한국정보전자통신기술학회논문지(jkiiect)'18-02, Vol.11 No.1
논문 18-11-01-76 http://dx.doi.org/10.17661/jkiiect.2018.11.1.76

Splunk 플랫폼을 활용한 유해 정보 탐지를 위한 빅데이터 분석 시스템 설계

이협건*, 김영운, 김기영, 최종석

Design of Splunk Platform based Big Data Analysis System for Objectionable Information Detection

Hyeop-Geon Lee*, Young-Woon Kim, Ki-Young Kim, Jong-Seok Choi

요 약 미래 경제 성장 동력으로 부상하고 있는 사물인터넷은 이미 생활과 밀접한 분야에서는 도입이 활발하게 이루어지고 있으나, 잠재된 보안위협은 여전히 잔존하고 있다. 특히 인터넷 상의 유해 정보는 스마트홈 및 스마트시티의 활성화로 인해 폭발적으로 설치된 CCTV에 할당된 IP 정보 및 심지어 접속 포트 번호들이 포털 검색 결과 및 페이스북, 트위터와 같은 소셜 미디어 등에 공개되어 간단한 툴로도 보다 쉽게 해킹이 가능하다. 사용자들이 많이 사용하는 포털 검색 데이터 및 소셜 미디어 데이터의 보안취약점 및 불법 사이트 정보들을 데이터 분석하여, 보안취약성 같은 위험 요소가 내포된 데이터 및 사회적 문제를 야기하는 불법 사이트에 대한 대응을 신속하게 수행할 수 있게 지원하는 빅데이터 분석 시스템이 필요하다. 본 논문에서는 빅데이터 분석 시스템 설계를 위해 하둡 기반 빅데이터 분석 시스템과 스파크 기반 빅데이터 분석 시스템 연구를 통해 요구사항을 도출하여 요구사항에 맞게 Splunk 플랫폼을 활용한 유해 정보 탐지를 위한 빅데이터 분석 시스템을 설계하였다.

Abstract The Internet of Things (IoT), which is emerging as a future economic growth engine, has been actively introduced in areas close to our daily lives. However, there are still IoT security threats that need to be resolved. In particular, with the spread of smart homes and smart cities, an explosive amount of closed-circuit televisions (CCTVs) have been installed. The Internet protocol (IP) information and even port numbers assigned to CCTVs are open to the public via search engines of web portals or on social media platforms, such as Facebook and Twitter; even with simple tools these pieces of information can be easily hacked. For this reason, a big-data analytics system is needed, capable of supporting quick responses against data, that can potentially contain risk factors to security or illegal websites that may cause social problems, by assisting in analyzing data collected by search engines and social media platforms, frequently utilized by Internet users, as well as data on illegal websites.

Key Words : Big Data, Splunk, Big Data Analysis System, IoT Security, Spark, Hadoop

1. 서론

인터넷의 확산은 아동과 청소년에게 온라인 교육을 비롯하여 게임과 음악 등 문화생활과 무한한 정보에 접근할 기회를 제공해 삶을 풍요롭게 하는 반면 다양한 역기능과 역효과를 동반하기도 한다. IT 기술을 발달로 인해 언제 어디서든지 쉽게 포털 사이트 및 소셜 미디어 접속이 가능하여 불법 콘텐츠의 유통, 사이버폭력,

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning(NRF-2017R1D1A1B03034729)
*Department of Data Analysis, Seoul Gangseo Campus of Korea Polytech
Department of Data Analysis, Seoil University
Department of Computer, Soongsil University
**Corresponding Author : Department of Data Analysis, Seoul Gangseo Campus of Korea Polytech(hglee67@kopo.ac.kr)
Received January 29, 2018 Revised February 06, 2018 Accepted February 08, 2018

