Topgear - Hadoop hive Movielens

Part 01 (Find the distinct movies which are associated with the Tag "fairy tale")

1. Downloading files in hadoop master

```
$wget http://files.grouplens.org/datasets/movielens/ml-10m.zip
```

2. unzip file

```
$unzip ml-10m.zip
```

3. remove zip file

```
$rm ml-10m.zip
```

4. copy the files into HDFS

```
$hadoop fs -copyFromLocal ml-10m
```

5. hive query to create database

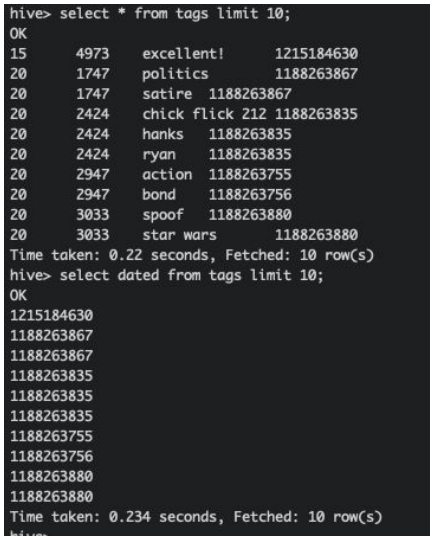
```
hive>create EXTERNAL table IF NOT EXISTS tags(userid int,movieid
int,tag string,dated bigint)
>ROW FORMAT SERDE
"org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe"
>WITH SERDEPROPERTIES ("field.delim"="::")
>STORED AS TEXTFILE;
```

```
hive> create EXTERNAL table IF NOT EXISTS tags(userid int,movieid int,tag string,dated bigint)
    > row format serde
    > 'org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe'
    > WITH serdeproperties ("field.delim"="::")
    > STORED as TEXTFILE;
OK
Time taken: 1.083 seconds
```

6.Load data from HDFS.

```
hive>Load data INPATH '/user/maria_dev/ml-10m/tags.dat'
OVERWRITE into TABLE tags;
```

```
hive> Load data INPATH '/user/maria_dev/ml-10m/tags.dat'
    > OVERWRITE into TABLE tags;
Loading data to table default.tags
Table default.tags stats: [numFiles=1, numRows=0, totalSize=3584119, rawDataSize=0]
OK
Time taken: 1.342 seconds
```



```
hive> select * from tags limit 10;
OK
15      4973    excellent!      1215184630
20      1747    politics        1188263867
20      1747    satire  1188263867
20      2424    chick flick 212 1188263835
20      2424    hanks   1188263835
20      2424    ryan    1188263835
20      2947    action  1188263755
20      2947    bond    1188263756
20      3033    spoof   1188263880
20      3033    star wars       1188263880
Time taken: 0.22 seconds, Fetched: 10 row(s)
hive> select dated from tags limit 10;
OK
1215184630
1188263867
1188263867
1188263835
1188263835
1188263835
1188263755
1188263756
1188263880
1188263880
Time taken: 0.234 seconds, Fetched: 10 row(s)
```

Adding hive jar for MultiDelimitSerde :

```
hive>add jar hive-contrib-2.3.3.jar
```

Creating movienames table and loading data into movies table :

```
hive>create EXTERNAL table IF NOT EXISTS movienames(movieid
int,moviename string,movie_catgry string)
> ROW FORMAT SERDE
"org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe"
> WITH SERDEPROPERTIES ("field.delim"="::")
> STORED AS TEXTFILE;

hive> Load data local INPATH 'ml-10m/movies.dat' OVERWRITE into TABLE
movienames;
Loading data to table default.movienames
```

```
hive> create EXTERNAL table IF NOT EXISTS movienames(movieid int,moviename string,movie_catgry string)
    > ROW FORMAT SERDE "org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe"
    > WITH SERDEPROPERTIES ("field.delim"="::")
    > STORED AS TEXTFILE;
OK
Time taken: 2.727 seconds
hive> describe movienames;
OK
movieid                 int                     from deserializer
moviename               string                  from deserializer
movie_catgry            string                  from deserializer
Time taken: 0.125 seconds, Fetched: 3 row(s)
hive> Load data local INPATH 'ml-10m/movies.dat' OVERWRITE into TABLE movienames;
Loading data to table default.movienames
OK
Time taken: 0.658 seconds
hive> select * from movienames limit 10;
OK
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
Time taken: 1.785 seconds, Fetched: 10 row(s)
```

