碩士學位論文

Elastic Stack을 활용한 보안관제 시각화 구현과 평가

Design and Evaluation Security Control Visualization using Elastic Stack

한밭大學校 情報通信專門大學院

컴퓨터工學科

尹 成 烈

2020년 8월

Elastic Stack을 활용한 보안관제 시각화 구현과 평가

Design and Evaluation Security Control Visualization using Elastic Stack

指導敎授 金 正 鎬

이 論文을 工學碩士學位
請求論文으로 제출함

2020년 5월

한밭大學校 情報通信專門大學院

컴퓨터工學科

尹 成 烈의 碩士學位 論文을 認准함

審査委員長 황 경 호 (인)

審 査 委 員 김 태 윤 (인)

審 査 委 員 김 정 호 (인)

2020년 6월

한밭大學校 情報通信專門大學院

목 차

표 목 차

그 림 목 차

국 문 요 약

Elastic Stack을 활용한 보안관제 시각화 구현과 평가

논문제출자 윤 성 열
지 도 교 수 김 정 호

급증하는 사이버 공격에 대비하여 침해사고 발생시 원인 파악과 신속 대응을 위해 공공, 민간, 군, 금융 각 분야에서 보안관제를 위한 대응 체계를 구축하고 있다. 공공 분야에선 국가정보원이 '국가사이버안전센터'를 설치하여 담당하고 있고 군 분야에서는 국방부에서 '국방정보전대응센터'를 설치하였다. 금융분야에서는 금융보안원(FSI)이 맡고 있고 민간분야에선 한국인터넷진흥원(KISA)이 '인터넷침해대응센터'를 설치하여 각 민간기업을 대상으로 분야별 보안관제, 사이버 침해 대응 등 각 분야에서 지원을 하고 있다. 그러나 보안관제 분야에서 광범위한 분야의 관제는 수행이 어려워 네트워크 보안 관제에만 머물러 있으며 기업은 보안관제를 수행하기 위해선 전문업체를 선정하거나 솔루션을 도입해야하는 실정이다.

본 연구에서는 민간기업들이 전체적인 보안관제 인프라를 구축 할 수 있도

록 오픈소스 빅데이터 솔루션을 이용하여 보안관제 체계를 구축하는 방법을 기술한다. 특히, 보안관제 시스템(SIEM:Security Information & Event Management)을 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. Elastic Stack을 활용해 보안로그를 단계별로 수집-분석-시각화 하여 대시보드를 만들고 대용량 로그를 입력 후 검색속도를 측정하였다. 이를 통해 Elastic Stack이 Splunk를 대체 할 수 있는 빅데이터 분석 솔루션으로서의 가능성을 발견했다.

Ⅰ. 서 론

1.1 연구 배경 및 필요성

2019년 개인정보보호&정보보안 컨퍼런스인 G-PRIVACY 2019에서 한국인터넷진흥원(KISA)의 발표자료에 의하면 2011년 9월 30일 개인정보보호법 시행 이후 2018년까지 개인정보 유출 신고는 502건으로 보고되고 있다. 이는 전년대비 약 4배가 증가한 수치로 IOT 등의 발전을 원인으로 해석하고 있으며 또한 사고원인을 분석한 결과 사이버 침해사고는 대부분 취약한 솔루션을 사용하는 기업의 웹사이트 및 인터넷 기반 신서비스가 해킹에 취약해 정보유출 되는 경우가 80%를 차지한다고 말했다. 이제는 정부, 공공기관 뿐만 아니라 일반 기업체들도 정보보호에 대한 중요성의 인식이 퍼지며 서비스 구축시 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.[1]

정보보안 관제를 실시간으로 빠르게 시각화하고 대응하기 위해선 빅데이터 솔루션의 도입이 필수적이다. 빅데이터 분석솔루션은 용량이 큰 데이터를 기존의 방법이나 도구로 수집, 저장, 분석 등이 어려운 비정형 또는 정형 데이터를 처리할 때 많이 이용된다.

하지만 일반 기업이 이러한 빅데이터 분석솔루션 도입 시 3가지 문제가 있다. 첫 번째 Splunk 같은 완성된 빅데이터 분석솔루션을 구매할 경우 고비용의 문제이다. 초기 구축했더라도 시간이 갈수록 늘어나는 로그로 인해 추가 라이센스 비용을 감당하기 어려 수 있다. 두 번째 문제는 완제품 솔루션 구매 시 자체 확장성이 부족해 조직에 최적화된 솔루션이 필요하면 문제가 발생한다. 조직에 최적화된 솔루션이 필요한 경우 많은 시간과

비용의 문제가 발생한다. 세 번째, 분석 데이터에 대한 시각화의 어려움이다. Hadoop과 Spark 같은 빅데이터 솔루션은 시각화를 구축할 때, 처음부터 모든 부분을 설계하고 개발해야 하며, 자체 솔루션 개발시 시간 및 비용 등이 문제일 것이며, 이것은 개발자를 보유하고 있지 않은 회사의 경우, 큰 장벽이 될 것이다.[2]

1.2 논문의 목적

본 논문에서는 빅데이터 분석시스템 도입 시 비용·개발시간을 단축 할 수 있는 하나의 방법으로 무료 오픈소스 솔루션 중 하나인 Elastic Stack에 대해 기술하고, 빅데이터 분석 솔루션중 산업계에서 주로 사용하는 제품인 Splunk와 검색 성능 비교실험을 진행하였다. 이를 통해 Elastic Stack이 Splunk를 대체 할 수 있는 솔루션인지 실험하였다. 오픈 소스코드 솔루션 Elastic Stack을 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 보안 이벤트 처리가 가능한 SIEM(Security information and event management) 솔루션으로의 발전 가능성도 확인할 수 있었다.

1.3 논문의 구성

본 논문은 5장으로 구성되어 있다. 1장은 논문의 배경과 보안관제 솔루션들의 문제점에 대해 설명하고 2장은 빅데이터 솔루션들의 기능을 소개하고 각 솔루션의 장·단점을 비교하였다. 먼저 산업계에서 주로사용하고 있는 Splunk와 본 논문에서 활용한 Elastic Stack을 소개하고 3장에서는 본 논문에서 다루고자 하는 오픈소스 빅데이터 솔루션인 Elastic Stack과의 비교를 통해 빅데이터 분석 솔루션을 통한 보안관제 솔루션 구축방법을 확인 할 것이다. 또한 Splunk의 세부 기능들을 Elastic Stack의 기능을 기술하며 비교하였고 Elasticsearch, Logstash, Kibana 3가지 솔루션의 역할과 세부 기능 등을 기술하였다. 4장에서는 Dashboard 구현을 통한 시각화 테스트와

100만줄부터 1000만줄 까지 로그 검색결과를 상용솔루션인 Splunk와 Elastic Stack을 비교하여 성능 테스트를 진행하였고 5장에서는 실험 결과 등을 통해 해당 시스템으로 보안관제 시스템 구축 후 기업에서 실제 활용이 가능한지 여부를 해석하였다.

Ⅱ. 관련 연구

2.1 보안관제에서 빅데이터

정보보안에 취약한 중소기업들을 위해 취약점 점검 지원, 보안컨설팅 등 다양한 지원사업은 추진되고 있다. 하지만 2019년부터 금융데이터를 클라우드 인프라(AWS, Azure 등)을 이용해서 처리할 수 있게 되어 많은 기업들이 서비스를 운용할 때 클라우드 인프라를 활용한다. 하지만 이를 악용하여 조직적이고 지능적으로 사이버 공격또한 발전하고 있다. 또한 제도적으로 복합한 정보보안관련 여러 가지 규제를 지키며 운영 하기란 어려운 일이다. 때문에 대부분의 기업들은 침해사고 발생시 사후대응에 바쁘고 사고 발생시마다 새로운 솔루션을 도입하고 이를 운영할 인력또한 추가 배치하게 된다. 하지만 침해대응은 현재의 사후 대응체계를 유지하는 것은 적절치 않다.[3]

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.

2.2 빅데이터 솔루션 활용 이유

기존의 보안관제 솔루션은 RDBMS(Relational Database Management) 기반으로 구축되는 경우가 많았다. 하지만 RDBMS는 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 index를 수행하는 RDBMS와는 반대 구조인 inverted index 구조로 데이터를 토큰화 시켜 저장한다. 때문에 비정형 데이터에 대해서도 Full Text 검색 속도가 RDBMS에 비해 매우 빠르다.