포르노, 온라인 게임 중독, 온라인 사기 등 아동에게 노출된 위험이 점점 더 다양해지며 심각해지고 있다. 특히 최근 발생된 CCTV 해킹 사건과 소라넷과 같은 음란 사이트 및 도박 사이트 등 유해 사이트들은 사회 질서를 어지럽히고 많은 사회적 문제를 발생시키고 있다. CCTV 해킹은 특정 사이트를 통해 해킹된 CCTV 내부의 영상이 공개되어, 개인 사생활 및 정보 보안에 큰 위협이 되고 있다. 뿐만 아니라, 해킹된 CCTV들은 디도스 공격하는 악성 좀비 PC 역할도 수행하였다. 이러한 CCTV 해킹은 구글과 같은 포털의 검색 데이터에서 CCTV의 IP 및 포트 정보 등 주요 정보 정보가 노출되어 간단한 프로그램으로 쉽게 해킹이 가능했다. 도박, 음란 등 유해 사이트들은 한국인터넷진흥원, 사이버경찰청, 방송통신심의위원회에서 발견될 때마다 URL 차단을 수행하고 있으나, 도메인 주소 변경 및 우회 경로 접속 기법을 통해 여전히 성행하고 있다. 따라서 구글과 같은 포털 사이트의 검색 데이터로부터 위험 요소가 내포된 유해 정보 차단을 위한 빅데이터 분석 기술이 필요하다.

이에 본 연구에서는 빅데이터 솔루션 중 세계 점유율 1위인 Splunk Enterprise를 이용하여 유해 정보 탐지를 위한 빅데이터 분석 시스템을 설계한다. 제안하는 빅데이터 분석 시스템은 인터넷 및 IT 기기의 기술이 발달함에 따라 증가되고 있는 보안취약성과 같은 위험 요소가 내포된 데이터 및 불법 사이트 데이터와 같은 유해 정보를 포털 사이트와 소셜 미디어로부터 빅데이터를 수집한 뒤, 그 데이터들을 빅데이터 분석하여 발견된 보안취약점 및 불법 사이트 정보에 대해 시각화를 수행한다. 제안하는 빅데이터 분석 시스템으로부터 생성된 시각화된 정보는 방송통신심의위원회, 사이버경찰청, 한국인터넷진흥원 및 관련 기업에게 활용되어 즉각적인 대응을 할 수 있도록 지원한다.

본 논문의 구성은 다음과 같다. 2장에서는 표준화 동향에 대해 살펴본다. 3장에서는 본 논문에서 제안하는 Splunk를 이용한 유해 정보 탐지를 위한 빅데이터 분석 시스템을 설계한다. 마지막 4장에서는 결론 및 향후 연구 과제를 제시한다.

2. 관련 연구

본 장에서는 기존에 사용되고 있는 하둡 기반의 빅데이터 분석 시스템과 스파크 기반 빅데이터 분석 시스템에 대해 살펴보고, 이 결과를 기반으로 빅데이터 분석 시스템에 필요한 요구사항을 도출한다.

2.1 하둡 기반 빅데이터 분석 시스템

하둡 기반 빅데이터 분석 시스템은 빅데이터 플랫폼 중 많이 알려진 빅데이터 모델로 다양한 하둡 서브 프로젝트들과 융합되어 사용되고 있다. 다양한 하둡 서브 프로젝트들과 융합되어 구성된 빅데이터 분석 시스템은 일반적으로 하둡 에코 시스템으로도 부른다.

하둡 기반 빅데이터 분석 시스템의 장점은 다양한 하둡 프로젝트들과 융합되어 시너지 높은 빅데이터 모델이 될 수 있다. 그러나 하둡 기반 빅데이터 분석 시스템의 단점은 앞서 설명한 다양한 하둡 프로젝트들과 융합에 있어 많은 문제점이 야기된다. 하둡 기반 빅데이터 분석시스템은 새롭게 배포되는 하둡 버전에 따라 구성되는 서브 프로젝트들의 호환성에 문제가 발생한다. 특히 하둡 2.8.2 버전은 하둡의 핵심 기술인 하둡분산파일시스템과 맵리듀스에서 얀이 추가되었다. 이로 인해 하둡 기반 빅데이터 분석 시스템은 다양한 빅데이터 기술들을 융합하고 유지 관리하는데 어려움이 많다.

2.2 스파크 기반 빅데이터 분석 시스템

스파크 기반 빅데이터 분석 시스템은 앞서 설명한 하둡 기반 빅데이터 분석 시스템의 문제점 중 하나인 실시간 빅데이터 분석을 제공하기 위해 개발된 빅데이터 모델이다.

