분산 하둡 환경에서 스프링 프레임워크 기반의 웹 스크래핑 데이터 수집 처리 시스템 개발

이명호
세명대학교 스마트IT학부
e-mail:mhlee@semyung.ac.kr

Development of Web Scraping Data Collection Processing System Based on Spring Framework in Distributed Hadoop Environment

Myeong-Ho Lee
School of SmartIT, Semyung University

요 약

다양한 디바이스의 출현으로 정형/비정형 데이터의 가파른 증가는 많은 영역에서 빅데이터를 새로운 활용에 따른 다양한 서비스를 요구하고 있다. 그러나 현재까지 대부분의 연구는 기술 동향 연구나 데이터 처리 설계 및 구현 연구나 샘플 데이터를 통하여 수집과 정제를 가정한 후, 처리 및 탐색, 분석 및 응용을 위한 연구였다.

따라서 본 연구에서는 스프링 프레임워크 환경에서 웹스크래핑을 통한 데이터 수집을 위하여 가상화 환경에서 3대의 리눅스 서버를 통하여 하둡 2.0을 기반으로 비정형 데이터를 수집한 후, 하둡 분산 파일 시스템과 HBase에 적재하고, 적재된 비정형 데이터를 기반으로 트위터 Okt 형태소 분석기를 통하여 정형화 데이터를 관계형 데이터베이스에 저장할 수 있게 설계하고 구현하였다. 향후 Frontend를 React.js/Vue.js, Thymeleaf, Cross Platform 환경의 Flutter로 확장하여 웹이나 모바일로 실제 적용 사례를 통하여 서비스 될 수 있는 풀스택 인공지능, 머신 러닝, 딥러닝 연구가 지속되어야 할 것이다.

1. 서론

대용량 빅데이터 처리를 여러대의 컴퓨터에서 분산 처리할 수 있게 해주는 프레임워크가 하둡(Hadoop)이다. 2011년 발표된 하둡 1.0은 맵리듀스를 실행할 때 맵리듀스 작업 갯수를 관리했기 때문에 클러스터 전체 사용률이 낮은 단점이 있었다. 하둡 2.0은 맵과 리듀스의 관계가 1:1 관계가 아니며 하나 이상의 컨테이너를 실행할 수 있기 때문에 잡들의 워크로드를 고려하여 리소스 설정을 진행할 수 있는 장점이 있다. 2017년 하둡 3.0이 발표되었다[1,2].

차세대 웹 표준를 대비하여 우리나라에서도 2008년도부터 오픈소스를 적극 활용한 전자정부 표준프레임워크를 구성하였다. 2022년 현재 Spring Framework 5.3.6과 Spring Boot 2.4.5를 지원하는 전자정부 표준프레임워크 실행환경 4.0.0 업데이트를 발표하였다[3].

현재까지 대부분의 연구는 기술 동향 연구나 데이터 처리 설계 및 구현 연구나 샘플 데이터를 통하여 수집과 정제를 가정한 후, 처리 및 탐색, 분석 및 응용을 위한 연구였다[4]. 그러나 스프링 프레임워크를 기반으로 실시간 스크래핑을 통하여 비정형 빅데이터를 수집한 후 형태소 분석기를 통하여 비정형 빅데이터를 정제하고, 의미있는 정형화된 빅데이터로 변환하여 관계형 데이터 베이스로 저장 처리하는 분산 하둡 플랫폼 설계 연구는 많이 미비하였다.

따라서 본 연구에서는 스프링 프레임워크 환경에서 3대의 가상 머신 서버을 통하여 웹스크래핑으로 하둡 2.0을 기반으로 키워드로 검색한 비정형 데이터를 수집하고 수집된 빅데이터를 HBase에 적재한다. 그리고 적재된 비정형 데이터를 기반으로 트위터 Okt 형태소 분석기를 통하여 정형화 데이터를 오라클 관계형 데이터베이스에 저장할 수 있게 설계하고 구현하였다.