	RDBMS	빅데이터 검색엔진
데이터 저장 방식	정규화	**역정규화**
전문(Full Tex) 검색 속도	느림	**빠름**
의미 검색	불가능	**가능**
Join	가능	**불가능**
수정 / 삭제	빠름	**느림**

<표 2-1> RDBMS,빅데이터 엔진비교

2.3 관련 도구들

2.3.1 Splunk

Splunk는 빅데이터 분석솔루션으로 현재 산업계에서 가장 많이 사용되는 솔루션이다. 시스템에서 발생하는 모든 데이터를 실시간으로 저장, 분류하고 한 곳에서 검색, 분석, 시각화 하여 수집된 데이터에 대하여 가시성과 대응 능력을 극대화하는 대량 분산 처리 모델의 단일 플랫폼 소프트웨어이다. 통신, 정보보안, 금융, 의료, 교육, 비영리단체, 공공 부분, 온라인서비스 등 다양한 분야에서 빅데이터 분석을 위해 활용 중이다.[5]

Splunk는 강력한 UI를 지원하고 사용자가 원하는 UI로 변경 가능하며 데이터 즉시 분석이 가능하다. Fortune 100대 기업 중 92명이 이용하고 있으며 그 외에 9,000여 개 이상의 기업, 서비스공급자와 정부가 Splunk 솔루션을 이용하고 있다. <그림 2-1>처럼 국내에서도 많은 기업과 공공기관에서 빅데이터 분석을 위해 Splunk를 이용하고 있다. 2020년에 Splunk는 7년 연속 '가트너 매직 쿼드런트'의 SIEM(Security Information & Event Management) 부분에서 리더로 선정될 정도로 보안업계에서도 많이 활용되고 있다.[6]

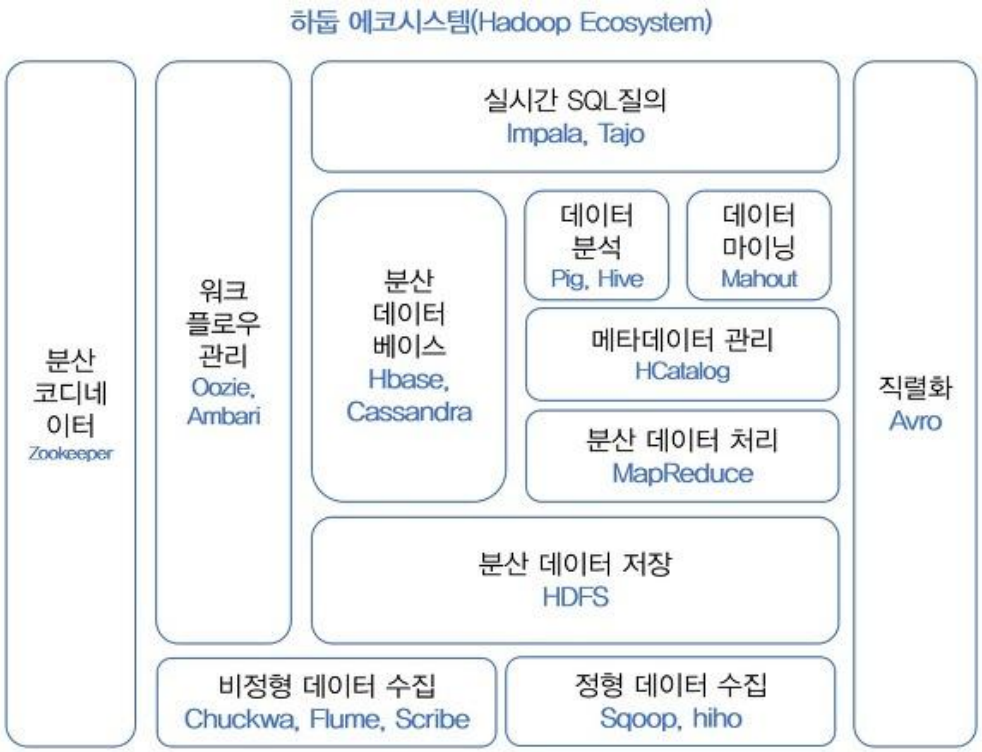
<그림 2-1> Splunk사 주요 국내고객

하지만 Splunk는 <그림 2-3>처럼 일일 로그 처리량으로 금액이 산정되는 유료 소프트웨어라는 것이다. Splunk가 고비용인 이유는 다양한 분야(IT, 통신, 기계, 건설, 조선, 금융, 의료, 제조업 등)에서 이용할 수 있도록 많은 기능이 포함되어 제작되었기 때문이다. 각종 시각화 기능, 수학적 통계, 보고서, 다국어, 백업 기능, 압축 등 많은 기능을 지원하기 때문에 사용하지 않는 많은 기능이 소프트웨어 가격으로 포함된 것으로 보인다. 일회성 침해사고 로그 분석이나 중소기업에는 큰 부담일 수밖에 없다.[7]

저비용의 1GB 정도의 영구 라이센스로 구축했더라도 지속적인 로그량 증가로 추가 라이센스를 구매해야 한다면 구매자 입장에서 부담이 될 것이다. 과거 보안 로그는 하루 처리량이 기가바이트 단위로 처리하던 데이터를 현재는 테라바이트 단위까지 처리해야 하는 상황이 있으므로 Splunk의 일일 처리량에 따른 금액 산정 정책은 큰 단점일 수밖에 없다.

2.3.2 Hadoop

Hadoop은 아파치 재단의 프로젝트로 빅데이터 처리 플랫폼으로 많이 알려져 있으며 Hadoop은 아파치 재단의 다양한 서브 프로젝트들과 융합되어 하둡 에코 시스템이라는 명칭으로 불린다.[8]



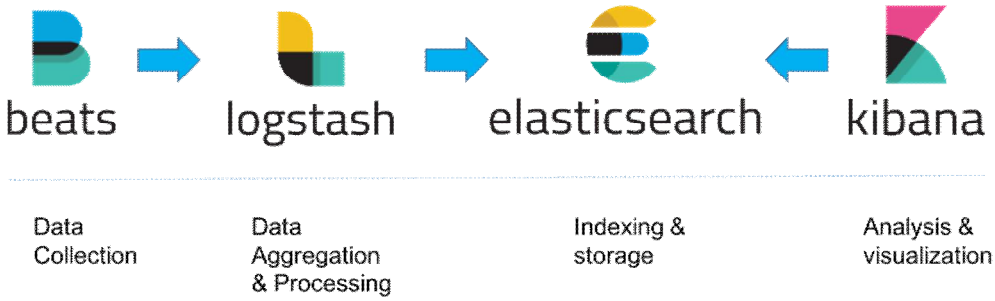
<그림 2-2> Hadoop Ecosystem

Hadoop은 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 Hadoop은 다양한 프로젝트 들과 융합되어 시너지 높은 모델이 될 수 있지만 다양한 프로젝트 간의 호환성과 보안관제에서 중요한 시각화에 대한 약점이 있어 보안관제 솔루션으로는 적합하지 않다.

2.3.3 Spark

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

2.3.4 Elastic Stack



<그림 2-3> Elastic Stack 구성

Elastic Stack은 <그림 2-3>처럼 Elasticsearch+Logstash+Kibana+filebeats로 구성되어 있다. Elastic Stack 중 Elasticsearch는 Apache Lucene를 바탕으로 개발된 검색엔진 솔루션이며. Logstash는 beats 등을 이용하여 수집한 각종 로그를 JSON 형태로 만들어 Elasticsearch로 전송하는 역할을 하고 Kibana는 Elasticsearch에 저장된 Data를 사용자에게 그래프, 테이블 등 시각화 형태로 보여주는 솔루션이다.

2.3.4 Elasticsearch

Elasticsearch는 셰이 배논(Shay Banon)에 의해 Lucene을 기반으로 만들어진 분산 검색 엔진이다. Lucene은 더그 커팅 (Doug Cutting)이

개발했고 손쉽게 검색 기능을 추가할 수 있게 도와주는 자바 형태의 검색 라이브러리이다. 인덱스 저장, 관리 그리고 쿼리 문을 수행하여 결과 값을 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

Elasticsearch는 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)

Elasticsearch는 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.

