Elastic Stack과 Kafka를 이용한 로그 분석 및 시각화 시스템 설계

윤용국, 최의인
한남대학교
ykyoon328@naver.com, eichoi@hnu.kr

Log analysis and visualization system design using Elastic Stack and Kafka

Yoon Yong Kuk, Choi Eui In
Hannam Univ.

요 약

IT기술의 발전에 따라 함께 발전하고 있는 사이버 공격으로 인해 기존보다 더 높은 보안성을 요구하고 있는 현 상황에 대응하기 위해 본 논문에서는 Elastic Search 검색 엔진을 활용한 사용자 애플리케이션, Elastic Stack 시스템을 구축한다. Apache Kafka를 통해 Elastic Stack 시스템의 Logstash와 Beat 사이의 임시 버퍼 스트림을 구현함으로써 안정성을 확보하여 웹 서버의 접근 로그를 분석하기 위한 분석환경을 구현하였다. 또한, 이러한 분석환경에 의해 도출된 분석결과를 실시간으로 모니터링하기 위해 Kibana를 활용하여 분석된 데이터를 통계 및 시각화하고 모니터링하기 위한 시각화 관제 시스템을 제안한다.

Ⅰ. 서 론

IT기술의 발전에 따라 현재 온라인 시장은 기업이나 공공기관, 교육시설 등 다양한 분야로 확대되어가고 있다. 이에 따라 생성되는 데이터양은 기하급수적으로 증가하고 데이터의 접근성이 높아져 기존보다 더 높은 보안성을 요구하고 있다.

이에 대한 예로 '2019'년 정보보호 실태조사'에 따르면 국내 기업체 2.8%가 랜섬웨어, 악성코드와 같은 사이버 공격을 받았으며, 전년 대비 0.5%나 증가했음을 보여주고 있다.[1] 또한, 국내 정보보안업계 대표 주자인 안랩(AhnLab)의 '2019년 사이버 공격 동향 통계'에 따르면 제조나 중공업 등의 대규모 산업을 노리는 공격횟수가 전년도보다 급격히 늘었으며, 특히 웹 기반의 공격 유형이 전체의 39%로 가장 많은 사이버 공격 유형으로 조사되고 있다.[2]

본 논문에서는 기존의 보안 문제를 개선하고 위와 같은 사이버 공격에 대응하기 위해 Elasticsearch 검색 엔진과 Apache Kafka를 활용하여 기존 방식보다 빠르고 안정적이게 웹 로그 데이터를 수집 및 분석할 수 있는 로그 데이터 분석 시스템을 구축하며, 분석된 로그 데이터를 시각화하여 실시간으로 모니터링이 가능한 시스템을 제안한다.

Ⅱ. 본론

2.1 관련 연구

2.1.1 로그

로그란 특정 사이트에 방문한 사용자가 언제, 어디서 어떤 활동을 하였는지에 대한 기록으로서, 보안사고나 시스템 장애 발생 시 시스템 전반의 상태를 확인할 수 있는 중요 데이터이다. IT기술의 발전으로 데이터양이 증가함에 따라 로그 데이터 또한 기하급수적으로 증가하였는데, 빅데이터가 주목받고 있는 현재 이러한 로그 데이터의 분석을 통해 발전하고 있는 사이버 공격에 대한 공격 패턴을 파악하거나, 온라인 쇼핑몰의 로그 데이터를 통한 시장 분석, 정치권의 여론 분석 등 다양한 분야에서 활용되고 있다.[3-4]

