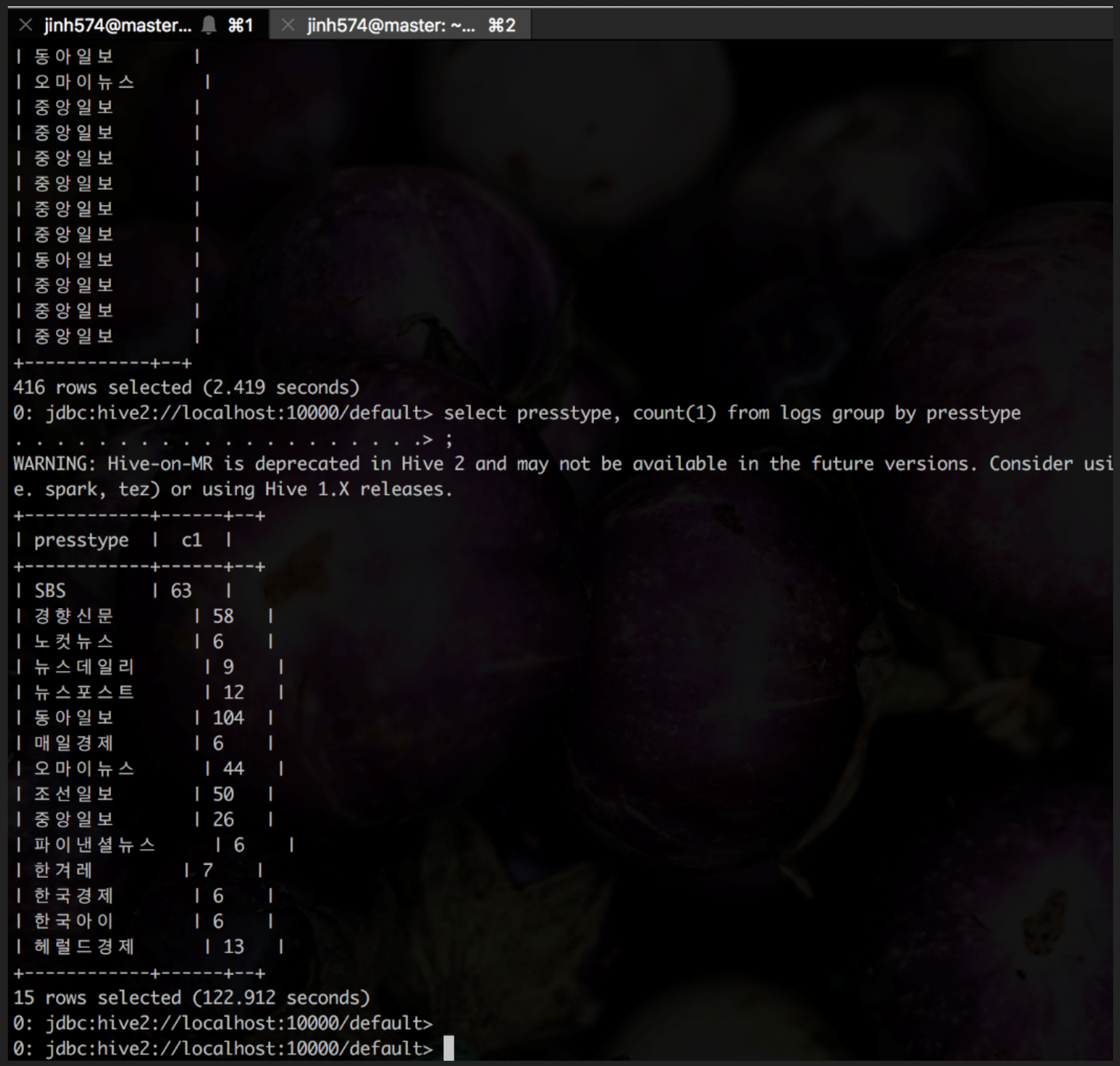
BIGDATA ANALYSIS WITH HADOOP

2017대선 BIGDATA 분석

INDEX

데이터 수집 예시

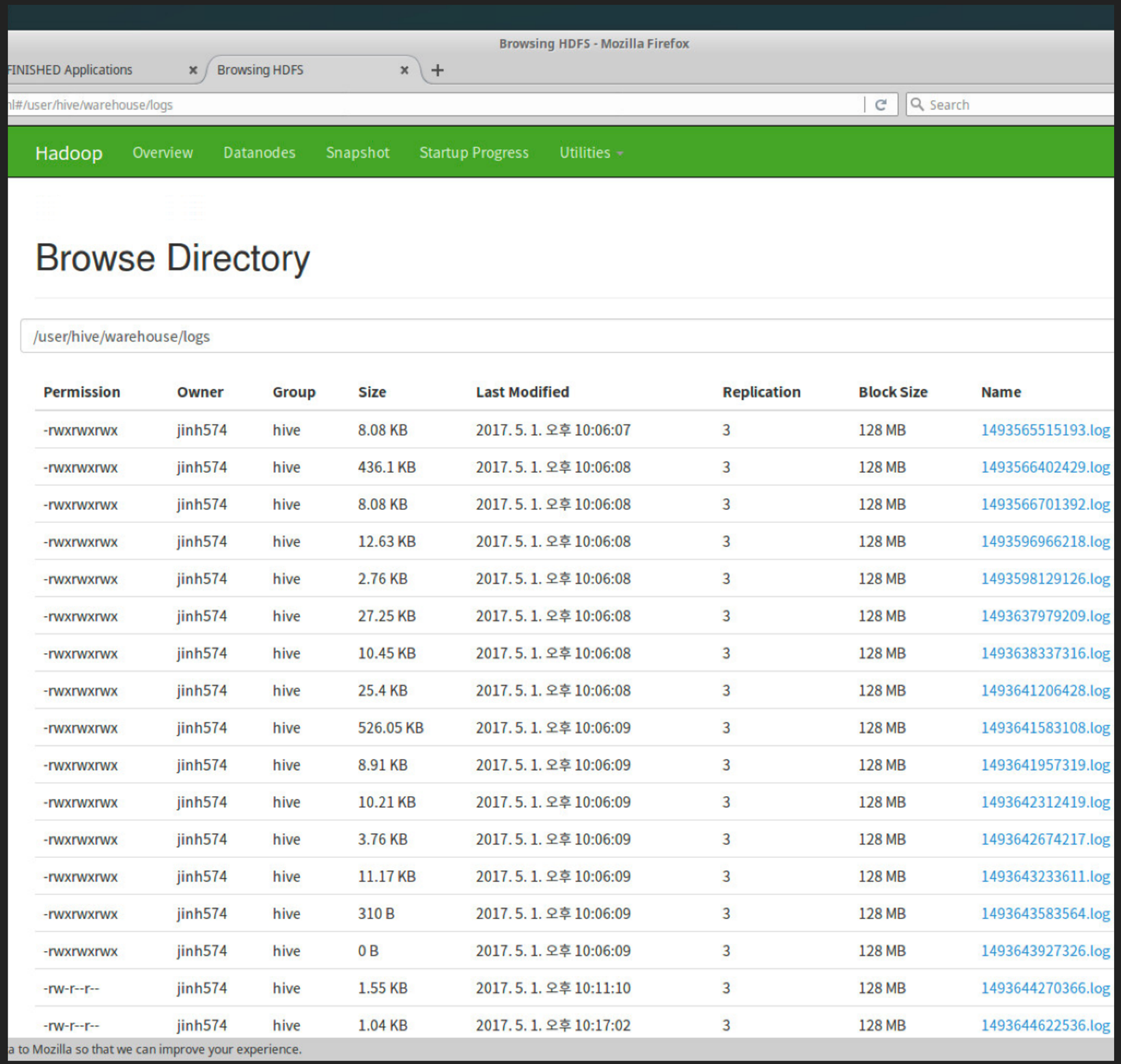
언론사 RSS 수집 초기
Hive를 통해 데이터 수 추출



HDFS BROWSER 예시

5분마다 수집된 RSS 피드를
HDFS에 저장 및 관리

추후, 맵리듀스잡으로
하루치 로그파일로 병합



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxrwxrwx	jinh574	hive	8.08 KB	2017. 5. 1. 오후 10:06:07	3	128 MB	1493565515193.log
-rwxrwxrwx	jinh574	hive	436.1 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493566402429.log
-rwxrwxrwx	jinh574	hive	8.08 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493566701392.log
-rwxrwxrwx	jinh574	hive	12.63 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493596966218.log
-rwxrwxrwx	jinh574	hive	2.76 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493598129126.log