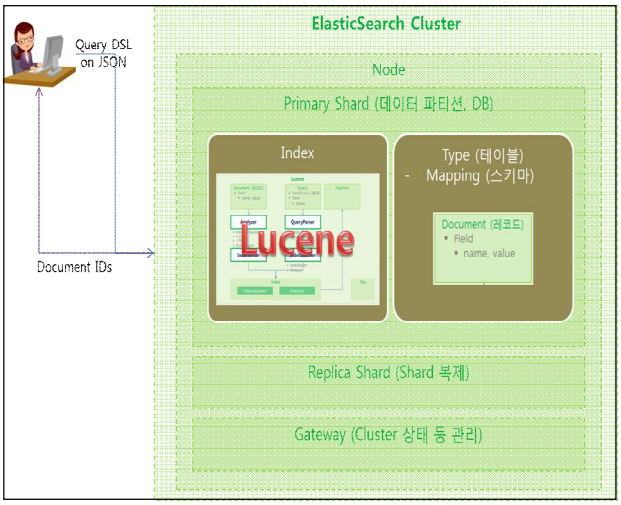
4) 전문 검색(Full text search)

Elasticsearch는 강력한 전문검색을 지원한다.

5) 문서 중심(Document Oriented)

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

Elasticsearch는 <그림 2-4>처럼 Cluster, Node, Shard, Replica, Gateway로 구성된다. Elasticsearch를 쉽게 이해하기 위해서는 사용되는 [표 2-1]과 같이 정의된다.



<그림 2-4> Elasticsearch 개념적 구성도

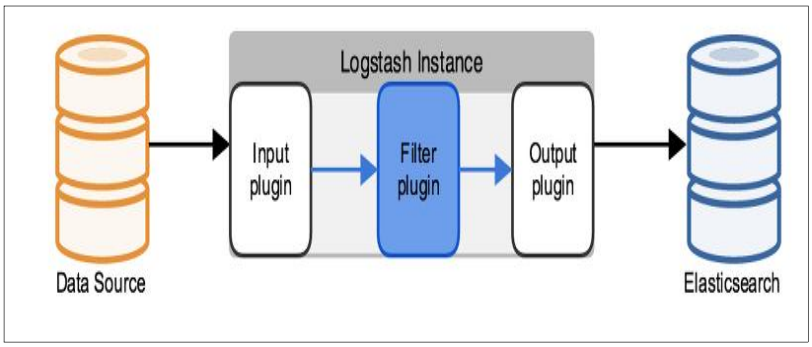
용어	상세
Cluster	Node의 집합으로 유일한 이름을 가짐
Node	Cluster를 이루는 물리적인 서버
Index (indice)	유사한 특징을 가진 문서들의 모음으로 DBMS에서데이터베이스와 유사한 개념 Term, Count, Docs로 구성
Shard	Index의 subset 개념으로 Lucene을 사용하여 구성 실제 데이터와 색인을 저장하고 있으며 분배와 병렬화 연산을 통해 높은 성능과 처리량을 가짐
Type (Document Type)	데이터 (Document)의 종류로 index 내에서의 논리적인 category/partition DBMS에서 테이블과 유사한 개념
Mapping	DBMS에서 테이블 스키마와 유사한 개념
Route	색인 필드 중 unique key에 해당하는 값을 routing path로 지정한 후, 이 path를 사용하여 인덱싱과 검색에 사용할 shard를 지정하여 성능을 향상할 수

	있음
Document	ElasticSearch에서 관리하는 기본적인 데이터(정보)의 저장 단위 JSON (JavaScript Object Notaion)으로 표현 DBMS에서 레코드와 유사한 개념
Field	Document를 구성하고 있는 항목으로 name과 value로 구성되어 있으며, DBMS에서 컬럼과 유사한 개념
Gateway	Cluster 상태, Index 설정 등의 정보를 저장
Query	검색어
TermQuery	검색어의 종류
Term	검색어의 항목
Token	검색어의 항목을 구성하는 요소

<표 2-2> Elasticsearch 용어 설명

2.3.5 Logstash

Elasticsearch는 뛰어난 검색엔진이지만 사용하려면 입력할 데이터를 JSON 형태로 가공해야 한다. Logstash는 JRuby로 만들어졌다. 로그 수집 및 가공을 위해 만들어졌으며 다양한 방식으로 데이터를 입력받아 Elasticsearch로 전달하는 역할을 한다.

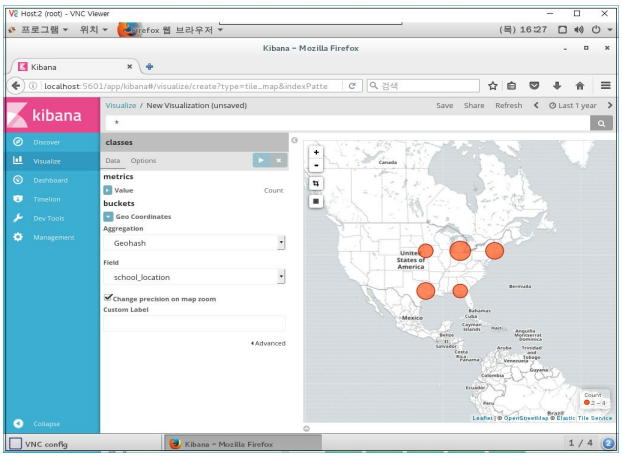


<그림 2-5> Logstash 작동 과정

Logstash의 작동 과정은 <그림 2-5>처럼 크게 Input, Filter, Output 단계로 이뤄진다. Input은 로그의 위치를 정의하고 데이터를 읽어온다. Filter는 읽어온 데이터를 가공한다. Filter를 이용하는 방법은 여러 가지가 있다. 정규표현식을 이용해 로그를 자르는 방법과 grok을 이용해 미리 정의된 정규표현식을 사용하는 방법 CSV를 이용해 데이터를 엑셀의 자르기 기능과 같이 일정한 패턴(콤마, 점) 등으로 자르는 방법과 XML을 이용한 방법 등 다양하다. Output에서는 필터링 된 보안 로그를 Elasticsearch로 전달하는 역할을 한다.

2.3.6 Kibana

<그림 2-6>는 Kibana 실행 화면이다. Kibana는 Logstash를 통해 Elasticsearch에 모인 로그 데이터를 쉽게 검색하고, 다양한 시각화 분석을 할 수 있게 도와주는 솔루션이다.



<그림 2-6> Kibana 실행 화면

Kibana의 주요 메뉴는 Discover, Visualize, Dashboard로 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 : Elasticsearch에 수집된 결과를 시각화 분석을 할 수 있다. 막대 그래프, Area chart, 테이블 등 여러 종류의 시각화 도구를 지원하고 있다.

3) Dashboard : Visualize를 통해 시각화한 객체를 모아 하나의 Dashboard에 배치하여 한눈에 확인할 수 있다.

2.4 보안관제 데이터별 시각화 방법

Elastic Stack은 다양한 기능을 활용해서 시각화를 구현할 수 있다. 하지만 데이터를 시각화 하기 위해선 데이터의 특성에 따라 적합한 표현 기법이 있다. 분석한 로그는 데이터 특성에 따라 시각화를 해야한다.[22]

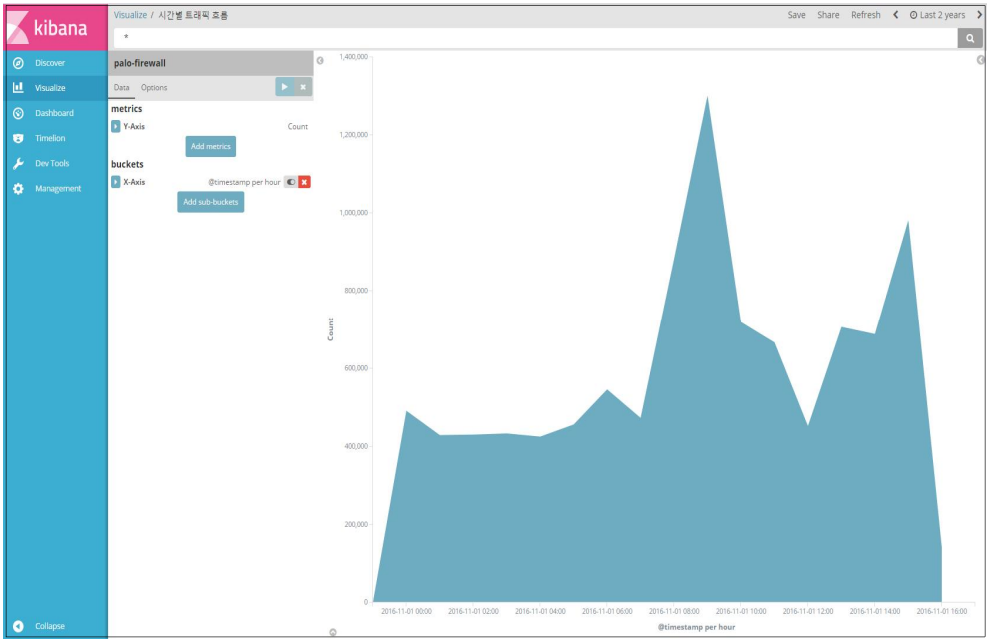
1) 표 : 표는 간단하며 이해하고 해석하기 쉬운 데이터 표현 기법이다. 침해사고 대응에서의 가장 기본은 출발지와 목적지의 파악이다. 해당 정보를 파악하기 위해서는 일반적으로 Src_ip, Dst_ip, Src_port, Dst_port, Protocol 등을 확인하는 것이 일반적이며 이를 표현하기에 가장 적절한 시각화 방법은 표다. 표는 자료를 정확하게 표현하고 다른 데이터 표현의 기초자료가 될 수있기 때문에 그래프 등과 비교했을 때 직관적인 표현 방법은 아니지만 표와 연계한다면 다양한 시각화 기법 등이 활용 될 수 있다.

