Hive를 사용한 movielen 분석

(이삭엔지니어링 인턴 김형근, fnfn9087@gmail.com)

(CDH 5.15.1에서 진행)

1. MovieLens Dataset 가져오기

```
$ wget http://files.grouplens.org/datasets/movielens/ml-1m.zip
$ unzip ml-m1.zip
$ cd ml-m1
```

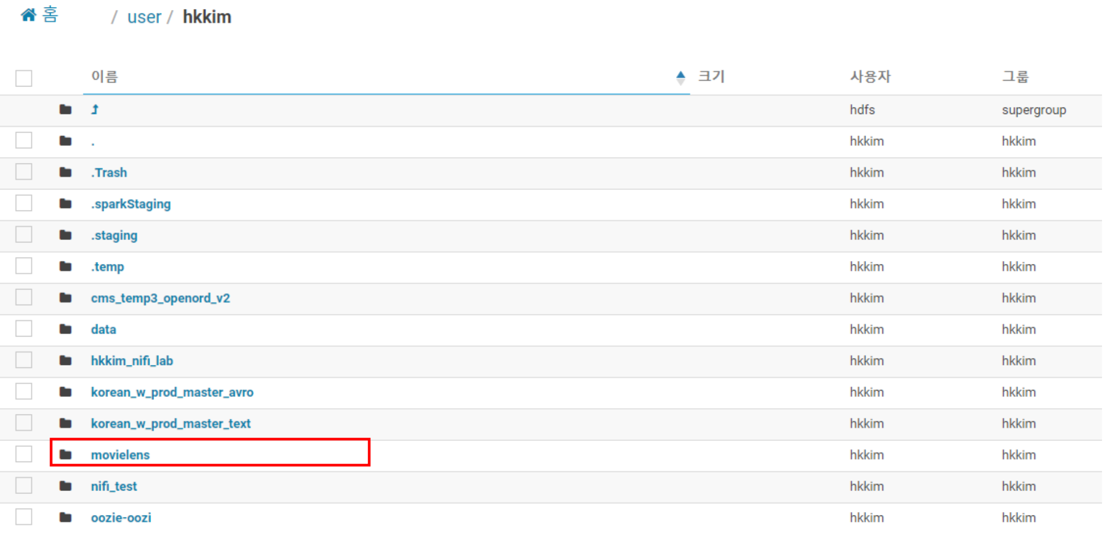
// ml-m1 디렉터리에는 movies.dat, users.dat, users.dat 라는 파일들이 존재할 것이다

```
$ sed -i 's/::/,/g' ml-1m/movies.dat
$ sed -i 's/::/,/g' ml-1m/users.dat
$ sed -i 's/::/,/g' ml-1m/ratings.dat
// 파일 내부의 구분자를 ":"에서 ","로 변경시킨다.

$ mv ml-1m/movies.dat /ml-1m/movies.csv
$ mv ml-1m/ratings.dat /ml-1m/ratings.csv
$ mv ml-1m/users.dat /ml-1m/users.csv
// 각 파일들의 파일형식을 csv로 바꾼다.
```

2. movielens 디렉터리 만들기

```
$ hdfs dfs -mkdir /user/hkkim/movielens
$ hdfs dfs -ls /user/hkkim
```



홈 / user / hkkim

		이름	크기	사용자	그룹
		↥		hdfs	supergroup
□		.		hkkim	hkkim
□		.Trash		hkkim	hkkim
□		.sparkStaging		hkkim	hkkim
□		.staging		hkkim	hkkim
□		.temp		hkkim	hkkim
□		cms_temp3_openord_v2		hkkim	hkkim
□		data		hkkim	hkkim
□		hkkim_nifi_lab		hkkim	hkkim
□		korean_w_prod_master_avro		hkkim	hkkim
□		korean_w_prod_master_text		hkkim	hkkim
□		movielens		hkkim	hkkim
□		nifi_test		hkkim	hkkim
□		oozie-oozi		hkkim	hkkim

3. SQL문 작성

(1) movies.sql

```
DROP DATABASE IF EXISTS movielens CASCADE;
CREATE DATABASE movielens;
USE movielens;
CREATE EXTERNAL TABLE movies (MovieID INT,
Title varchar(60),
Genres varchar(60))
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY "\n"
STORED AS TEXTFILE
LOCATION '/user/hkkim/movielens/ml-1m/mvs.txt';
LOAD DATA INPATH '/user/hkkim/movielens/ml-1m/movies.csv' INTO TABLE
movies;
SELECT * FROM movies LIMIT 10;
```