Selecting distinct movies with movieid and movie name with tag fairy tales.

```
hive> select movieid,moviename from movienames where movieid in
(select distinct movieid from movietags where tag="fairy tale");
```

Result

```
531  Secret Garden, The (1993)
588  Aladdin (1992)
837  Matilda (1996)
1022 Cinderella (1950)
2143 Legend (1985)
4973 Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001)
5618 Spirited Away (Sen to Chihiro no kamikakushi) (2001)
7064 Beauty and the Beast (Belle et la bête, La) (1946)
8360 Shrek 2 (2004)
36401 Brothers Grimm, The (2005)
45730 Lady in the Water (2006)
1032 Alice in Wonderland (1951)
2116 Lord of the Rings, The (1978)
34150 Fantastic Four (2005)
```

2096 Sleeping Beauty (1959)
2125 Ever After: A Cinderella Story (1998)
8952 Falling Angels (2003)
1197 Princess Bride, The (1987)
3159 Fantasia 2000 (1999)
4370 A.I. Artificial Intelligence (2001)
2291 Edward Scissorhands (1990)
4306 Shrek (2001)
8580 Into the Woods (1991)
48394 Pan's Labyrinth (El Laberinto del fauno) (2006)
54259 Stardust (2007)
594 Snow White and the Seven Dwarfs (1937)
595 Beauty and the Beast (1991)
42734 Hoodwinked (2006)

Final result

```
hive> select movieid,moviename from movienames where movieid in (select distinct movieid from movietags where tag="fairy tale");
Query ID = nitishpandey_20190602063734_37f2ce59-9c73-4005-93e4-e36efd5e506a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1559453325848_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.53 s
--------------------------------------------------------------------------------
OK
531     Secret Garden, The (1993)
588     Aladdin (1992)
837     Matilda (1996)
1022    Cinderella (1950)
2143    Legend (1985)
4973    Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001)
5618    Spirited Away (Sen to Chihiro no kamikakushi) (2001)
7064    Beauty and the Beast (Belle et la bête, La) (1946)
8360    Shrek 2 (2004)
36401   Brothers Grimm, The (2005)
45730   Lady in the Water (2006)
1032    Alice in Wonderland (1951)
2116    Lord of the Rings, The (1978)
34150   Fantastic Four (2005)
2096    Sleeping Beauty (1959)
2125    Ever After: A Cinderella Story (1998)
8952    Falling Angels (2003)
1197    Princess Bride, The (1987)
3159    Fantasia 2000 (1999)
4370    A.I. Artificial Intelligence (2001)
2291    Edward Scissorhands (1990)
4306    Shrek (2001)
8580    Into the Woods (1991)
48394   Pan's Labyrinth (El Laberinto del fauno) (2006)
54259   Stardust (2007)
594     Snow White and the Seven Dwarfs (1937)
595     Beauty and the Beast (1991)
42734   Hoodwinked (2006)
Time taken: 18.809 seconds, Fetched: 28 row(s)
hive>
```

part 02 (Find the top 10 ranking movies for the year (say 2000))