차단된 IP Table

SourceAddress: Descending	DestinationAddress: Descending	DestinationPort: Descending	IPprotoc Descend
		80	tcp
		6881	tcp
		6881	udp
		6881	tcp
		6881	udp
		6881	tcp
		6881	udp
		6881	tcp
		6881	udp
		6881	tcp

<그림 2-7> 보안관제 표 예시

2) 선도표 : 선도표는 연결된 점으로 정보를 나타내는 그래프로, 연속된 선 또는 직선으로 표현된다. 흔히 시간적 간극이 있는 데이터의 트랜드를 시각화 하는데 사용되며 보안관제에선 로그 데이터의 시간별 용량, 접속 추이 등을 분석할 때 활용 가능한 지표로 활용된다. 구성될 데이터에 따라 면적도표로도 표현이 되며 연속적으로 값을 보는게 중요한 보안관제에선 필수적인 시각화 방법이다.

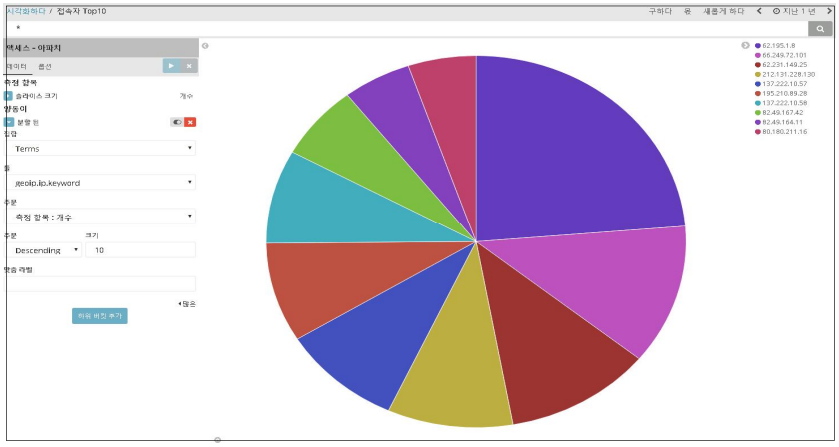


<그림 2-8> 선도표, 면적도표 예시

3) 막대도표 : 막대도표는 시각화에서 가장 많이 쓰이는 기법으로 '막대그래프' 라고도 한다. 막대도표는 대부분 연속성 보다는 분산데이터 분석에 많이 쓰이며 보안관제에서는 접속 지역비교 등을 위해 많이 사용된다. 막대 도표는 단순한 직선 막대로 표현하기 때문에 각 항목을 비교하기가 좋은 시각화 방법이다. 특히 하나의 특이점이 존재할 경우 이를 즉각적으로 발견가능 하기 때문에 매우 많이 활용된다.

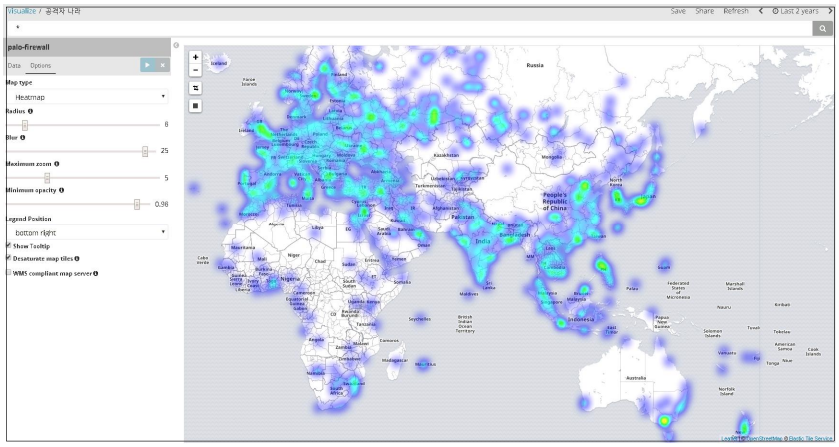
<그림 2-9> 막대도표 예시

4) 파이도표 : 파이 도표는 일반적으로 '원형 도표'라고 하며 공통의 속성 또는 특성을 분류하여 퍼센트를 나타내기 위해서 사용한다. 보안관제에선 지역을 확인, 프로토콜 분포도 등을 표현하기 위해 많이 사용된다. 파이 도표는 전체를 더한 것이 하나의 원을 이루므로 각 항목이 전체에서 차지하는 비율을 짐작하기 쉽기 때문에 보안관제에서 많이 활용되는 시각화 기법이다.



<그림 2-10> 파이차트 예시

5) 산포도 : 데이터의 군집을 발견하기 위해 많이 활용되며 데이터 간의 관계 분석에 용이하기 때문에 보안관제에서도 많이 활용된다. 보안관제에서는 주로 지도로 많이 표시하며 국가별 트래픽 현황을 한눈에 볼 수 있어 용이하다.



<그림 2-11> 산포도 예시

Ⅲ. Elastic Stack 시스템 구축

3.1 실행환경

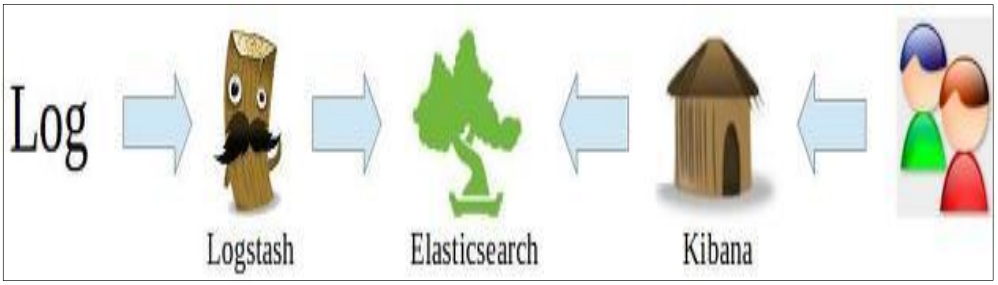
이번 장에서는 Elastic Stack 환경구축하기 위한 방법에 대해 기술하였다. 대용량 방화벽 로그를 활용해 로그검색, 대시보드, 파이차트 등 시각화 표현을 구현한 것에 관해 기술한다. [표 3-1]과 같은 환경으로 구성되며, VMware 환경 기반으로 구성하고 OS는 리눅스(CentOS 6.8)를 사용하였다. Elastic Stack은 Elasticsearch+Logstash+Kibana로 구성되기 때문에 각각 설치해야 하며, 설치 후 최적화 설정 과정도 진행한다.

