

빅데이터 처리시스템 개발

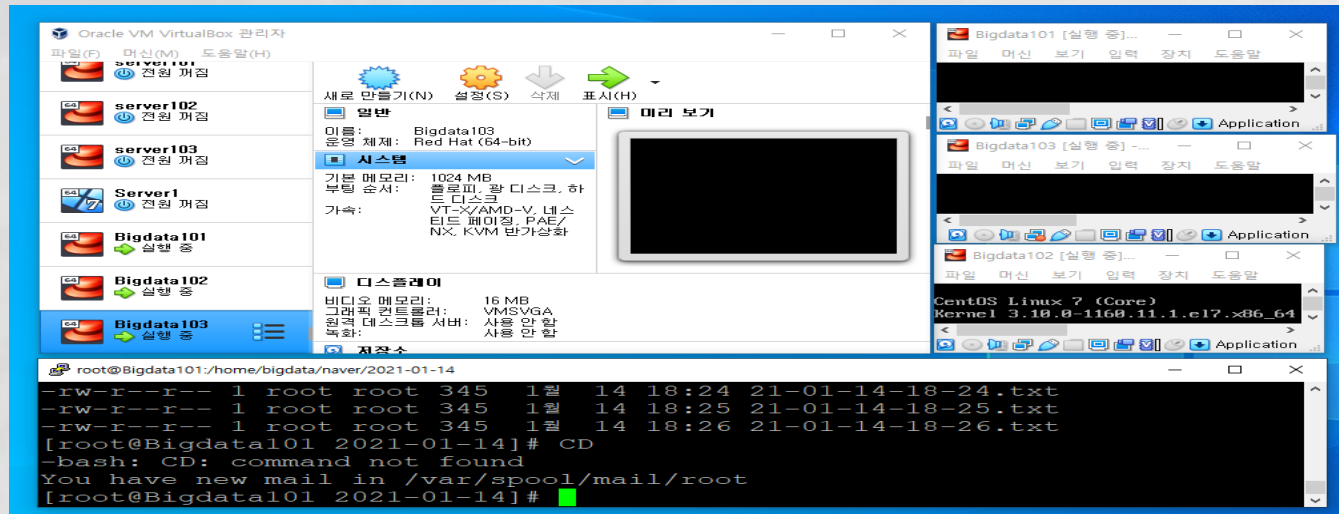
김은표

목차

- Hadoopo 실행
- MapReduce 실습
- Hive 실습

Hadoopo 실행

- 가상PC 3개 작동 후 PuTTY를 이용하여 메인 드라이버 접속
- <http://192.168.56.101:50070/> 접속 확인
- [Live Nodes](#) 확인하기






Configured Capacity:	18.56 GB
DFS Used:	8.24 MB (0.04%)
Non DFS Used:	12.2 GB
DFS Remaining:	6.35 GB (34.2%)
Block Pool Used:	8.24 MB (0.04%)
DataNodes usages% (Min/Median/Max/stdDev):	0.04% / 0.04% / 0.05% / 0.00%
<u>Live Nodes</u>	3 (Decommissioned: 0, In Maintenance: 0)

MapReduce 실습


● MapReduce 실행

1. start-all.sh
2. hdfs dfs -mkdir /MapReduce
3. hdfs dfs -put /root/2_7_naver.py /MapReduce

Browse Directory

Show entries

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	1.52 KB	Jan 14 15:36	3	128 MB	2_7_naver.py	




Showing 1 to 1 of 1 entries

Hadoop, 2020.



Hive 실습

- Hive용 HDFS 디렉터리 생성
 - `hdfs dfs -mkdir /hive`
 - `hdfs dfs -mkdir /hive/warehouse`

Browse Directory

Show entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Jan 13 15:47	0	0 B	naver_in	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Jan 13 15:32	0	0 B	user1	