-rwxrwxrwx	jinh574	hive	27.25 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493637979209.log
-rwxrwxrwx	jinh574	hive	10.45 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493638337316.log
-rwxrwxrwx	jinh574	hive	25.4 KB	2017. 5. 1. 오후 10:06:08	3	128 MB	1493641206428.log
-rwxrwxrwx	jinh574	hive	526.05 KB	2017. 5. 1. 오후 10:06:09	3	128 MB	1493641583108.log
-rwxrwxrwx	jinh574	hive	8.91 KB	2017. 5. 1. 오후 10:06:09	3	128 MB	1493641957319.log
-rwxrwxrwx	jinh574	hive	10.21 KB	2017. 5. 1. 오후 10:06:09	3	128 MB	1493642312419.log
-rwxrwxrwx	jinh574	hive	3.76 KB	2017. 5. 1. 오후 10:06:09	3	128 MB	1493642674217.log
-rwxrwxrwx	jinh574	hive	11.17 KB	2017. 5. 1. 오후 10:06:09	3	128 MB	1493643233611.log
-rwxrwxrwx	jinh574	hive	310 B	2017. 5. 1. 오후 10:06:09	3	128 MB	1493643583564.log
-rwxrwxrwx	jinh574	hive	0 B	2017. 5. 1. 오후 10:06:09	3	128 MB	1493643927326.log