하드웨어
CPU : Processor 1 / Core : 2
MEMORY : 8GB
OS : CentOS 6.8 (vmware)
소프트웨어
logstash-6.3.0
kibana-6.3.0
elasticsearch-6.3.0
java-1.8.0

<표 3-1> Elastic Stack 시스템 구성

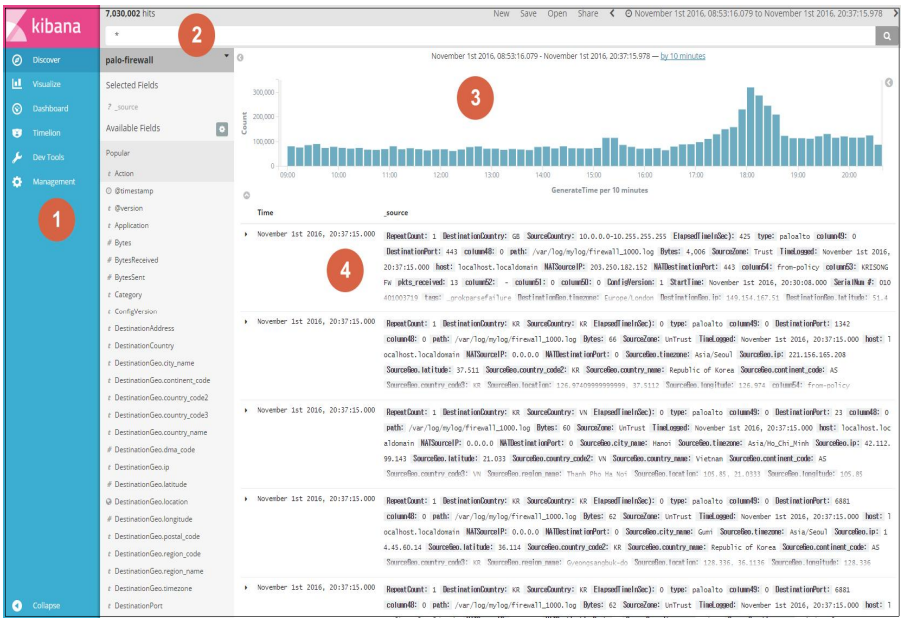
3.2 로그 분석 시스템 구축

Elastic Stack으로 구성한 로그 분석시스템은 <그림 3-1>과 같이 구성되었다. 로그가 수집되면 Logstash에서 필터링을 거친 후 Elasticsearch 저장된다. 저장된 데이터를 Kibana 조회를 통해 Elasticsearch에 접근해 원하는 데이터를 조회할 수 있다.



<그림 3-1> 로그 분석시스템 구성

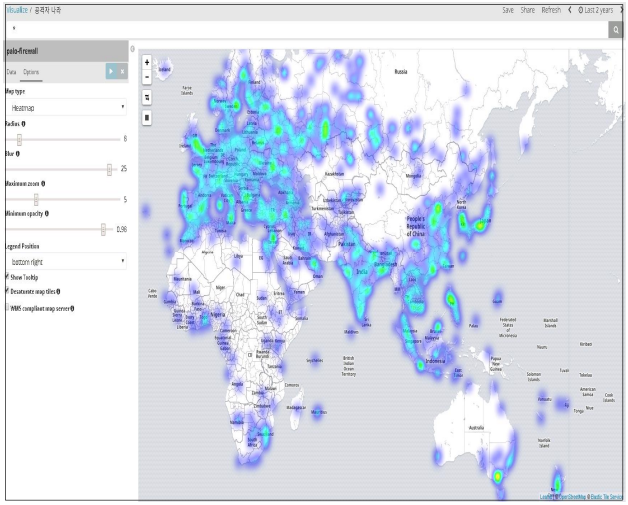
<그림 3-2>는 로그 분석시스템 구축 후 Kibana를 통해 접속한 화면이다. 왼쪽 1번 영역은 검색, 각종 시각화분석, 대시보드 설정 등의 메뉴가 구성되어 있다. 2번 영역은 로그 검색을 할 수 있고 JSON 형태의 명령어로 질의가 가능하다. 3번 영역은 트래픽을 그래프 형태로 보여준다. 어느 시간대 트래픽이 많은지 확인할 수 있다. 간단한 클릭으로 시간대 조회도 가능하다. 4번 영역은 검색된 로그의 정보가 출력된다.



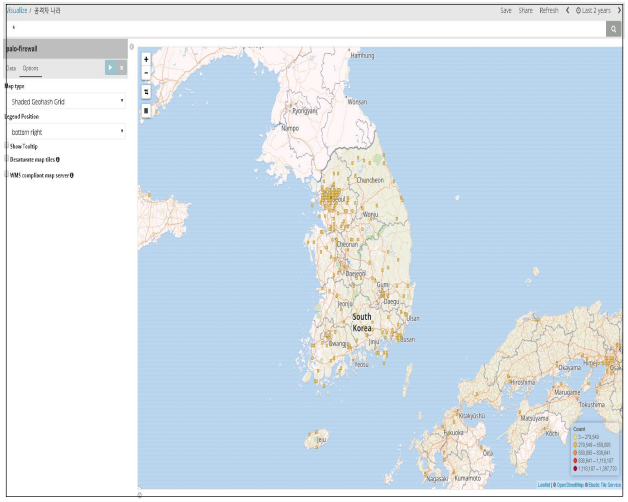
<그림 3-2> 로그분석시스템 접속 화면

3.3 GeoIP를 활용한 로그 시각화 해석

Elastic Stack을 활용해 <그림 3-3>처럼 국가별 로그 접속 추이를 직접 단번에 확인할 수 있다. <그림 3-4>처럼 오른쪽 옵션을 메뉴를 통해 다른 형태로 변경도 가능하여 DDoS 공격 분석 시 나라별 접근 현황을 시간대별로 확인할 수 있다.



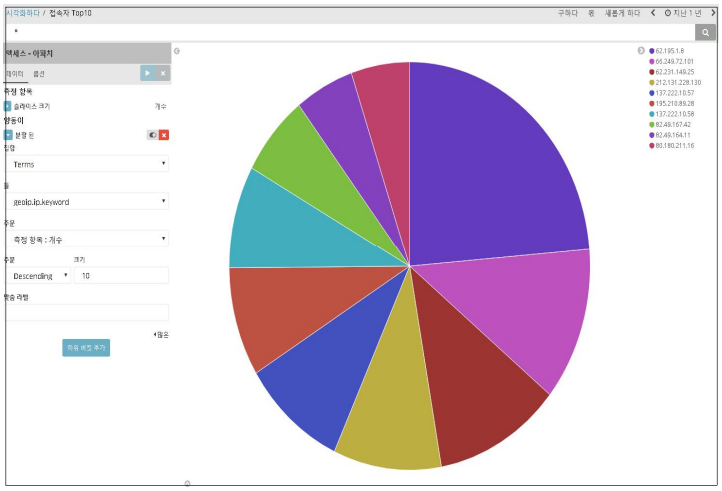
<그림 3-3> GeoIP 활용 로그 시각화 예시1



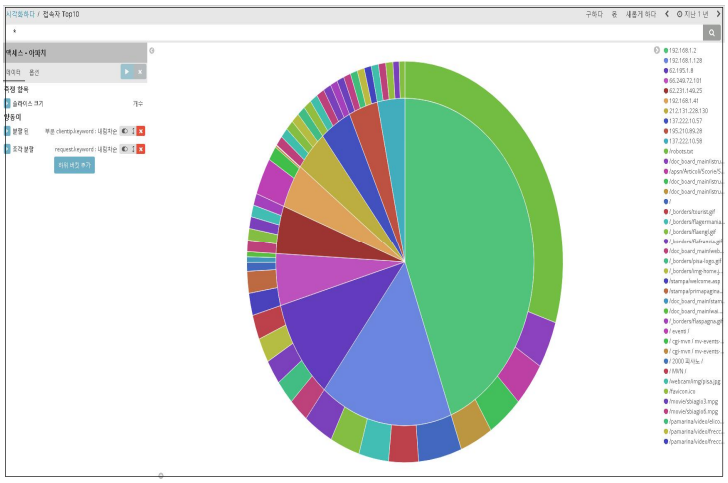
<그림 3-4> GeoIP 활용 로그 시각화 예시2

3.4 Pie chart를 활용한 시각화 해석

Pie chart는 전체에 대한 각 부분의 비율을 부채꼴 모양으로 나타낸 그래프이다. 각 부채꼴의 중심각이 전체에서 차지하는 비율을 나타내며, 전체의 비율 및 통계적 수치를 한눈에 볼 수 있다는 장점이 있다. 로그 분석 시에는 <그림 3-5> <그림 3-6>처럼 접근IP TOP 10, 접근 포트 TOP 10 등으로 활용 범위는 다양하다.