2. 구현 기술의 현황

2.1 Web Scraping

웹 스크래핑(Web Scaping)이란 웹 사이트 상에서 원하는 정보를 자동으로 추출하여 수집하는 기술이다. 웹 크롤링도 일종의 웹 스크래핑 기술이다. 웹 크롤링은 조직적으로 자동화된 방법으로 WWW를 탐색하는 컴퓨터 프로그램인 웹 크

롤러가 하는 작업을 말하며 여러 인터넷 사이트의 페이지들을 브라우징하는 행위이다. 파싱이란 웹 페이지들의 자료에서 데이터의 특정 패턴들을 추출한 후, 추출된 정보를 가공하는 것이다[5]. 본 연구에서는 스프링 프레임워크 상에서 웹 스프레핑을 통하여 수집된 비정형 빅 데이터들을 형태소 분석기를 통하여 원하는 데이터들로 정형화하여 분류하도록 한다.

2.2 형태소 분석기

형태소 분석이란 형태소 보다 큰 어절이나 문장을 최소 의미 단위로 분절하는 과정이다. 한국어 자연어 처리(NLP, Natural Language Processing)에서 형태소 분석/품사 태깅을 통한 토큰화는 아주 중요하다. 현재 오픈 소스의 한국어 형태소 분석기(Morphological Analyzer) 중에서 분석 결과로 관심이 있는 것으로 MeCab[6], Okt[7], khaiii[8] 등이 있다. 본 연구에서는 성능이 뛰어난 Okt 형태소 분석기를 사용하여 수집된 비정형 데이터를 오라클 RDBMS에 정형화된 데이터를 저장하도록 한다.

2.3 Spring Framework

전자정부 표준 프레임워크로 채택하여 운영하는 프레임워크가 Spring Framework이다. 스프링 프레임워크의 핵심 요소는 애플리케이션 수준의 비즈니스 로직에만 집중할 수 있도록 한다[9]. Spring MVC는 각 컴포넌트들의 역할이 명확하게 분리하여 Backend와 Frontend가 동시에 개발할 수 있게 할 수 있는 디자인 패턴이다.[10].

따라서 본 연구에서는 분산 빅데이터 풀스택 플랫폼를 위하여 전자정부 표준 프레임워크인 스프링 프레임워크를 기반으로 설계하였다.

2.4 MyBatis

MyBatis는 데이터베이스 레코드에 기본 타입과 Map 인터페이스 그리고 Java POJO를 설정해서 매핑하기 위해 XML과 애노테이션을 사용할 수 있다[11]. 본 연구에서는 스프링 프레임워크 환경에서 웹스크래핑을 통하여 수집 및 처리된 비 정형 데이터를 정형화하여 향후 분석 및 응용을 위하여 영구 저장하기 위한 효율적인 SQL Mapper로 MyBatis를 적용하였다.

2.5 Hadoop

하둡에서 NameNode는 HDFS의 Master 역할을 하며, Slave 역할을 하는 DataNode에게 I/O작업을 할당한다. Secondary NameNode는 NameNode의 블록정보를 합칠 때 사용하기 위한 Node이다.

Resource Manager는 Cluster전반의 자원 관리와 테스크들의 스케쥴링을 담당한다. Application Manager는 Node Manager에서 특정 작업을 위해서 Application Master를 실행하고 상태를 관리한다. Resource Tracker는 Container가 아직 살아 있는지 확인하기 위해서, Application Master 재시도 최대 횟수, 그리고 Node Manager가 죽은 것으로 간주 될 때까지 얼마나 기다려야 하는지 등과 같은 설정 정보를 가지고 있다. DataNode에서 실제 데이터가 저장된다[12].