스파크 기반 빅데이터 분석 시스템의 장점은 메모리를 활용하여 빅데이터 저장·처리를 수행하기 때문에 맵리듀스를 활용하는 하둡 기반 빅데이터 분석 시스템에 비해 빅데이터 처리 성능은 약 30배 이상 차이난다. 스파크 기반 빅데이터 분석 시스템의 단점은 빅데이터를 분석하기 위해 원천 데이터를 RDD로 변경하여 메모리로 데이터를 처리하기 때문에 빅데이터 분석을 위해 인프라 구축하는 발생되는 비용이 매우 크다. 따라서 스파크 기반 빅데이터 분석 시스템은 실시간 분석이 필요

한 경우의 서브 분석에만 사용되고 있다.

2.3 빅데이터 분석 시스템 요구사항

빅데이터 분석 시스템이 갖추어야 할 사항들은 크게 3가지로 구분된다.

첫째, 빅데이터 분석 시스템들이 처리하는 데이터의 양과 서버의 규모는 기존의 분석 시스템들과는 달리 그 규모가 매우 크다. 또한 빅데이터 분석 시스템의 대부분의 핵심적인 사항들은 비용적인 측면에서의 효율성, 지속적으로 증가하는 데이터들의 수용, 빈번하게 발생하는 장애에 대한 대처와 인프라 관리의 편리성이 요구된다.

둘째, 빅데이터 분석 시스템은 빅데이터 수집·저장·분석·시각화로 구성되며, 각 항목별 다양한 서브 프로젝트들이 존재한다. 다양한 서브 프로젝트들은 서로 융합되어 사용되기 때문에 서브 프로젝트의 버전에 대한 변화에 따른 유연한 호환성이 요구된다.

셋째, 빅데이터 분석 시스템은 빠른 데이터 분석이 필요한 부분에 선택적인 실시간 분석 기능이 적용되어야 한다. 실시간 빅데이터 분석은 많은 인프라 환경 구축 비용이 발생하기 때문에 메모리를 기반으로한 빅데이터 분석이 필요한 경우에만 사용되어야만 한다.

3. 제안하는 빅데이터 분석 시스템

제안하는 빅데이터 분석 시스템은 앞서 설명한 빅데이터 분석 시스템의 요구사항에 맞춰 설계하며, 개발한 빅데이터 수집 알고리즘이 적용된 Crawling Manager를 이용하여 포털 사이트 및 소셜 미디어의 빅데이터를 수집하고, 그 수집된 빅데이터를 Splunk를 이용하여 빅데이터를 분석한다. 또한 분석된 빅데이터를 기반으로 구현‘한 시각화 프레임워크를 이용해 시각화를 수행한다. 제안하는 빅데이터 분석 시스템은 크게 Collector Layer, Splunk Layer와 User Interface Layer로 구성된다. [그림 1]은 제안하는 빅데이터 분석 시스템의 구조를 나타낸다.

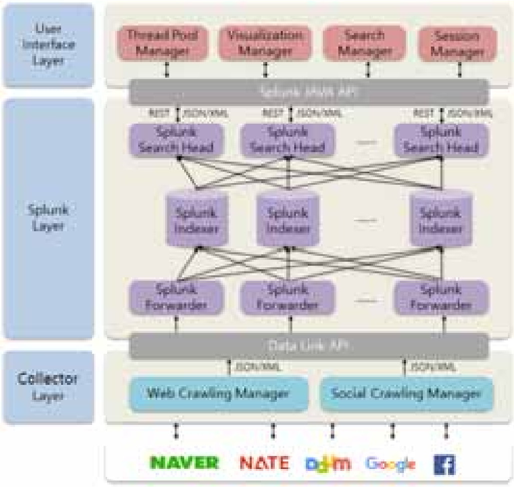

그림 1. 제안하는 빅데이터 분석 시스템 구조
Fig. 1. Architecture of proposal big data analysis system

3.1 Collector Layer

Collector Layer는 분석에 필요한 포털 사이트 및 소셜 미디어 데이터를 수집하는 영역으로 데이터 수집 기능과 수집된 데이터 전달 기능을 수행한다. 데이터 수집 기능은 연구 개발한 수집 알고리즘이 적용된 Web Crawling Agent와 Social Crawling Agent를 이용하여 수행한다. Web Crawling Agent는 포털 사이트의 데이터를 수집하는 역할을 수행하며, Social Crawling Agent는 페이스북의 데이터를 수집하는 역할을 수행한다. Crawling Manager를 2가지로 구분해서 구현하는 이유는 포털 사이트의 데이터와 소셜 미디어 데이터의 구조의 차이가 많아 최적화된 역할을 수행하기 위해 구분하여 구현한다.