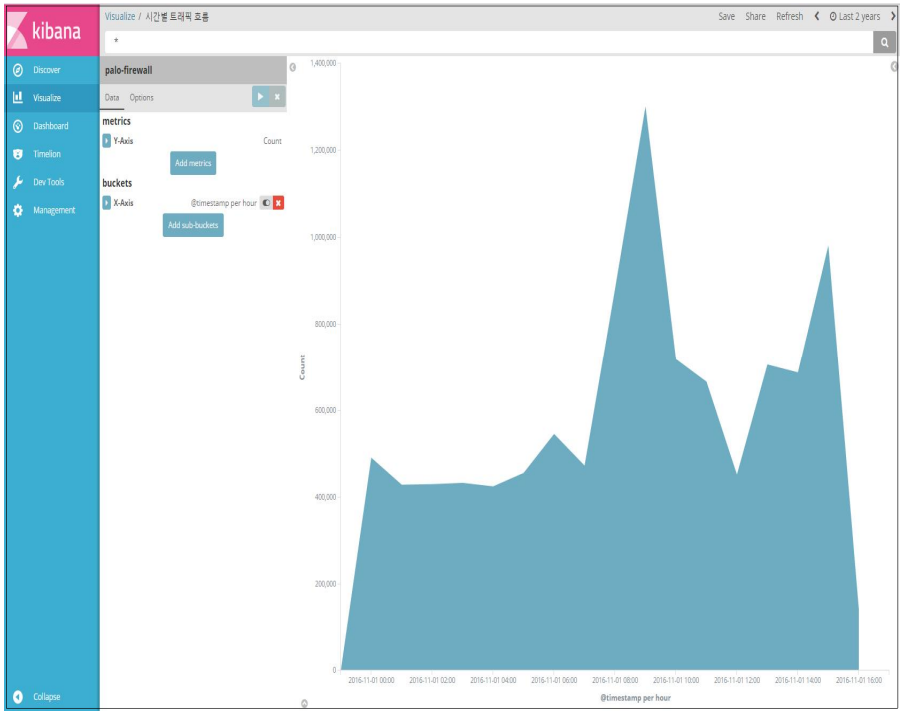
<그림 3-5> Pie chart활용 로그 시각화 예시1



<그림 3-6> Pie chart활용 로그 시각화 예시2

3.5 Area chart를 활용한 시각화 해석

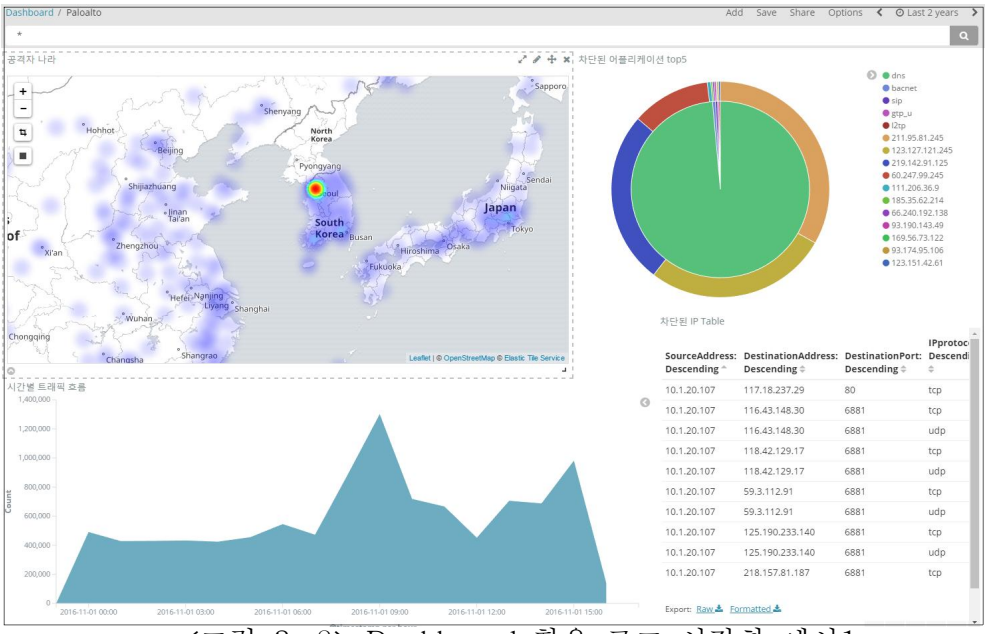
<그림 3-7> Area chart는 로그의 시간에 변화에 트레픽 변화, 접속 추이 등을 분석할 때 주로 사용가능 하다.



<그림 3-7> Area chart 활용 로그 시각화 예시

3.6 Dashboard를 활용한 시각화 해석

<그림 3-8>, <그림 3-9>는 Dashboard로 로그와 관련된 현황 분석을 단번에 확인할 수 있다. 어느 곳에서 접속을 많이 했는지 또는, 어느 시간대에 접속이 많이 이루어졌는지를 로그 분석 시 한눈에 확인할 수 있다.



<그림 3-8> Dashboard 활용 로그 시각화 예시1

<그림 3-9> Dashboard 활용 로그 시각화 예시2

Ⅳ. 빅데이터 솔루션 비교

4.1 비교항목 선정

이번장에서는 Elastic Stack과 상용 솔루션인 Splunk의 기능 비교를 진행하였다. 미국의 정보 기술 연구 및 자문회사인 가트너에서 보안관제 솔루션에 대해 사용자별 비교 조사를 진행하였다. Elastic Stack은 총 40명, Splunk는 82명이 응답 하였으며 기능, 사용의 용이성, 서비스 지원, 가격, 기술지원 총 5가지 항목을 대표로 비교 하였다.

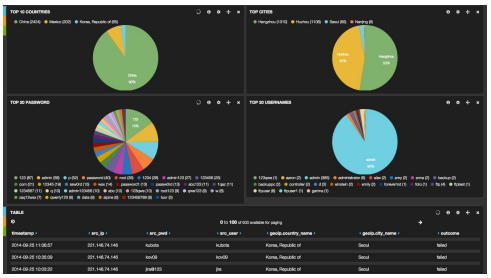
전체 평가 점수는 Elastic Stack은 4.3점, Splunk는 4.6점으로 총점은 Splunk가 높았지만 제품 추천 여부에 대해서는 이용자 중 95%가 Elastic Stack을 추천했고, Splunk는 89%가 추천하겠다고 응답하였다.

1) 기능 : 빅데이터 분석 솔루션에서 기업에 맞춤형 요구 사항을 충족시키는 건 매우 중요하다. 기능의 요소로는 맞춤형 제작 여부, 검색쿼리, 경고/알림, 데이터 시각화 등이 포함 된다.
2) 사용의 용이성 : 다양한 기업들이 활용함에 있어 용이성은 굉장히 중요하다. 아무리 좋은 솔루션이여도 사용법이 어렵다면 결국 구축하기 위한 비용 및 시간을 투자해야하기 때문이다.
3) 서비스 지원 : 빅데이터 솔루션을 활용하여 보안관제센터 구축시에 발생하는 문제를 해결 할 수 있는 커뮤니티는 필요하다. 개발사가 아닌 이를 이용하는 사용자간에도 자유롭게 의견을 나눌 수 있는 커뮤니티또한 중요한 요소이다.
4) 가격 및 지원 : 기업입장에서 솔루션 도입시 반드시 고려해야할 사항이다. 보안관제는 보통 기업에 이익을 가져다주는 분야가 아닌만큼 최소 비용으로 구축을 해야한다.
5) 기술지원 : RESful API를 활용하고 이를 문서화시켜 편하게 만들 었는지, 최초 도입후 확장시에 용이한지 또한 매우 중요한 요소이다.

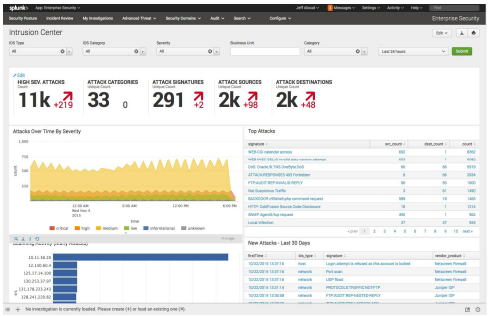
4.2 비교결과

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

Elasitc Stack



Splunk



<그림 4-1> 시각화 Dashboard 비교

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

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

4) 가격 및 지원 : 오픈소스 프로젝트인 Elastic Stack은 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 Splunk도 동일하며 Splunk는 로그 용량별로 가격을 책정한다.