2.1.2 Elasticsearch

Elasticsearch는 Apache Lucene을 기반으로 개발한 오픈소스 분산 검색 엔진이다. 2010년 처음 발표되었으며, JSON 기반의 비정형 데이터 및 정형 데이터, 위치 정보, 메트릭 등의 다양한 유형의 데이터를 사용자가 원하는 방식으로 검색하고 결합할 수 있도록 분산 검색과 분석을 지원하고 있다. 또한, 많은 양의 데이터를 빠르고 거의 실시간(Near Real Time)으로 저장, 검색, 분석할 수 있기 때문에 인기 있는 Database 엔진으로 활용되고 있다. Elasticsearch는 비교적 간단한 설정으로 여러 대의 PC에서 분산환경을 구성할 수 있으며, 데이터의 용량과 PC 사양에 따라 원본 데이터와 복제본을 알맞게 설정하여 수평적 분산환경을 구성하기 때문에 단일서버보다 안정성이 있는 운영 환경을 제공한다. 이러한 특성으로 인해 Elasticsearch는 검색 엔진으로서 단독으로 사용되기도 하지만, 사용자 애플리케이션으로써 활용하기 위해 Kibana와 Logstash를 연동하여 Elastic Stack 시스템을 구성하여 활용된다.[5]

2.1.3 Logstash

Logstash는 다양한 종류의 로그 데이터를 수집하고 입출력 필터를 구성하여 가공처리를 할 수 있는 Elasticsearch의 플러그인 중 하나이다. 로그 데이터 수집 도구인 Filebeat가 등장하기 이전에는 Logstash에서 Input-Filter-Output의 파이프라인을 구성하여 직접 데이터를 수집 및 분류하여 Elasticsearch에 데이터를 전송하였다. 그러나 Logstash에서 담당하는 기능이 많아지고, 자원 소모량이 많아짐에 따라 데이터 수집 역할을 하는 Beats가 등장하게 되었다. 따라서 본 연구에서는 Logstash에서 직접적인 데이터를 수집하지 않고 Kafka에서 데이터를 가져오고, 플러그인을 사용한 Filter를 제작하여 Elasticsearch로 출력하는 파이프라인을 구성하도록 한다.[3]

2.1.5 Kibana

Kibana는 오픈소스 웹기반 분석 및 시각화 도구로서, Elasticsearch 서버를 운영하면서 간단하게 커스텀 대시보드를 구성하여 현재 데이터가 얼마나 들어오고 데이터 성향에 따라 증감을 확인하는 등 다양한 표현방식

을 구성하는 도구이다. 본 연구에서는 Kibana를 통해 Elasticsearch에 저장되는 많은 양의 데이터를 탐색하고 실시간으로 데이터를 분석 및 시각화 할 수 있는 시스템을 구현한다.

2.1.4 Apache Kafka

Apache Kafka는 LinkedIn에서 개발된 분산 메시지 시스템으로서 2011년에 오픈 소스로 공개되었다. Kafka는 대용량의 실시간 로그 처리에 특화되어 설계된 메시지 시스템으로서 분산 시스템을 기본으로 설계되었기 때문에 기존 메시지 시스템에 비해 분산 및 복제 구성을 손쉽게 할 수 있는 단점을 가지고 있다.

2.2 실시간 웹 로그 분석 및 모니터링 시스템 구현

```
cluster.name: test
node.name: Masternode // PC 2 : Node1, PC 3: Node2
path.data: D:₩ELK₩elasticsearch-7.13.0₩cluster₩data
path.logs: D:₩ELK₩elasticsearch-7.13.0₩claster₩logs
network.host: xxx.xxx.xxx.140 // PC 2 : 139, PC 3 : 48
discovery.seed_hosts:["xxx.xxx.xxx.140","xxx.xxx.xxx.139","xxx.xxx.xxx.48"]
cluster.initial_master_nodes: ["Masternode","Node1","Node2"]
```

