HDFS Exercises

HDFS 실습을 위한 입력 데이터 셋

Grouplens 에서 제공되는 **movie_lens dataset**

영화 정보 및 평점 관련 데이터 셋 제공

20,000,263 개의 평점

465,564 개의 영화 태그 및 **27,278** 개의 영화 정보 보유

https://grouplens.org/datasets/movielens/

Structure of MovieLens Data

File Name

Structure

movies.csv

```
kmubigdatalab@kmubigdata-cluster-m:~/ml-20m$ head movies.csv
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
7,Sabrina (1995),Comedy|Romance
8,Tom and Huck (1995),Adventure|Children
9,Sudden Death (1995),Action
10,GoldenEye (1995),Action|Adventure|Thriller
```

movieId, title, genres

tags.csv

```
kmubigdatalab@kmubigdata-cluster-m:~/ml-20m$ head tags.csv
18,4141,Mark Waters,1240597180
65,208,dark hero,1368150078
65,353,dark hero,1368150079
65,521,noir thriller,1368149983
65,592,dark hero,1368150078
65,668,bollywood,1368149876
65,898,screwball comedy,1368150160
65,1248,noir thriller,1368149983
65,1391,mars,1368150055
65,1617,neo-noir,1368150217
```

userId, movieId, tag, timestamp

ratings.csv

```
kmubigdatalab@kmubigdata-cluster-m:~/ml-20m$ head ratings.csv
1,2,3.5,1112486027
1,29,3.5,1112484676
1,32,3.5,1112484819
1,47,3.5,1112484727
1,50,3.5,1112484580
1,112,3.5,1094785740
1,151,4.0,1094785734
1,223,4.0,1112485573
1,253,4.0,1112484940
1,260,4.0,1112484826
```

userId, movieId, rating, timestamp

bigDATA Lab.
KOOKMIN UNIVERSITY

Downloading MovieLens Data

$ wget https://s3.ap-northeast-2.amazonaws.com/kmubigdata-movielensdata/ml-20m.zip

$ unzip ml-20m.zip

$ rm -f ml-20m.zip

$ head ml-20m/movies.csv

```
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
7,Sabrina (1995),Comedy|Romance
8,Tom and Huck (1995),Adventure|Children
9,Sudden Death (1995),Action
10,GoldenEye (1995),Action|Adventure|Thriller
```

Handful HDFS commands

HDFS 의 명령어

포맷

$ hdfs [SHELL_OPTIONS] COMMAND

Usage

$ hdfs

// shows the usage

Handful HDFS commands

HDFS 관련해서는 대부분의 명령어는 hdfs dfs [COMMAND [COMMAND_OPTIONS]] 형식

Hdfs 의 파일 시스템과 연관된 옵션들을 보여줌

```
$ hdfs dfs
// shows generic options
```

Handful HDFS commands

Command

```
$ hdfs dfs -ls <args>
```

Example

```
$ hdfs dfs -ls /
```

Handful HDFS commands

Command

$ hdfs dfs -mkdir [-p] <paths>

입력으로 들어온 이름의 디렉토리를 생성

Example

$ hdfs dfs -mkdir -p /dataset/movielens/

$ hdfs dfs -ls /dataset/

Handful HDFS commands

Command

$ hdfs dfs -put <localsrc> <dst>

Copy srcs from local file system to hdfs.

Example

$ hdfs dfs -put ml-20m/ratings.csv /dataset/movielens/

$ hdfs dfs -ls /dataset/movielens/

Handful HDFS commands

Command

$ hdfs fsck <path>

파일시스템의 상태를 확인가능하게 해줌

Example

$ hdfs fsck /dataset/movielens/ratings.csv

ratings.csv 파일의 전체 파일 크기는 얼마인가? 평균 블락 크기는 무엇인가? 블락의 갯수는 몇개이며, **HDFS** 에서의 기본 블락 크기는 얼마로 설정되어 있을까요?

HDFS 에서 블락 크기 설정

명령어를 통해서 블락 사이즈의 변경 가능

$ hdfs dfs -Ddfs.block.size=1048576 -put ml-20m/ratings.csv /

위 명령어는 블락사이즈를 **1MB**로 설정을 하게되고, **rating.csv** 파일의 블락 갯수는 더 많이 증가하게 된다. **Fsck** 명령어를 통해서 나온 결과물을 확인해보세요.

$ hdfs fsck /ratings.csv

Handful HDFS commands

Command

$ hdfs dfs -get <localsrc> <dst>

HDFS 에 있는 파일을 로컬 디스크로 복사

Example

$ hdfs dfs -get /dataset/movielens/ratings.csv /tmp/

$ ls /tmp/ratings.csv

Handful HDFS commands

Command

$ hdfs dfs -cat URI

Stdout 에 파일내용을 출력

Example

$ hdfs dfs -cat /dataset/movielens/ratings.csv | less // type “q” to quit

Handful HDFS commands

Command

$ hdfs dfs -tail URI

파일의 끝부분 출력

Example

$ hdfs dfs -put ml-20m/movies.csv /dataset/movielens/

$ hdfs dfs -tail /dataset/movielens/movies.csv

Handful HDFS commands

Command

$ hdfs dfs -df [-h] URI

HDFS 에서 디스크 가용 용량을 보여줌

Example

$ hdfs dfs -df -h /

Handful HDFS commands

Command

$ hdfs dfs -cp <source> <dest>

HDFS 내부에서 source 경로에서 dest 경로로 파일을 복제 함

Example

$ hdfs dfs -cp /dataset/movielens/*.csv /

$ hdfs dfs -ls /

Handful HDFS commands

Command

$ hdfs dfs -getfacl <path>

파일과 디렉토리의 Access Control List (ACL) 를 출력해줌

Example

$ hdfs dfs -getfacl /movies.csv

Handful HDFS commands

HDFS 의 파일 replica 갯수를 바꿀 수 있음

Command - $ hdfs dfs -setrep <numReplicas> <path>

numRplicas 만큼으로 path 의 파일을 설정

Example

$ hdfs dfs -put ml-20m/tags.csv /dataset/movielens/

$ hdfs fsck /dataset/movielens/tags.csv # 이명령어를 통해서 replica 갯수 확인

$ /dataset/movielens/tags.csv #파일의 replica 갯수를 바꾸는 명령어를 실행 (원본 파일의 현재 replica 갯수와 다르게 임의로 설정. 만약 replica 갯수를 3으로 설정한다면 어떻게 될까요? 그 이유를 서술해주세요.

$ hdfs fsck /dataset/movielens/tags.csv

Handful HDFS commands

Command

$ hdfs dfs -rm [-f] [-r|-R] URI

주어진 파일을 지움

Example

$ hdfs dfs -rm /ratings.csv /movies.csv

Handful HDFS commands

Command

$ hdfs dfsadmin -report

HDFS 시스템 상태를 알려주는 명령어

Example

$ hdfs dfsadmin -report

결과물을 캡쳐하고 출력되는 내용이 어떤 것을 의미하는지 보고서에 작성해주세요.