하둡 보안 기술과 취약점 분석

김아용* · 하의륜*· 김한길** · 박만섭** · 정회경***
*배재대학교

Hadoop Security Technologies and Vulnerability Analysis

A-Yong Kim* · He Yilun* · Han-Kil Kim** · Man-Seub Park** · Hoe-Kyung Jung***
*Paichai University
E-mail : janlssary@pcu.ac.kr, mrheyilun@gmail.com, MIE-khg0482@yahoo.co.kr, ceo@egluon.com, hkjung@pcu.ac.kr

요 약

스마트폰이 보급되어 빅 데이터(Big Data) 시대를 맞이하였고, 페이스북(FaceBook)이나 트위터(Twitter)같은 SNS(Social Network Service)를 실생활에서 일상화되어 사용하고 있다. 여기서 발생하는 SNS의 비정형 데이터를 버리지 않고 분석 및 추출하고 활용하기 위해서 아파치 재단에서 개발된 하둡(Hadoop)을 활용하고 있다. 하둡은 대량의 자료를 처리할 수 있는 오픈 소스 프레임워크(Open Source Framework)이다. 하둡은 국내의 기업에서도 도입하고 있으며 현재 개발 및 상용하고 있다. 그러나 하둡은 기술 개발에 비해 보안 분야는 미흡하다는 지적을 받고 있다. 이에 본 논문에서는 하둡의 보안 기술과 취약점을 분석하고 보안을 향상시키는 방법을 제안한다.

ABSTRACT

And were the prevalence of smartphones is the Big Data era, such as Facebook or Twitter, SNS (Social Network Service) routine is used in the real world. Take advantage of the analysis, and to extract and utilize developed in the Apache Foundation Hadoop (Hadoop) without abandoning the SNS unstructured data here. Hadoop is an open source framework that can handle large amounts of data. Hadoop has been introduced in the domestic corporate and commercial development and Compared to the technology development Hadoop has been pointed out that the lack of security sector. In this paper, we propose a method to enhance the security and vulnerability analysis of security technologies and Hadoop.

키워드

하둡, 보안, 오픈소스, 취약점

Ⅰ. 서 론

리서치 전문 회사인 가트너(gartner)에서는 2013년 10대 전략 기술(IT Trend)에 전략적 빅데이터(Strategic Big Data)를 포함시켰다[1]. 빅데이터란 IT 기술이 발달하면서 여러 IT 기기(Device), SNS에서 생성되는 데이터를 말한다. 여기서 생성되는 데이터는 정형데이터 보단 비정형데이터가 비율이 높다. 그리고 생성되는 데이터 발생량은 예측하기 힘들 정도의 거대한 데이터이다. 이러한 빅데이터는 기존의 처리방식으로는 감당할 수 없어서 버려져 왔거나 많은 비용을 들여서 처리해야 했던 데이터이다. 하지만 하둡이 등장하면서 비정형데이터는 IT 금맥이 되기 시작하였다. 하둡은 아파치(Apache) 재단에서 더그 커팅에 의해 시작하여 개발된 오픈소스 프레임워크이다. 하둡은 구글(Google)에서 논문으로 발표된 구글 파일 시스템(GFS)과 맵리듀스(MapReduce)를 본 받아서 개발되었다. 하둡은 거대한 파일을 분산하여 여러 서버에 보내서 처리하기 때문에 처리속도가 빠르고 저비용으로 처리할 수 있다는 장점을 가지고 있다. 하지만 이러한 생태계를 갖고 있기 때문에 보안을 신경써야하는 부분도 늘어난다는 단점도

가지고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 하둡과 맵리듀스, 하둡 분산 파일 시스템을 살펴보고, 3장에서는 하둡이 사용하고 있는 인증 방법과 취약점을 설명한다. 4장에서는 상용화 되어 있는 툴을 설명하고 마지막 5장에서는 결론 및 향후연구과제에 대해 기술한다.

Ⅱ. 관련연구

2.1 하둡(Hadoop)

아파치 하둡은 하둡 분산 파일시스템(HDFS : Hadoop Distributed File System)과 맵리듀스로 구성된 아파치 루씬(Lucene) 프로젝트중 하나로 오픈소스 웹 검색엔진인 아파치 너치(Nutch)에서 탄생하였다. 하둡은 자바(Java)로 개발된 오픈 소프트웨어로서 JDK(Java Development Kit)만 설치되어 있으면 이기종간의 제약을 받지 않으므로 서버 확장(몇 대에서 수천 대)이 자유로운 편이며, 하둡에 여러 프레임워크를 연동하여 사용할 수 있다.

2.2 맵리듀스(MapReduce)

하둡의 맵리듀스는 방대한 양의 데이터를 병렬로 처리하게 해주는 프로그래밍 모델이다. 맵리듀스란 이름은 함수형 언어의 맵(Map)과 리듀스(Reduce) 함수에서 유래되었으며, 하둡 분산 파일 시스템에 저장된 데이터를 읽고 분석한 후에 결과를 다시 하둡 분산 파일 시스템에 저장한다. 맵리듀스 프레임워크는 잡트래커(JobTracker)가 단일 마스터와 클러스터 당 태스크트래커(Task-Tracker)가 슬레이브(slave)로 구성된다. 마스터(Master)는 잡의 태스크를 스케줄링하는 기능을 수행하고 실패한 태스크를 다시 실행하게 하고 모니터 기능을 수행한다. 슬레이브는 마스터의 지시대로 작업을 수행한다[2]. 하둡 프레임 워크는 자바만 지원하지만 맵리듀스 프로그램은 자바뿐만 아니라 루비, 파이썬, C++로도 지원해준다.