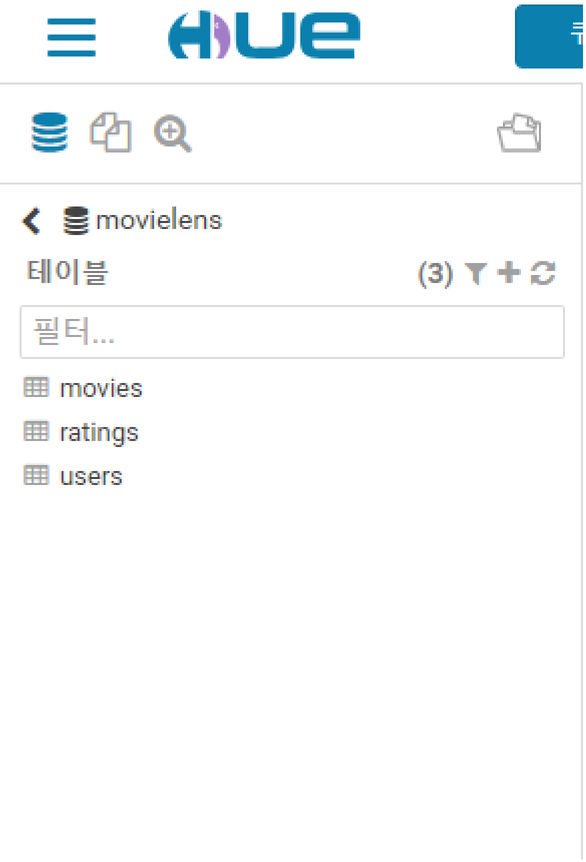
(2) ratings.sql

```
USE movielens;
CREATE EXTERNAL TABLE ratings (UserID INT,
MovieID INT,
Rating INT,
Timestamp STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY "\n"
STORED AS TEXTFILE
LOCATION '/user/hkkim/movielens/ml-1m/rts.txt';
LOAD DATA INPATH '/user/hkkim/movielens/ml-1m/ratings.csv' INTO
TABLE ratings;
SELECT * FROM ratings LIMIT 10;
```

(3) users.sql

```
USE movielens;
CREATE EXTERNAL TABLE users (UserID INT,
Gender STRING,
Age INT,
Occupation INT,
ZIP INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY "\n"
STORED AS TEXTFILE
LOCATION '/user/hkkim/movielens/ml-1m/usr.txt';
LOAD DATA INPATH '/user/hkkim/movielens/ml-1m/users.csv' INTO TABLE
users;
SELECT * FROM users LIMIT 10;
```

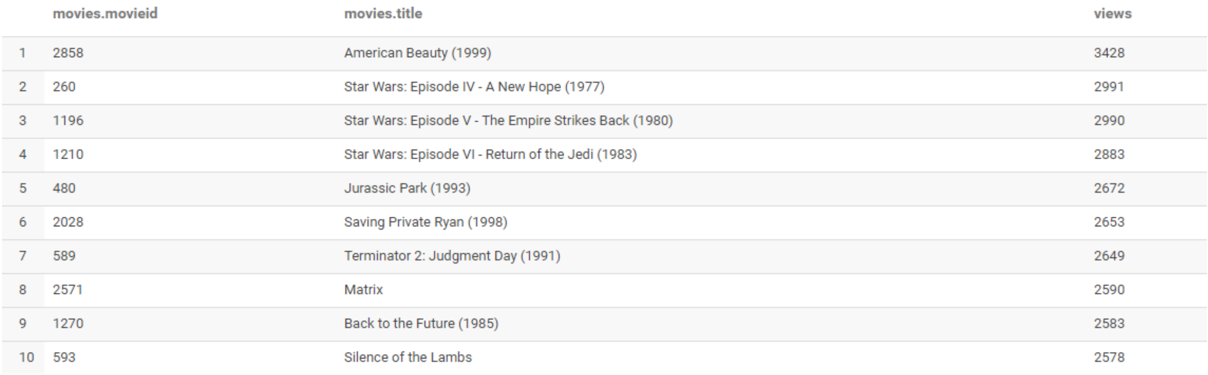
```
$ hive -f movies.sql
$ hive -f ratings.sql
$ hive -f users.sql
// 위와 같이 작성 HiveQL문을 실행시킨다.
```