-rw-r--r--	jinh574	hive	1.55 KB	2017. 5. 1. 오후 10:11:10	3	128 MB	1493644270366.log
-rw-r--r--	jinh574	hive	1.04 KB	2017. 5. 1. 오후 10:17:02	3	128 MB	1493644622536.log

HIVEQL 분석 예시

파티셔닝으로 물리적으로 분리된
테이블을 분석

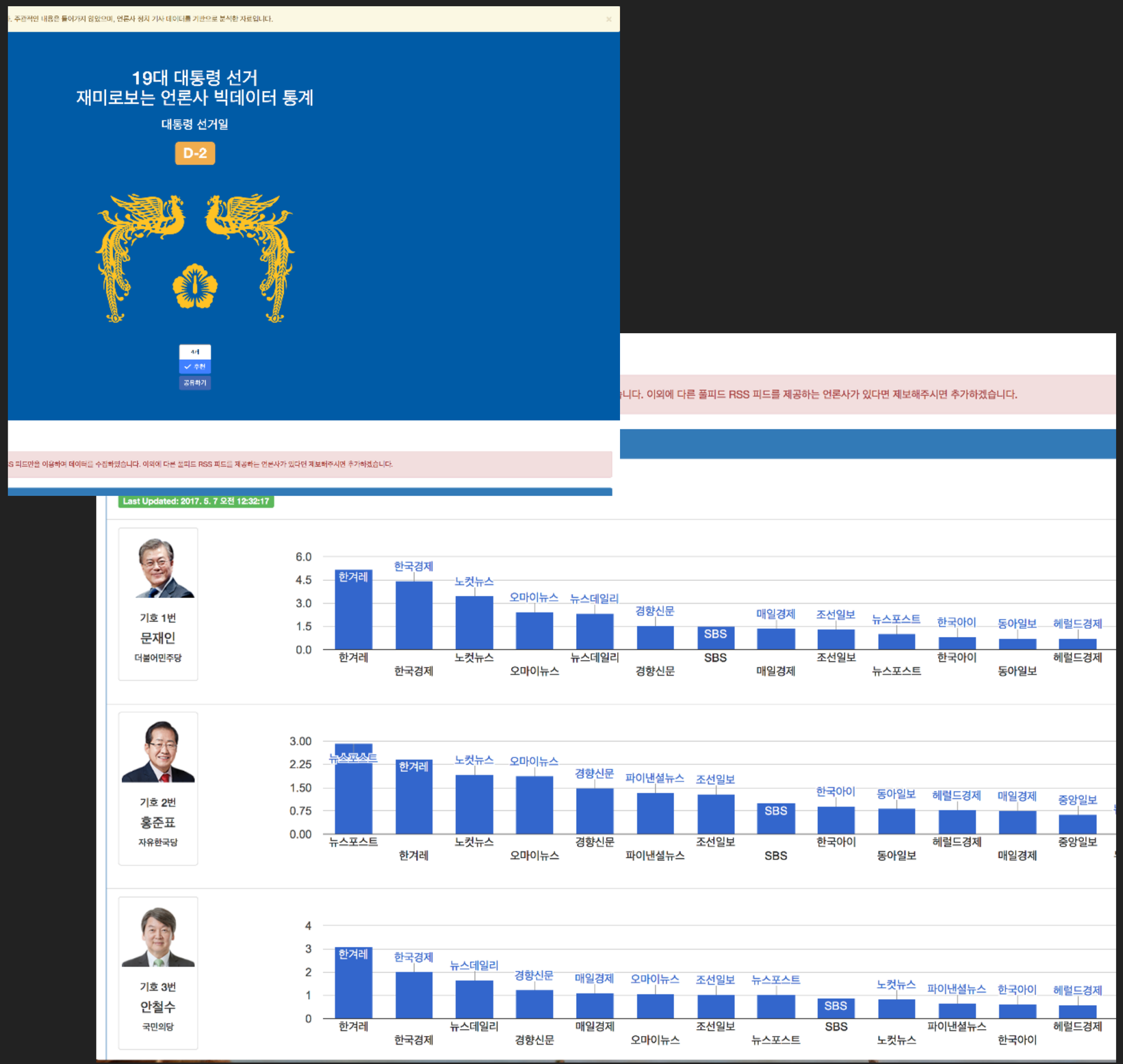
후보별 쿼리를 통해
언론사별 언급횟수 계산

```
INSERT OVERWRITE INTO TABLE candidates PARTITION (candidate = '1')
SELECT a.presstype, a.count AS word_count, b.count AS total_count, a.count/b.count AS avg_count FROM
    (SELECT presstype, count(1) AS count FROM
        (logs LATERAL VIEW explode(split(contents, " ")) exploded_table AS word)
    WHERE word LIKE '%문재인%' OR word LIKE '%문 후보%'
    GROUP BY presstype) a
JOIN
    (SELECT presstype, count(1) AS count FROM
        logs GROUP BY presstype) b
ON (a.presstype = b.presstype);
```

