Management Prediction System based on Network Log and SNMP”, *Journal of Digital Contents Society* Vol. 18, No. 4, pp 747-751, Jul. 2017

[13] ElasticSearch [Internet]. Available: https://www.elastic.co/kr/products/elasticsearch

[14] S. W. Son, K. S. Kim, J. W. Choi, and G. S. Lee, “Development of Managing Security Services System Protection Profile”, *Journal of Digital Contents Society* Vol. 16, No. 2, pp 345-353, Apr. 2015

[15] J. W. Yoon, C. Y. Park, and U. S. Song, “Building the Educational Practice System based on Open Source Cloud Computing”, *Journal of Digital Contents Society* Vol. 14, No. 4, pp 505-511, Dec. 2013

현 정 훈 (Jeong-Hoon Hyun)

1996년 : 경북대학교 공과대학 전자공학과 학사
2016년 ~ 현재 : 고려대학교 빅데이터 응용 및 보안학과(석사과정)

1995년~현 재: (주)코스콤 IT리스크관리부 정보보호운영팀장
※관심분야 : 빅데이터 정보보호, 네트워크 인프라 설계 및 구축, 네트워크 트래픽 분석, 침해대응 보안관제 등

김 형 중 (Hyoung-Joong Kim)

1978년 : 서울대학교 전기공학과 학사
1986년 : 서울대학교 제어계측공학과(공학석사)
1989년 : 서울대학교 제어계측공학과(공학박사)

1989년~2006년: 강원대학교 교수
2006년~현 재: 고려대학교 정보보호대학원 교수
※관심분야 : 컴퓨터보안, 패턴인식, 가역정보은닉, 머신러닝, 빅데이터 분석 등