2.3 하둡 분산 파일 시스템(HDFS)

하둡 분산 파일 시스템은 이기종간의 하드웨어로 구성된 클러스터에서 대용량 데이터 처리를 위하여 개발된 분산 파일 시스템이다. 기존의 대규모 처리 시스템을 구축하려면 데이터 유실을 방지하기 위해 스토리지와 네트워크 이중화 작업, 병목현상을 방지하기 위하여 고속의 SAN을 구축 해야 되서 비효율적이였다. 하지만 하둡 분산 파일 시스템을 이용하면 이러한 작업 없이 구축이 가능하여 비용대비 효율을 얻을 수 있다.

하둡 분산 파일 시스템은 파일을 블록 단위로 분할하여 여러 데이터 노드에 저장한다. 이때 하드웨어 장애를 대비하여 다른 데이터 노드에 파일을 복제한다. 복제본의 수는 설정보다 적으면 자동적으로 복사본을 만들어 내며 데이터노드 장애가 발생하면 해당 데이터는 다른 노드에서 불러와 실행하여 기존 스토리지의 기능을 대신해준다.

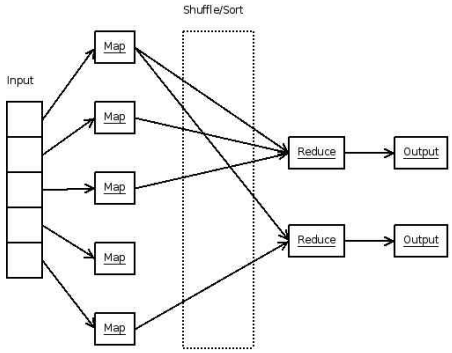

그림 1. Hadoop MapReduce[2]

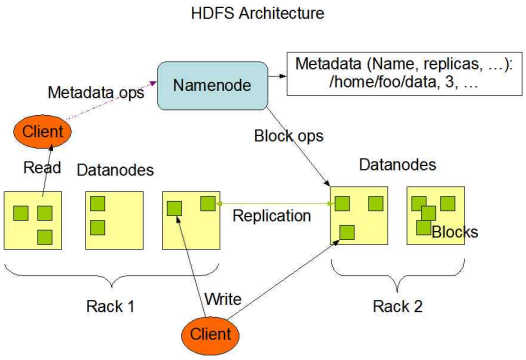

그림 2. HDFS Architecture[2]

Ⅲ. 하둡 보안 인증과 취약점

3.1 하둡 보안 인증

하둡의 초기 버전은 안전한 환경에서 사용한다고 가정하여 데이터 접근 방지 보다는 불의의 사고로 인해 데이터의 유실을 막기 위한 접근 제한 조치가 설계되어 보안 기능이라고 보기 어려웠다[3]. 하둡의 보안 기능은 Hadoop 1.0에 추가되었으며 커버로스 인증과 위임 토큰을 사용하여 보안을 강화하게 되었다.

3.2 하둡 보안 취약점

하둡 에코시스템(EchoSystem)은 여러 프레임워크로 구성된다. 이러한 구성은 복잡한 생태계를 구성하여 취약점이 높아진다. 만약 에코시스템 중에서 보안 부분의 취약점이 발생하면 공격자에 의해 노출되기 쉽다. 또한 하둡은 자바로 구현되어 있기 때문에 자바의 취약점을 갖고 있으며, 커버로스를 이용하여 인증을 하기 때문에 커버로스의 취약점도 갖고 있다.

Ⅳ. 하둡 상용 툴

빅데이터에 관련된 벤처기업 및 대기업에서는 하둡을 기반으로 솔루션을 개발하고 현재 상용화중이다. EMC사의 Security Analytics[4], Gruter사의 qoobah[5], Intel사의 Apache Had- oop[6] 등이 있다. 이러한 솔루션은 빅데이터를 도입해야 하지만 기술 및 시간이 부족한 회사에는 유용하게 이용할 수 있다. 하지만 상용화되어 있는 툴이라서 비용은 지불해야 하는 단점이 있다.

Ⅴ. 결론 및 향후연구과제

본 논문에서는 빅데이터를 처리하기 위해 많이 사용하고 있는 하둡의 보안 문제를 정리해보고 상용화 되어 있는 제품을 소개하였다. 하둡은 기존의 시스템을 교체 하는게 아니라 기존 시스템에서 추가되는 개념으로 생각하여 도입을 무리하게 진행하지 말고 신중하게 결정하여 도입해야 한다. 하둡 보안을 무리하게 진행하면 성능 저하가 발생할 수도 있으므로 성능 저하를 최소화하면서 보안을 강화하는 방향으로 연구해야 한다. 향후 연구과제는 하둡에서 보안을 덜 중시해도 되는 부분과 보안이 꼭 필요한 부분으로 분리하여 효율적이면서 보안을 구성할 수 있는 연구 과제이다.

참고문헌

[1] Gartner, http://www.gartner.com
[2] Apache Hadoop, http://hadoop.apache.org
[3] T. White, Hadoop: The Definitive Guide
[4] EMC, http://www.emc.com
[5] Gruter, http://www.gruter.com
[6] Intel, http://www.intel.com