Showing 1 to 2 of 2 entries

Hive 실습

○ Hive 설치

- `cd /home/bigdata/`
- `ln -s apache-hive-2.3.7-bin/ hive`

○ Hive 설정 후 실행

```
[root@Bigdata101 bigdata]# ll
합 계 20
-rw-r--r-- 1 root root 14817 1월 8 12:10 User1.java
drwxr-xr-x 8 root root 200 1월 7 16:39 apache-flume-1.9.0-bin
drwxr-xr-x 10 root root 184 1월 12 17:01 apache-hive-2.3.7-bin
lrwxrwxrwx 1 root root 22 1월 7 15:17 flume -> apache-flume-1
.9.0-bin
lrwxrwxrwx 1 root root 14 1월 5 17:30 hadoop -> hadoop-2.10.1

Logging initialized using configuration in file:/home/bigdata/apache-h
2.3.7-bin/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the fut
versions. Consider using a different execution engine (i.e. spark, tez
using Hive 1.X releases.
hive>
```

Hive 실습

- Hive 실행 후 Hive 기본 내부테이블 `Naver_in` 생성

```
hive> SHOW TABLES;
OK
naver_in
user1
user2
Time taken: 0.037 seconds, Fetched: 3 row(s)
hive> CREATE TABLE `Naver_in` (
  > `rank` Int,
  > `keyword` String,
  > `rdate` String
  > )
  > row format delimited
  > fields terminated by ","
  > tblproperties("skip.header.line.count"="1");
(CALLED: Execution Error: return code 1 from org.apache
```

Hive 실습

- Naver_in 테이블을 OVERWRITE 후 SELECT * FROM `Naver_in`을 실행 시켜 결과를 출력시킨다.

```
hive> LOAD DATA INPATH '/naver/2021-01-04/*'
Loading data to table default.naver_in
OK
Time taken: 1.675 seconds
hive> SELECT * FROM `Naver_in`;
OK
1      정 인 이 사 건      210104174154
2      박 봄              210104174154
3      윤 갑 근          210104174154
4      add              210104174154
5      현 대 로 템        210104174154
6      아 키 네 이 터      210104174154
7      박 봄 키          210104174154
8      나 경 원          210104174154
9      정 인 이 진 정 서    210104174154
10     구 혜 선          210104174154
```


Hive 실습

- Hive 외부 테이블 생성하기 위해 Naver_ex 테이블 생성
- 생성된 외부 테이블을 SELECT * FROM Naver_ex 실행

```
hive>  
>  
> CREATE EXTERNAL TABLE `Naver_ex` (  
> `rank` Int,  
> `keyword` String,  
> `rdate` String  
> )  
> row format delimited  
> fields terminated by ","  
> location "/naver/2021-01-04/"  
> tblproperties("skip.header.line.count"="1");  
OK  
Time taken: 0.127 seconds  
hive> SELECT * FROM `Naver_ex`;  
OK  
Time taken: 0.209 seconds  
hive> 
```

Hive 실습

- 내부 테이블 `Naver_in`의 키워드 Total 출력하는 코드와 결과입니다.

```
hive> SELECT keyword, SUM(1) as total FROM Naver_in  
> GROUP BY keyword ORDER BY total DESC;
```

박 봄	39
나 경 원	39
정 인 이 사 건	39
윤 갑 근	39
아 키 네 이 터	39
박 봄 키	39
현 대 로 템	34
구 혜 선	27
나 경 원 딸	22
양 치 승	17
심 형 래	15
현 대 위 아	14
add	13
큰 스 탄 틴	7
정 인 이 진 정 서	3
마 원	2
송 파 구 의 원	2

Hive 실습

- 외부 테이블 `Naver_in` 의 키워드 Total 출력하는 코드와 결과입니다.
- 정상적으로 들어갔는지 경로 확인(192.168.56.101:8088)

```
hive> SELECT keyword, SUM(1) as total FROM Naver_ex  
> GROUP BY keyword ORDER BY total DESC;
```

```
Stage-Stage-1: Reduce: 1 Cumul  
HDFS Write: 96 SUCCESS  
Stage-Stage-2: Map: 1 Reduce: 1  
ad: 5584 HDFS Write: 87 SUCCESS  
Total MapReduce CPU Time Spent: 1  
OK  
Time taken: 38.668 seconds  
hive> exit
```

application 1610679770562 0004	root	SELECT keyword, SUM(1) as total FROM ...DESC(Stage- 2)	MAPREDUCE	default	0	Fri Jan 15 12:52:29 +0900 2021
application 1610679770562 0003	root	SELECT keyword, SUM(1) as total FROM ...DESC(Stage- 1)	MAPREDUCE	default	0	Fri Jan 15 12:52:15 +0900 2021