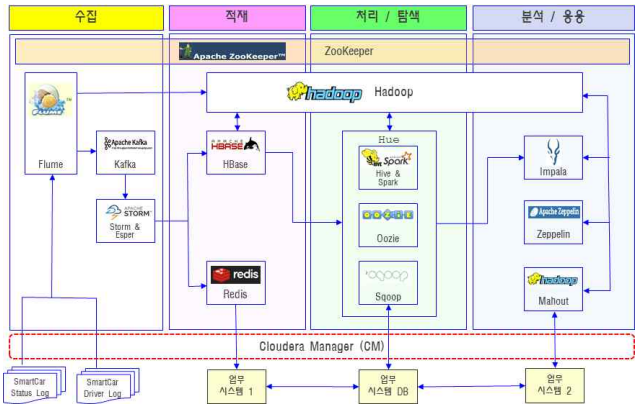
2.6 Zookeeper

Zookeeper는 분산 환경에서 하나의 서버에만 서비스가 집중되지 않도록 서버들 간의 조정으로 서비스를 분산 처리하게 해주며, 서버들 간의 환경 설정을 통합적으로 관리해 준다. 클라이언트가 ResionServer와 통신하려면 Zookeeper를 통해야 한다[13].

3. 개발 플랫폼의 설계

3.1 하둡 에코 시스템 구성도

빅데이터 수집 및 처리 구조는 [그림 1]과 같이 수집, 적재, 처리/탐색, 그리고 분석/응용 단계로 구성되어 있다.



[그림 1] 하둡 에코시스템 구성도

3.2 스프링 기반의 웹스크래핑 풀스택 개발 환경

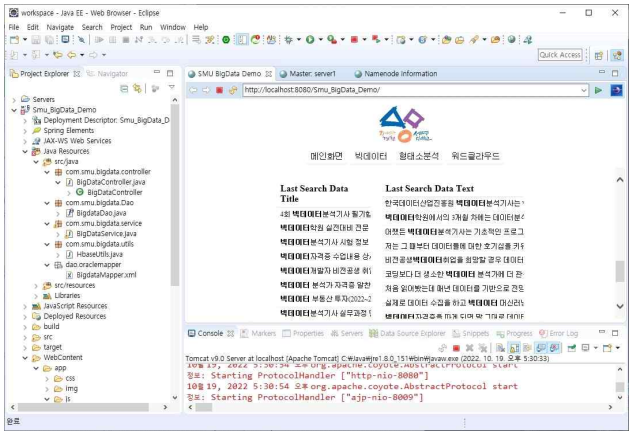
스프링 프레임워크 기반의 웹스크래핑 풀스택 개발 환경은 가상화 환경에서 리눅스 서버 3대로 구축되었다. 서버 1에서는 하둡 시스템과 HBase 및 주키퍼가 동작되도록 ResourceManager, NameNode, Secondary NameNode, HResionServer 그리고 QuoeumPeerMain 으로 구성하고, 서버 2, 서버 3에서는 NodeManager, DataNode, HResionServer 그리고 QuorumPeerMain 이 동작하도록 구성하였다. [표 1]은 본 연구에서 구성된 스프링 프레임워크 기반의 웹스크래핑 풀스택 개발 환경이다.

[표 1] 분산 빅데이터 구축 개발 환경

Items	Contents
Virtualization	Oracle VM VirtualBox 6.x
Server 1, 2, 3 O/S	Ubuntu Linux Ubuntu 16.04.3 LTS
IDE Tools	Eclipse IDE for Enterprise Java and Web Developers
Web Container	Apache Tomcat 9.0.13
Java Development Kit	Linux x64 Java 1.8.0_161
Framework	Spring Framework 4.x
Hadoop Ecosystem	Hadoop 2.6.x
	HBase 1.2.x
	Zookeeper 3.x
RDBMS	Oracle DataBase 11g 11.2.0