3.2 Splunk Layer

Splunk Layer는 Splunk Enterprise를 이용하여 빅데이터 분석하는 영역으로 빅데이터 분석 기능과 외부 시스템 연동 기능을 수행한다. 빅데이터 분석 기능은 Splunk의 주요 모듈인 Splunk Forwarder, Splunk Indexer, Splunk Search Head를 이용하여 수행한다. 외부 시스템 연동 기능은 Splunk에서 제공하는 JAVA API를 이용하여 외부 시스템인 User Interface Layer와 연동하며 REST 방식을 통해 데이

터 전달을 수행한다. Splunk Forwarder는 Splunk에서 데이터 수집을 위해 사용되는 모듈로 수집된 데이터를 Splunk Indexer로 전달하는 기능을 수행한다. Splunk Forwarder는 Collector Layer의 Data Link API로 전달받은 수집된 데이터를 Splunk Indexer에 적합한 데이터 형태로 변환하여 저장하는 역할을 수행한다. 제안하는 빅데이터 분석 시스템에서 포털 데이터와 소셜 미디어 데이터를 Splunk Forwarder를 이용하여 수집하지 않고 별도의 Collector Layer를 이용하여 구현한 이유는 Splunk는 Crawling을 지원하지 않으며, 기존 구축된 시스템에 수집 모듈을 설치하여 수집하는 방식이기 때문이다. Splunk Indexer는 Splunk가 데이터 분석을 수행하기 위한 데이터가 저장되는 저장소이다. Splunk 처리 방식 특성상 Splunk Indexer를 여러 노드로 분산시켜 구현할수록 분석 성능은 향상된다. Splunk Search Head는 Splunk에서 데이터 분석을 수행하기 위해 사용되는 모듈로 Splunk에서 지원하는 SPL(Search Processing Language, 검색 처리 언어)를 이용하여 데이터 분석을 수행한다. Splunk의 처리 속도는 SPL의 작성 방식에 따라 성능 차이가 크게 발생하기 때문에 유해 정보 탐지를 위한 최적화된 SPL을 개발한다. 또한 Splunk의 처리 속도는 Splunk의 3가지 주요 검색 패널인 Source, Sourcetype과 Host에 따라 시스템을 적절하게 분리하여 사용해야만 요구하는 속도가 나타난다.

3.3 User Interface Layer

User Interface Layer는 사용자 편의성을 높이기 위한 영역으로 Visualization Manager, Search Manager, Thread Pool Manager와 Session Manager로 구성된다. Visualization Manager는 시각화를 수행하는 모듈로 Splunk로부터 전달받은 빅데이터 분석 결과를 기반으로 사용자에게 보다 쉽게 이해시키기 위해 차트 및 표 등 데시보드 형태로 정보를 제공한다. Visualization Manager 구현 기술은 jsChart 등 JSON, XML 데이터 처리가 가능한 시각화 프레임워크를 이용한다. Search Manager는 빅데이터 분석을 위한 SPL을 작성하고 요청하고 결과를 받는 역할을 수행한다. 또한 Search Manager는 검색 빈도수가 높은 SPL을 저장하여 편리성을 제공한다. Thread Pool Manager는 시각화 수행 및 SPL 요청 등 동시 실행중인 Thread를 관리하는 역할을 수행한다. Session Manager는 사용자들의 접속 관리하는 역할을 수행한다.

4. 성능평가

본 논문에서의 성능평가는 제안하는 시스템이 데이터 처리율 및 타 시스템과의 연계 호환성에 적합한지 타당성 분석을 수행한다.

4.1 데이터 처리율

데이터 처리율은 수집된 유해 데이터를 수집하는 Collector Layer에서부터 Splunk Layer까지 빅데이터 처리과정을 측정한다. [표 1]은 데이터 처리율에 사용되는 변수를 나타낸다.