분석 데이터 서비스 예시

Spring프레임워크로
분석된 정보 가공

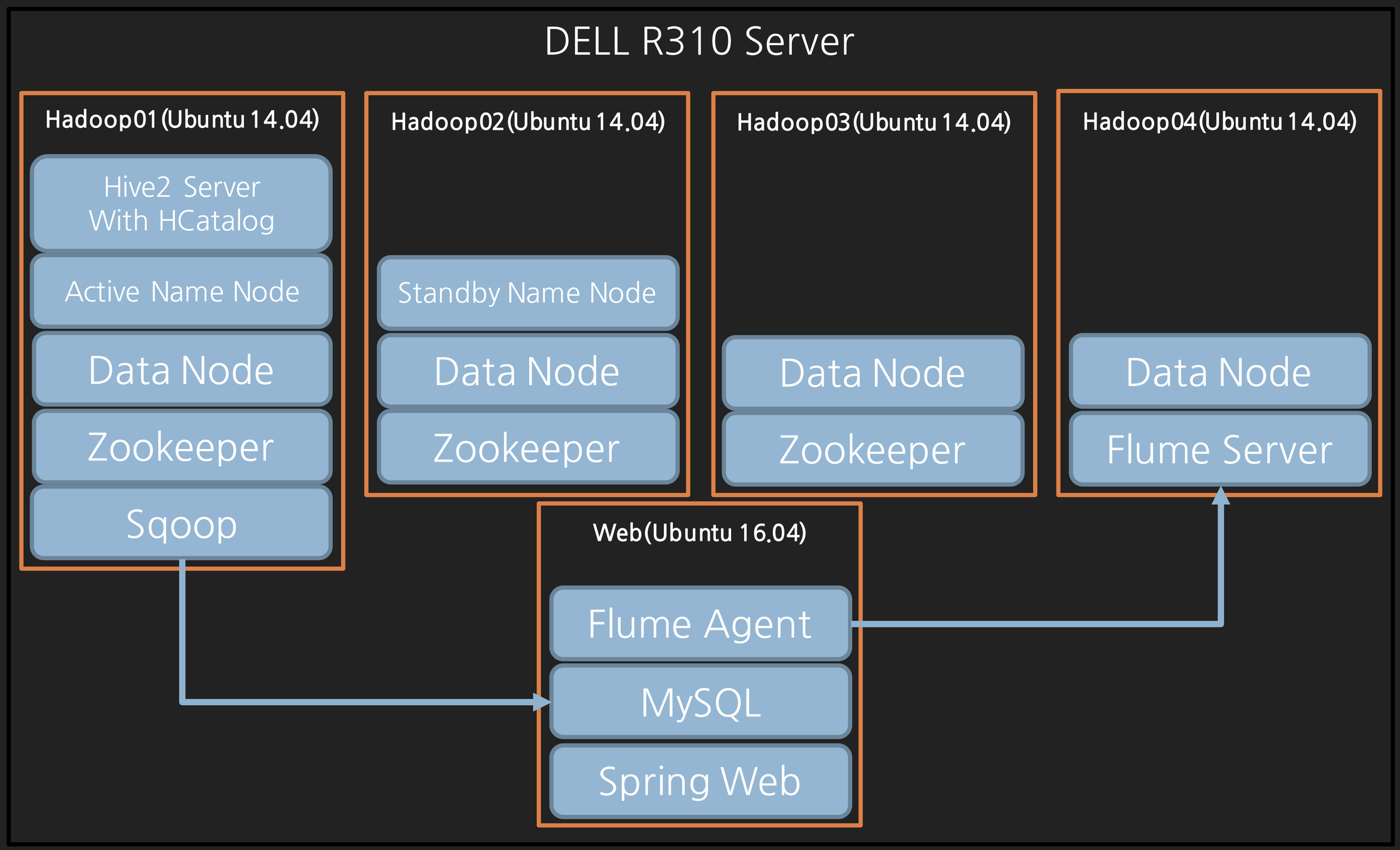
분석된 정보는 ‘Sqoop’을 통해
서비스 MySQL로 복사 후 서비스



분석 데이터 서비스 예시

분석된 정보는
후보별, 언론사별 기준을 나누어
시각화하여 제공

서비스 구성도


DELL R310 Server
Hadoop01(Ubuntu 14.04)
Hive2 Server With HCatalog
Active Name Node
Data Node
Zookeeper
Sqoop
Hadoop02(Ubuntu 14.04)
Standby Name Node
Data Node
Zookeeper
Hadoop03(Ubuntu 14.04)
Data Node
Zookeeper
Hadoop04(Ubuntu 14.04)
Data Node
Flume Server
Web(Ubuntu 16.04)
Flume Agent
MySQL
Spring Web

개요

대한민국 헌정사상 처음으로 대통령이 탄핵되는 사태를 맞이하여 19대 대통령 선거를 하게 되었습니다. 이에 기사에서 언급되는 대통령 후보자에 따른 후보 지지율 예측과 언로사별 후보자 지지에 대한 정보를 알고 싶었습니다.

본 프로젝트는 하둡 에코시스템 생태계에 대한 전반적인 이해와 기반이 되는 HDFS와 맵리듀스에 대한 이해에 중점을 두었습니다. 빅데이터 과학자로서 필요로 하는 역량을 키우기 위해 데이터 수집부터 시각화까지 전체적인 수행절차를 프로젝트를 통해 수행한 프로젝트입니다.

기능

- Zookeeper를 통해 하둡 네임노드에 HA를 적용하여 안정적인 서비스 구축
- 5분마다 RSS 주소에 접근하여 새로운 피드의 제목, 내용, 링크, 생성날짜 정보를 수집