Index Volume	Perpetual License (per GB)	Annual Term License (per GB)	Volume Purchase Discount
1GB Per Day	$4,500	$1,800	0%
10GB Per Day	$2,500	$1,000	44%
50GB Per Day	$1,900	$760	58%
100GB Per Day	$1,500	$600	67%
>100GB Per Day	Contact sales for custom pricing with additional volume discounts		

<그림 4-2> Splunk 가격정책

5) 기술지원 : Splunk는 개발시에 자주 사용되는 언어용 SDK를 제작하여 배포할 뿐만 아니라 200개 이상의 RESTful API를 제공하며 문서화 또한 훌륭 하다. Elasticsearch또한 RESTful API를 제공하며 개발언어에 대하여 SDK를 제공한다. 하지만 Splunk는 API를 자체적으로 개발하여 확장하기에는 완성형 솔루션이라 어려운 편이지만 Elasticsearch는 오픈소스 솔루션이므로 다양한 맞춤형 APP을 개발할 수 있다.[23]

제품명	Elastic Stack	Splunk
평가자수	40명	82명
기능	4.5	4.7
사용의용이성	4.3	4.6
서비스지원	4.3	4.5
가격 및 지원	4.4	4.2
기술지원	4.1	4.5

<표 4-1> Elastic Stack vs Splunk 비교설문

4.3 성능 테스트

4.3.1 실험 환경 구성

<표 4-1>은 Paloalto 방화벽 로그의 1개 라인의 샘플이다. 해당 Paloalto 방화벽 로그를 Elasticearch에 저장해야 한다. Splunk는 Fulltext 형태로 일단 하드디스크에 저장 후 Mapping(매핑) 설정하는 방식이지만 Elasticsearch는 로그 저장 전 Logstash와 Elasticearch에서 각 필드를 Mapping 후 하드디스크에 저장하는 방식이다.

```
<14> Dec 8 14:33:09 1,2015/12/08
14:33:09,001701002739,TRAFFIC,start,1,2015/12/08
14:33:08,203.230.46.147,104.20.5.36,0.0.0.0,0.0.0.0,IPS,,,web-browsing,vsy
s1,Main-wire,Main-wire,ethernet1/14,ethernet1/13,Log_Forwarding_Prof
ile,2015/12/08
14:33:08,198088,1,7944,80,0,0,0x0,tcp,allow,1172,1106,66,4,2015/12/08
14:33:09,0,any,0,5622694095,0x0,Korea Republic Of,United States
```

<표 4-2> 방화벽 로그 예시

Paloalto 로그를 저장하기 위해서는 먼저 <표 4-1>의 원본 로그를 [표 4-2]처럼 방화벽 로그 정의서를 만들어야 한다. 이렇게 로그를 분리해줘야 로그가 정상적으로 저장되는지 알 수 있다.

NO	Field Date	Field Name	Field Type
1	<14>	PRI	PRI
2	43040	Event Time	이벤트발생 시간
3	KRISONGFW 1	Domain	장비명
4	42674.99999	ReceiveTime	로그수집 시간
5	10401003719	SerialNum #	시리얼 번호
6	TRAFFIC	Type	타입
7	drop	Threat-ContentType	서브타입
8	1	ConfigVersion	버전
9	42674.99999	GenerateTime	로그시간
10	10.2.10.131	SourceAddress	출발지 IP

11	210.103.76.7	DestinationAddress	도착지 IP
12	0.0.0.0	NATSourceIP	nat 출발지 IP
13	0.0.0.0	NATDestinationIP	nat 도착지 IP
14	Rule_86_SNAT_Port_Deny	Rule	룰
15	-	SourceUser	출발지 USER
16	-	DestinationUser	도착지 USER
17	not-applicable	Application	어플리케이션
18	vsys1	VirtualSystem	가상시스템
19	Trust	SourceZone	출발지 zone
20	UnTrust	DestinationZone	도착지 zone
21	ethernet1/11	InboundInterface	들어오는 인터페이스
22	-	OutboundInterface	나가는 인퍼페이스
23	Log-Forwarding	LogAction	행위
24	42674.9999884259	TimeLogged	none
25	0	SessionID	세션 ID
26	1	RepeatCount	응답횟수
27	46075	SourcePort	출발지 포트
28	110	DestinationPort	도착지 포트
29	0	NATSourcePort	nat 출발지 포트
30	0	NATDestinationPort	nat 도착지 포트
31	0x0	Flags	flags
32	tcp	IPprotocol	프로토콜
33	deny	Action	행위
34	74	Bytes	바이트
35	74	BytesSent	보낸 바이트
36	0	BytesReceived	받는 바이트
37	1	Packets	패킷
38	42675	StartTime	none
39	0	ElapsedTimeInSec	none
40	any	Category	세션 범주
41	0	Padding	direcation
42	381910764	seqno	주소 영역
43	0x0	actionflags	플레이그
44	10.0.0.0-10.255.255.255	SourceCountry	출발지_나라
45	KR	DestinationCountry	도착지_나라
46	0	cpadding	URL카테고리 허용여부
47	1	pkts_sent	보낸 패킷사이즈

48	0	pkts_received	받은 패킷사이즈
49	policy-deny	session_end_reason	세션 종료이유

<표 4-3> 방화벽 로그 정의서

원본 로그는 각각의 필드가 쉼표(,)로 구분되어 있다. CSV(comma separated value) 형태로 데이터가 구분된 것이다. [표 4-3]은 Logstash에서 Paloalto 방화벽 로그를 Elasticsearch에 저장하기 위한 설정이다[13]. 각각의 필드는 쉼표로 구분하며 columns를 이용해 각각의 필드 이름을 설정한다. 그리고 기본적으로 정의되지 않은 필드는 String으로 저장되나 convert를 이용하면 integer 등 다른 속성을 설정할 수 있다.

4.3.2 검색 성능 비교

빅데이터 기반 로그 솔루션 2개의 검색 성능을 비교하였다. 빅데이터라고 하기엔 데이터가 적지만 최대 1억건(40GB) 정도의 방화벽 로그데이터 Elasticsearch, Splunk 각각 index 작업을 거친후 특정IP를 Full Text검색 하였다. 검색 결과 미세하지만 Splunk의 검색속도가 우세 하였다.

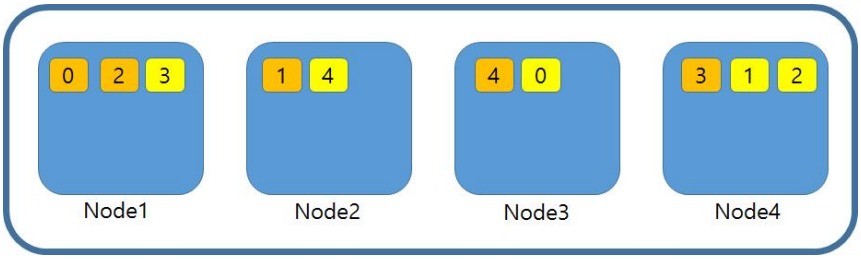
ES응답시간(ms)	Splunk응답시간(ms)	용량(GB)
20	20	5
38	40	10
52	48	15
77	75	20
98	92	25
112	108	30
139	128	35
157	146	40

<표 4-4> 검색 성능 비교

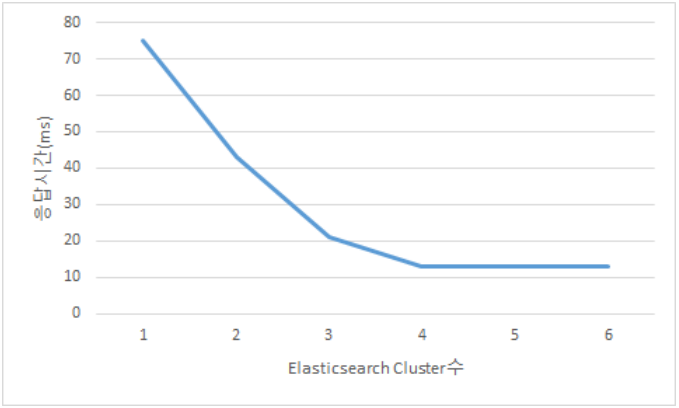
<그림 4-3> 검색 성능 비교

4.3.3 Elasticsearch Cluster Node

Elasticsearch는 기본적으로 분산 저장이 가능하다. Cluster Node의 개수가 늘어날수록 시스템이 안정화 되고 검색처리에 대한 부하를 분산하기 때문에 검색속도는 빨라지게 된다. 그림을 보면 Cluster의 수가 늘어날수록 응답시간이 줄어드는 것을 확인 할 수 있다. 로그데이터의 용량(40GB 고정) 이 크지 않아 Cluster Node의 수가 4개일 때 부터는 응답시간이 거의 유사하다. 위의 검색성능 비교에서 로그데이터 용량이 약 25GB일 때 100ms의 응답속도가 나왔다. 만약 하루에 100GB의 로그데이터가 쌓이는 기업이라면 Cluster Node를 4개정도 구축하여 구축한다면 무리없이 사용할 수 있는 수준으로 실험 결과가 도출되었다.