표 1. 데이터 처리율에 사용되는 변수
Table 1. Parameter for using data throughput

구분	정의 변수
데이터 처리 실패로 인한 손실 데이터	l
평균적으로 발생하는 손실 횟수	α
수집되는 데이터	n
Collector Layer 데이터 처리율	cp
Splunk Layer 데이터 처리율	sp

Collector Layer에서의 데이터 처리율은 수식(1)과 같이 나타낸다.

$$CP_l = \frac{\alpha^{l-\alpha}}{l!} = \rho\gamma^{-\alpha}\prod_{l=1}^{l}\left(\frac{\alpha n^2}{l}\right)\ \alpha = n \quad \text{수식(1)}$$

Splunk Layer에서의 데이터 처리율은 수식(2)과 같이 나타낸다.

$$SP_l = \frac{\alpha^{l-\alpha}}{l!} = \sum_{n}^{\alpha^{l-n}} cp\frac{l}{n} = {}^{-\alpha}\prod_{l=1}^{l}\left(\frac{\alpha}{l}\right) \quad \text{수식(2)}$$

데이터 처리율은 앞서 설명한 cp와 sp의 처리율을

데이터 처리횟수로 나눠 분석한다. 데이터 처리율은 수식(3)과 같이 나타낸다.

$$D_t = \sum_{n=1}^{\infty} \frac{(cp+sp)}{n!} \qquad \text{수식(3)}$$

[그림 2]는 데이터 처리율을 나타낸다.

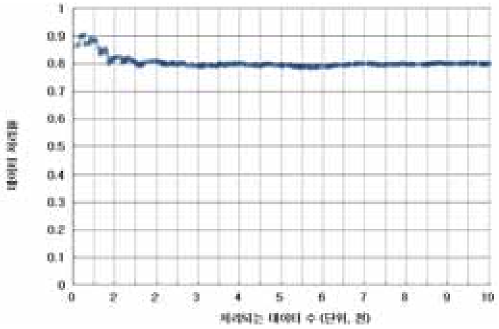

그림 2. 데이터 처리율
Fig. 2. Data throughput

분석 결과, 데이터 처리율은 동시에 처리되는 데이터가 약 1,000개인 경우에는 0.9로 빅데이터 처리에 상당히 높은 데이터 처리율을 보였다. 또한 동시 처리되는 데이터가 약 1,600개부터 약 0.87로 서서히 감소하였으며, 약 2,000개부터 약 0.8로 처리율이 초반보다 약 0.1 감소하였다. 그러나 동시에 처리되는 데이터 수가 약 10,000개까지 증가되어도 데이터 처리율은 약 0.8로 데이터 처리율이 크게 감소하지 않았다. 따라서 빅데이터 분석 시스템의 특성상 많은 데이터를 수집, 저장, 분석함에 있어 제안하는 시스템은 데이터 처리율에 있어 적합함을 알 수 있다.

4.2 타 시스템과의 연계 타당성

타 시스템과의 연계 타당성은 앞서 설계한 Collector Layer와 Splunk Layer의 연계 타당성과 Splunk Layer와 User Interface Layer의 연계 타당성을 분석한다.

Collector Layer와 Splunk Layer의 연계 타당성은 Collector Layer에서 사용되는 웹 크롤러와 소셜 크롤러를 통해 데이터를 파일형태로 저장하고, 그 수집된 데이터를 Splunk Layer의 Splunk Forwarder를 통해 데이터를 Splunk Layer로 전달한다. Splunk Forwarder는 파일 형태의 데이터를 전달하기에 적합하며, 웹 크롤러와 소셜 크롤러가 파일을 생성할 때마다 전달한다. 또한 Splunk Forwarder는 전달하는 파일을 설정한 Splunk Indexer에 적합한 형태로 변환하기 때문에 Collector Layer와 Splunk Layer의 연계는 적합하다.

Splunk Layer와 User Interface Layer의 연계 타당성은 Splunk Layer에서 제공하는 Java API를 활용하여 User Interface Layer와 연계한다. User Interface Layer는 Splunk Layer로부터 JSON 및 XML 형태로 데이터를 전달하기 때문에 다양한 형태로 데이터 표현이 가능하며, 데이터 처리가 가능하다. 따라서 Splunk Layer와 User Interface Layer의 연계는 적합하다.