- 수집된 RSS 피드를 Flume을 통해 HDFS에 저장
  - Hive의 맵리듀스 작업의 효율을 위해 5분마다 수집된 데이터를 하루단위로 병합하여 생성 관리
- 1시간마다 수집된 데이터를 Hive를 통해 분석하고 외부테이블에 생성 저장
- 분석된 외부테이블 데이터를 Sqoop을 통해 서비스용 MySQL로 복사
- Spring웹프레임워크를 이용, 분석된 데이터를 시각화하여 서비스

개발 환경 및 프로젝트 기간

개발환경

- Ubuntu 14.04(하둡노드) * 4, Ubuntu 16.04(Web)
- Hadoop 2.7.2
- Hive 2.1.1
- Zookeeper 3.4.10
- Flume 1.7
- Sqoop 1.4.6
- MySQL 5.6
- Spring Boot 1.3
  - Spring Web, Spring Security, Spring JDBC

프로젝트 기간

- 2017년 04월 ~ 2017년 05월 09일까지

구현방법

- 5분마다 Java ROME Library를 이용하여 RSS 수집 후 텍스트파일 저장
- Hadoop HDFS 구축, Zookeeper를 통해 네임노드 HA 구성
- 손쉬운 맵리듀스 잡을 실행하기 위해 Hive 구성
  - 단, 파티셔닝을 통해 물리적으로 후보별 데이터를 분리 저장
  - HCatalog를 이용하여 파티셔닝된 데이터를 추상화하여 Sqoop에 이용
- 수집된 RSS를 Flume을 통해 서비스용 MySQL에 복사
- Crontab을 이용 주기적으로 Hive분석, 파일병합, Sqoop전송을 관리
- Spring웹프레임워크 이용하여 RESTful 서비스 제공

GIT ADDRESS

HTTPS://GITHUB.COM/JINH574/JAVA-COLLECTRSSDATA