4. 영화 순위 분석

(1) Top 10 영화

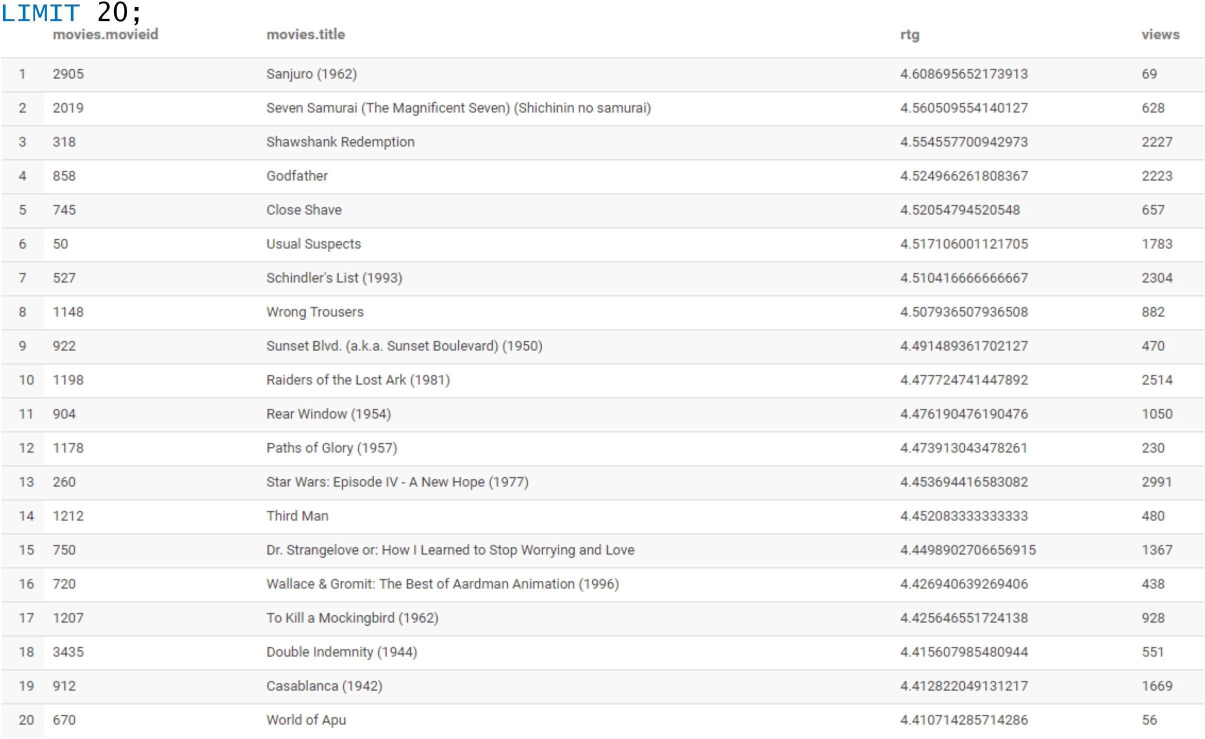
```
SELECT movies.MovieID,movies.Title,COUNT(DISTINCT ratings.UserID) as
views
FROM movies JOIN ratings ON (movies.MovieID = ratings.MovieID)
GROUP BY movies.MovieID, movies.Title
ORDER BY views DESC
LIMIT 10;
```



	movies.movieid	movies.title	views
1	2858	American Beauty (1999)	3428
2	260	Star Wars: Episode IV - A New Hope (1977)	2991
3	1196	Star Wars: Episode V - The Empire Strikes Back (1980)	2990
4	1210	Star Wars: Episode VI - Return of the Jedi (1983)	2883
5	480	Jurassic Park (1993)	2672
6	2028	Saving Private Ryan (1998)	2653
7	589	Terminator 2: Judgment Day (1991)	2649
8	2571	Matrix	2590
9	1270	Back to the Future (1985)	2583
10	593	Silence of the Lambs	2578

(2) 관람횟수가 40번 이상인 Top 20 영화

```
SELECT movies.MovieID,movies.Title,AVG(ratings.Rating) as
rtg,COUNT(DISTINCT ratings.UserID) as views
FROM movies JOIN ratings ON (movies.MovieID = ratings.MovieID)
GROUP BY movies.MovieID,movies.Title
HAVING views >= 40
ORDER BY rtg DESC
LIMIT 20;
```



	movies.movieid	movies.title	rtg	views
1	2905	Sanjuro (1962)	4.608695652173913	69
2	2019	Seven Samurai (The Magnificent Seven) (Shichinin no samurai)	4.560509554140127	628
3	318	Shawshank Redemption	4.554557700942973	2227
4	858	Godfather	4.524966261808367	2223
5	745	Close Shave	4.52054794520548	657
6	50	Usual Suspects	4.517106001121705	1783
7	527	Schindler's List (1993)	4.510416666666667	2304
8	1148	Wrong Trousers	4.507936507936508	882
9	922	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)	4.491489361702127	470
10	1198	Raiders of the Lost Ark (1981)	4.477724741447892	2514
11	904	Rear Window (1954)	4.476190476190476	1050
12	1178	Paths of Glory (1957)	4.473913043478261	230
13	260	Star Wars: Episode IV - A New Hope (1977)	4.453694416583082	2991
14	1212	Third Man	4.452083333333333	480
15	750	Dr. Strangelove or: How I Learned to Stop Worrying and Love	4.4498902706656915	1367
16	720	Wallace & Gromit: The Best of Aardman Animation (1996)	4.426940639269406	438
17	1207	To Kill a Mockingbird (1962)	4.425646551724138	928
18	3435	Double Indemnity (1944)	4.415607985480944	551
19	912	Casablanca (1942)	4.412822049131217	1669
20	670	World of Apu	4.410714285714286	56

Reference : https://towardsdatascience.com/getting-started-with-hive-ad8a93862f1a