<그림 4-4> Elasticsearch Cluster Node 예시



<그림 4-5> Elasticsearch Cluster Node 테스트

Ⅴ. 결론

사이버 침해사고가 증가함에 따라 실시간으로 침해사고를 발견하고 대응할 수 있는 정보보안관제의 중요성 또한 높아졌다. 실시간으로 쌓이는 데이터를 저장·분석하고 이를 시각화 하여 가독성을 높일 수 있는 빅데이터 기반의 솔루션은 보안관제에서 거의 필수적으로 되었다.

본 논문에서는 오픈소스 기반의 빅데이터 솔루션인 Elastic Stack을 활용해 로그 분석시스템을 구축하는 방안을 기술하였다. 보안 솔루션의 로그를 수집하여 분석하였고, 이를 대시보드 형태로 시각화 하였다. 기존의 RDBMS(Relational Database Management System) 방식으로는 속도이슈가 있었던 대용량 로그에 대한 분석을 NOSQL인 Elasticsearch를 사용함으로써 검색속도가 개선되었고, 시각화 툴인 Kibana를 활용하여 실시간 보안이벤트 처리가 가능한 대시보드 형태로 구현하였다. 사이버 침해사고 발생시 가장 중요한 사항이 빠른 대응인 만큼 Elastic Stack을 활용한 대용량 로그 분석이 유료 솔루션과 비슷한 성능을 발휘한다는 것을 확인하였다.

본 논문은 오픈소스만을 활용하여 보안관제센터를 구축할 수있다는 점에서 의미가 있다. 기존의 상용화된 보안솔루션 제품들과 빅데이터 분석 솔루션들은 도입시에 비용을 지불하고 추후 증설 및 확장개발 등이 필요할 때 솔루션업체에 의존할 수 밖에 없다. 하지만 오픈소스를 활용함으로서 다양한 커뮤니티 등에서 지속적으로 발전할 수 있고 정부, 공공기관, 대기업뿐만 아닌 일반 중소기업들도 솔루션을 활용하여 구축할 수 있다.

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

참고 문헌

[1] 한국인터넷진흥원, “2019년 개인정보 실태점검 이슈와 계획”
[2] 김성진, 김강석, “빅데이터 분석 기술(Haddop/Hive) 기반 네트워크 정상행위 규정 방법, 2017
[3] 한빛미디어, 네트워크 보안 시스템 구축과 보안 관제, 2016
[4] 인포더북스, 차세대 정보보호 인재 양성을 위한 보안관제 실무가이드, 2017
[5] 스플렁크사 메인홈페이지, Splunk, https://www.splunk.com/ko_kr/products/ splunk-enterprise.html, Access 2020
[6] 데일리시큐, Splunk 7년 연속 SIEM부분 리더 선정, https://www.dailysecu.com/news/articleView.html?idxno=107058, 2020
[7] Larry Dignan, Splunk adds unlimited plan to enterprise pricing mix, http://www.zdnet.com/article/splunk-adds-unlimited-plan-to-enterprise-pricing-mix/, Access 2017
[8] 위키북스, 시작하세요! 하둡 프로그래밍, 2014
[9] 이상용, "아파치 엘라스틱서치 기반 로그스태시를 이용한 보안로그분석 시스템," 대전대학교 석사학위논문, 2016
[10]정호욱, Elasticsearch 검색엔진,한빛미디어(P1-5), 2014
[11] DMZ환경 구축 http://jmoon.co.kr/53
[12] kibana 시각화, https://www.youtube.com/watch?v=xPjNtd8xUZo&feature=youtu.be
[13] 정규식 패턴. https://gist.github.com/fairchild/3030472
[14] logstash에 적용하는 정규식 패턴, https://github.com/logstash-plugins/logstash-patterns-core/blob/master/patterns/grok-patterns
[15] Elk Stack 구축, https://okdevtv.com/mib/elk/elk

[16] 유기순, 임설화, 김학범 "통합로그 관리 시스템의 기술 동향과 발전 방향", 정보보호학회지 23권 제6호 2013. 12(P95)
[17]글라이더 위키, ELK 설치하기, http://libqa.com/wiki/807, Access 2017.05.25
[18] 웹서버 로그분석, https://www.youtube.com/watch?v=hTP4QEoNaLM
[19] 방화벽 설치, http://jmoon.co.kr/54?category=656679
[20] Geo-ip 사용법, http://gyrfalcon.tistory.com/entry/Logstash-geoip%EB%A5%BC-%EC%82%AC%EC%9A%A9%ED%95%B4-apache-access-log%EB%A5%BC-%EC%A7%80%EB%8F%84%EC%97%90-%ED%91%9C%EC%8B%9C%ED%95%98%EA%B8%B0
[21] 현정훈, “ 오픈소스 ELK Stack 활용 정보보호 빅데이터 분석을 통한 보안관제 구현”, 2017
[22] 한빛미디어, “데이터 시각화의 구현과 분석”, 2016
[23] Grartner , https://www.gartner.com/reviews/market/security-information-event-management/compare/elasticsearch-vs-splunk, 2019

ABSTRACT

Design and Evaluation Security Control Visualization using Elastic Stack

Seong-Yeol Yun

Dept. of Computer Science
Graduate School of Information & Communications
Hanbat National University
Advisor : Jeongho Kim

In preparation for a rapidly increasing cyber attack, we are establishing a response system for security control in public, private, military, and financial sectors in order to identify the cause and respond quickly in the event of an infringement incident. In the public sector, the National Intelligence Service established and managed the National Security Center, and in the military sector, the Ministry of National Defense established the National Defense Information Response Center. In the financial sector, the Financial Security Service (FSI) is in charge, and in the private sector, the Korea Internet Security Agency (KISA) has established an 'Internet Infringement Response Center' to provide security support for each private enterprise in each field, such as cybersecurity and response. Doing. However, in the field of security control, it is difficult to perform control in a wide range of fields, so it remains only in network security control. Companies need to select a

professional company or introduce a solution to perform security control.

This study describes how to build a security control system using open source big data solutions so that private companies can build an overall security control infrastructure. In particular, when building a security information system (SIEM: Security Information & Event Management), the infrastructure was built using Elastic Stack, one of the free open source big data analysis solutions, as a way to reduce cost and development time. , We conducted a comparative experiment with Splunk, a product that is frequently introduced into the industry. Security logs were collected, analyzed, and visualized step by step using the Elastic Stack to create a dashboard, and after entering a large log, the search speed was measured. Through this, we discovered the possibility of Elastic Stack as a big data analysis solution that can replace Splunk.

감사의 글

어느새 2년하고도 6개월간의 석사과정을 마치고 학위 논문을 제출하게 되었습니다. 미흡하지만 학위 논문을 마치면서 지난 시간 동안 저에게 도움을 주신 많은 분들께 감사의 말씀을 전합니다.

누구보다도 많은 도움을 주시고, 부족한 점이 많은 저를 언제나 세심하고 꼼꼼한 손길로 지도해주신 김정호 교수님께 진심으로 감사드립니다. 또한 논문심사로 방향정립과 개선을 위해 황경호 교수님과 김태훈 교수님께도 감사드립니다. 덕분에 직장 생활을 병행한 대학원 생활을 무사히 마칠 수 있었습니다. 또한, 학사과정부터 석사과정까지 격려와 조언을 해주신 김윤중 교수님, 김차종 교수님, 이재흥 교수님, 진영택 교수님, 김영찬 교수님, 이현빈 교수님께도 깊은 감사의 말씀 드립니다.

직장 생활 와중에도 무사히 대학원 과정을 마칠 수 있도록 도움과 격려를 아끼지 않으신 구본일 대표님, 진인장 연구원님, 손성준 과장님께도 진심으로 감사드립니다. 또한, 한국인터넷진흥원으로 이직 후 업무와 학업을 병행하는데 많은 도움을 주신 김도균 팀장님, 김병섭 선임님, 이승호 주임님, 권다운 주임님, 박한울 주임님께도 감사의 말씀을 드립니다. 또한, 항상 제 대학원 과정을 응원하고 바쁜 시간 중에도 저에게 힘이 되어준 친구들에게도 감사하다고 전하고 싶습니다.

마지막으로 제가 대학원 과정을 무사히 마칠 수 있도록 든든한 버팀목이 되어준 가족들에게 감사함을 전하며 기쁨을 함께하고자 합니다. 어머니, 아버지, 효열이 언제나 제 편이 되어주시는 가족들에게 더욱 자랑스러운 아들, 형이 되도록 항상 발전하는 사람이 되겠습니다.

2020년 6월

윤 성 열