그림 1 Elasticsearch 분산환경 환경설정

그림 2 Kibana를 통한 로그 데이터 모니터링 화면

본 논문에서는 Elasticsearch 분산 검색 엔진을 활용하여 웹 로그 분석 시스템을 설계하고 구현한다. Logstash와 Beat를 활용하여 웹로그를 실시간으로 수집 및 필터링하고 데이터를 Elasticsearch에 전송함으로써 로그 데이터 수집 파이프라인을 구축하도록 한다. 또한 Logstash와 Beat 사이에 Apache Kafka를 임시 버퍼 스트림으로 활용함으로써 시스템의 안정성을 확보하도록 한다. Apache Kafka는 단일서버로 구성하여, 72시간 동안 데이터를 저장하는 버퍼 스트림 역할을 하도록 설정하며, 데이터가 1GB를 넘길 경우 신규 데이터를 덮어씌워 데이터양으로 인한 문제가 생기지 않도록 설정한다. 분석된 로그 데이터는 Kibana 에서 통해 대시보드를 설계 및 구축함으로써 분석된 로그 데이터를 IP, 접속 기간, 중복 로그 비율 등 사용자가 모니터링이 가능하도록 데이터를 시각화한다. 또한, Kibana에서 Field별로 데이터 조회를 위해 사용되는 KQL(Kibana Query Language)와 Filtering을 활용하여 비정상 로그를 탐지하고 기존 시스템보다 빠른 데이터 출력이 가능하도록 개량 패턴을 적용한다.

Ⅲ. 결 론

온라인 시장의 규모가 확대되면서 기업이나 공공기관, 학교뿐만 아니라 개인적으로도 웹 서버를 사용하여 사이트를 오픈하는 사례가 점차 증가하고 있다. 다양한 웹 서버 관련 서비스가 개발됨에 따라 웹서버를 다양한 용도로 사용하고 있으며 이에 따른 수많은 웹 로그 데이터가 발생하고 있다. 이러한 사용 사례가 증가함에 따라 데이터의 접근성이 높아지고, 웹 기반의 공격도 점차 증가함으로써 기존보다 더욱 높은 보안성을 요구하고 있다. 로그 데이터 분석은 이러한 문제를 해결할 수 있는 방법 중 하나로써 방대한 양의 로그 데이터들을 필터링하고 분석하여, 외부 위협, 이상 행위 등을 탐지하는데 활용하는 것이 가능하다

본 논문에서는 Elastic Search 분산 검색엔진을 활용하여 Elastic Stack 시스템을 구축하여 실시간으로 웹 로그를 분석하고, 시각화하여 모니터링 할 수 있는 시스템을 구축하고 성능을 확인하였다. 또한, Apache Kafka를 Logstash와 Beat 사이의 임시 버퍼 스트림으로 활용함으로써 안정성을 확보하였다. 또한, 이러한 분석환경에 의해 도출된 분석결과를 실시간으로 시각화하여 모니터링하는 Kibana의 Query와 Filtering 기본 패턴을 개량함으로써 기존 Elastic Stack 시스템보다 빠른 데이터 출력이 가능하도록 하여 기존 연구와의 차별점을 두었다.

향후 연구 방향으로는 웹 로그 데이터를 Machine Learning에 접목하여, 수집된 로그 데이터 분석 내용을 통해 새로운 공격 패턴에 대한 로그 패턴을 파악하고 클라이언트 비정상 접근 로그 기반 행동 패턴을 파악함으로써 보안성 향상과 불필요한 모니터링 자원 소모 절감 효과를 얻을수 있도록 연구를 진행할 예정이다.

ACKNOWLEDGMENT

This research was supported by the Hannam University fund

참 고 문 헌

[1] 과학기술정보통신부, 한국정보보호산업협회, 2019 정보보호실태조사, pp.33-34, 2020.

[2] 김평화, “2019년 웹 취약점, 미디어 대상 사이버 공격 시도 두드러졌다”, IT조선, 2020.01.16.

[3] 이봉환, 양동민 (2018). 아파치 엘라스틱서치 기반 로그스태시를 이용한 보안로그 분석시스템. 한국정보통신학회논문지, 22(2), 382-389

[4] B. M. Choi, J. H. Kong, S. S. Hong, and M. M. Han, “The Method of Analyzing Firewall Log Data using MapReduce based on NoSQL,” Journal of the Korea Institute of Information Security & ryptology, vol. 23, no. 4, pp. 667-677, Aug. 2013

[5] Elastic Search: Introduction, Basics, Architecture and Usage of Elastic Search [Internet]