5. 결론

본 논문에서는 Splunk 플랫폼을 활용한 유해 정보 탐지를 위한 빅데이터 분석 시스템 설계하였다. 제안하는 시스템은 인터넷 및 IT 기기의 기술이 발달함에 따라 증가되고 있는 보안취약성과 같은 위험 요소가 내포된 데이터 및 불법 사이트 데이터와 같은 유해 정보를 포털 사이트와 소셜 미디어로부터 빅데이터를 수집한 뒤, 그 데이터들을 빅데이터 분석하여 발견된 보안취약점 및 불법 사이트 정보에 대해 시각화를 수행한다. 제안하는 빅데이터 분석 시스템으로부터 생성된 시각화된 정보는 방송통신심의위원회, 사이버경찰청, 한국인터넷진흥원 및 관련 기업에게 활용되어 즉각적인 대응을 할 수 있도록 지원한다. 향후 본 설계를 기반으로 시스템이 구현이 필요하다.

REFERENCES

[1] Hye-Jung Chang and Do-Nyun Kim, "A Study on big data utilization for implementation of the resident participation type safe community planning of the smart

city," Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 9, No. 5, pp. 478~495, Oct, 2016.

[2] In-Hak Joo, "Spatial Big Data Query Processing System Supporting SQL-based Query Language in Hadoop," Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 10, No. 1, pp. 1~8, Feb, 2017.

[3] Eun-Hee Jeong and Byung-Kwan Lee, "A Design of Hadoop Security Protocol using One Time Key based on Hash-chain," Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 10, No. 4, pp. 340~349, Aug, 2017.

[4] Jae-Hyuck Kwak, Sangwan Kim, Taesang Huh and Soonwook Hwang, "Implementation and Performance Analysis of Hadoop MapReduce over Lustre Filesystem," KIISE Transactions on Computing Practices, Vol. 21, No. 8, pp. 561~566, Aug, 2015.

[5] Deoksang Kim, Hyeonsang Eom and Heonyoung Yeom, "Performance Optimization in GlusterFS on SSDs," KIISE Transactions on Computing Practices, Vol. 22, No. 2, pp. 95~100, Feb, 2016.

[6] Jik-Soo Kim, Nguyen Cao, Seoyoung Kim and Soonwook Hwang, "Design of a Large-scale Task Dispatching & Processing System based on Hadoop," Journal of KIISE, Vol. 43, No. 6, pp. 613~620, Jun, 2016.

[7] HyunJo Lee, TaeHoon Kim and JaeWoo Chang, "A MapReduce-based kNN Join Query Processing Algorithm for Analyzing Large-scale Data," Journal of KIISE, Vol. 42, No. 4, pp. 504~511, Apr, 2015.

[8] Areum Lee, Jiseon Bang and Yoonhee Kim, "A Design of a TV Advertisement Effectiveness Analysis System Using SNS Big-data," KIISE Transactions on Computing Practices, Vol. 21, No. 9, pp. 579~586, Sep, 2015.

저자약력

이 협 건(Hyeop-Geon Lee) [종신회원]

2011.03 – 2015.08,
승실대학교 일반대학원
컴퓨터학과 공학박사
2015.12 – 현재,
한국폴리텍대학 서울강서캠퍼스
데이터분석과 조교수

<관심분야>
빅데이터, 실시간 분석, 데이터분석

김 영 운(Young-Woon Kim) [종신회원]

2004.09 – 2018.08,
승실대학교 일반대학원
컴퓨터학과 공학박사
2015.12 – 현재,
한국폴리텍대학 서울강서캠퍼스
데이터분석과 조교수

<관심분야>
빅데이터, 실시간 분산 처리, 데이터분석

김 기 영(Ki-Young Kim) [종신회원]

2004.03 - 현재,
서일대학교 소프트웨어공학과
부교수

<관심분야>
빅데이터, 무선통신, 사물인터넷

최 종 석(Jong-Seok Choi) [정회원]

2010.03 – 2012.02,
승실대학교 일반대학원
공학석사
2013.03 – 2015.02,
승실대학교 일반대학원
공학박사수료

<관심분야>
빅데이터, 무선통신, 5G