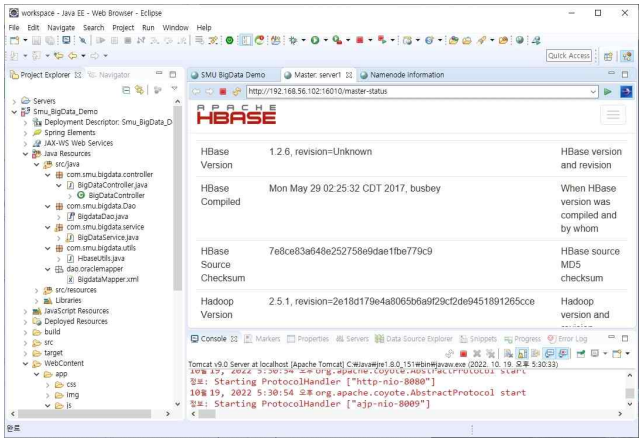
4. 웹스크래핑 빅데이터 풀스택 시스템의 구현

본 연구에서 구현한 웹스크래핑 빅데이터 풀스택 시스템을 구현한 화면은[그림 2]와 같다.



[그림 2] 웹스크래핑 빅데이터 풀스택 시스템 주 화면

웹스크래핑 한 후 데이터를 HBase에 적재하여 실행되고 있는 화면은 [그림 3]과 같다.



[그림 3] 웹스크래핑 후 데이터를 적재한 HBase 화면

5. 결 론

향후 다양한 디바이스의 출현으로 정형/비정형 데이터의 가파른 증가는 많은 영역에서 빅데이터를 새로운 활용에 따른 다양한 서비스를 요구하고 있다. 그러나 급속히 증가하는 빅데이터를 활용하고 적용하는 사례들은 많지만 대부분 실무 사례를 중심으로 소개하는 것이었다. 또한 스프링 프레임워크 환경에서 웹 스크래핑을 통한 데이터를 수집하고 적재한 후, 형태소 분석을 통하여 정재한 정형 데이터를 관계형 데이터베이스에 저장하고 처리하는 하둡 시스템 활용사례들은 많이 부족하였다.

따라서 본 연구에서는 스프링 프레임워크 환경에서 3대의 가상 머신 서버를 통하여 웹스크래핑으로 비정형 데이터를 수집한 후, 수집된 비정형 데이터를 형태소 분석기를 이용하여 정형화된 빅데이터를 관계형 데이터베이스에 저장할 수 있게 설계하고 구현하였다.

향후 Frontend를 React.js/Vue.js, Cross Platform 환경의 Fultter로 확장하여 웹이나 모바일로 실제 적용 사례를 통하여 서비스 될 수 있는 풀스택 인공지능, 머신 러닝, 딥러닝 연구가 지속되어야 할 것이다.

REFERENCES

[1] Apache Hadoop 2.10.1, https://hadoop.apache.org/docs/r2.10.1/
[2] Apache Hadoop 3.1.4, https://hadoop.apache.org/docs/r3.1.4/
[3] National Information Society Agency. eGovFrame, https://www.egovframe.go.kr/home/sub.do? menuNo=32
[4] H. J. Kim, "Design and Implementation of an Efficient Web Services Data Processing Using Hadoop-Based Big Data Processing", Journal of the Korea Academia-Industrial cooperation Society, 16(1), pp. 726-734, 2015.
[5] Wikipedia, Web scraping, https://en.wikipedia.org/wiki/Web_scraping
[6] Atlassian, Bitbucket, MeCab-ko, https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/
[7] Twitter, Open Korea Text(Okt), https://github.com/open-korean-text/open-korean-text
[8] Kakao Tech. Khaiii, https://tech.kakao.com/2018/12/13/khaiii/
[9] Wikipedia, Spring Framework. https://en.wikipedia.org/wiki/Spring_Framework
[10] I. M. Lee, Spring 3.1 of Toby(Vol. 1). Acorn, 2012.
[11] MyBatis, https://mybatis.org/mybatis-3/index.html
[12] http://www.edureka.in/blog/introduction-to-hadoop-2-0-and-advantages-of-hadoop-2-0/
[13] Apache Zookeeper, http://zookeeper.apache.org/