creating movieRatings table

```
hive> create external table movieRatings(userid int,movieid
int,rating int,dated bigint)
> ROW FORMAT SERDE
"org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe"
> WITH SERDEPROPERTIES ("field.delim"="::")
> STORED AS TEXTFILE;
```

```
hive> create external table movieRatings(userid int,movieid int,rating int,dated
    > ROW FORMAT SERDE "org.apache.hadoop.hive.contrib.serde2.MultiDelimitSerDe"
    > WITH SERDEPROPERTIES ("field.delim"="::")
    > STORED AS TEXTFILE;
OK
Time taken: 4.1 seconds
hive> describe movieRatings;
OK
userid                  int                     from deserializer
movieid                 int                     from deserializer
rating                  int                     from deserializer
dated                   bigint                  from deserializer
Time taken: 0.08 seconds, Fetched: 4 row(s)
```

loading data into the table from HDFS

```
hive>LOAD data INPATH "ml-10m/ratings.dat" into TABLE movieratings;
```

```
hive> LOAD data INPATH "ml-10m/ratings.dat" into TABLE movieratings;
Loading data to table default.movieratings
OK
Time taken: 0.695 seconds
hive> add jar hive-jars/hive-contrib-2.3.3.jar;
Added [hive-jars/hive-contrib-2.3.3.jar] to class path
Added resources: [hive-jars/hive-contrib-2.3.3.jar]
hive> select * from movieratings limit 10;
OK
1       122     5       838985046
1       185     5       838983525
1       231     5       838983392
1       292     5       838983421
1       316     5       838983392
1       329     5       838983392
1       355     5       838984474
1       356     5       838983653
1       362     5       838984885
1       364     5       838983707
Time taken: 0.224 seconds, Fetched: 10 row(s)
hive>
```

creating view using movieratings table to get year from unix timestamp

```
hive> create view ratingswithdate as
    > select movieid,rating,FROM_UNIXTIME(dated,"Y") as year from
movieratings;
```

```
hive> create view ratingswithdate as
    > select movieid,rating,FROM_UNIXTIME(dated,"Y") as year from movieratings;
OK
Time taken: 0.226 seconds
hive> select * from ratingswithdate limit 10;
OK
122     5       1996
185     5       1996
231     5       1996
292     5       1996
316     5       1996
329     5       1996
355     5       1996
356     5       1996
362     5       1996
364     5       1996
Time taken: 0.203 seconds, Fetched: 10 row(s)
```

creating table ratingswithyear2000 containing all the movies and its ratings of year 2000 only.

```
hive> create table ratingswithyear2000 as
    > select * from ratingswithdate where year="2000";
```

```
hive> create table ratingswithyear2000 as
    > select * from ratingswithdate where year="2000";
Query ID = nitishpandey_20190603163128_d55f0c68-ca0f-4522-b12f-e6556bb793bd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1559577722921_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 33.15 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hadoop-m/user/hive/warehouse/ratingswithyear2000
OK
Time taken: 34.309 seconds
hive> select * from ratingswithyear2000 limit 10;
OK
2       3       2000
11      5       2000
17      5       2000
25      3       2000
39      4       2000
47      4       2000
50      5       2000
60      4       2000
105     3       2000
110     5       2000
Time taken: 0.151 seconds, Fetched: 10 row(s)
```

Getting top 10 moovie ids which are most popular (most rated)

```
create view topmovieids as
    > select movieid,count(movieid) as ratingscount from
ratingswithyear2000 group by movieid order by ratingscount desc;
```

```
hive> create view topmovieids as
    > select movieid,count(movieid) as ratingscount from ratingswithyear2000 group by movieid order by ratingscount desc;
OK
Time taken: 0.205 seconds
hive> select * from topmovieids limit 10;
Query ID = nitishpandey_20190603164406_d673a36b-7851-4938-9d63-c6bde1d93861
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1559577722921_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.20 s
--------------------------------------------------------------------------------
OK
2858    4281
260     4012
1196    3905
1210    3720
2028    3585
593     3556
480     3523
2571    3493
589     3470
608     3459
Time taken: 18.079 seconds, Fetched: 10 row(s)
hive>
```

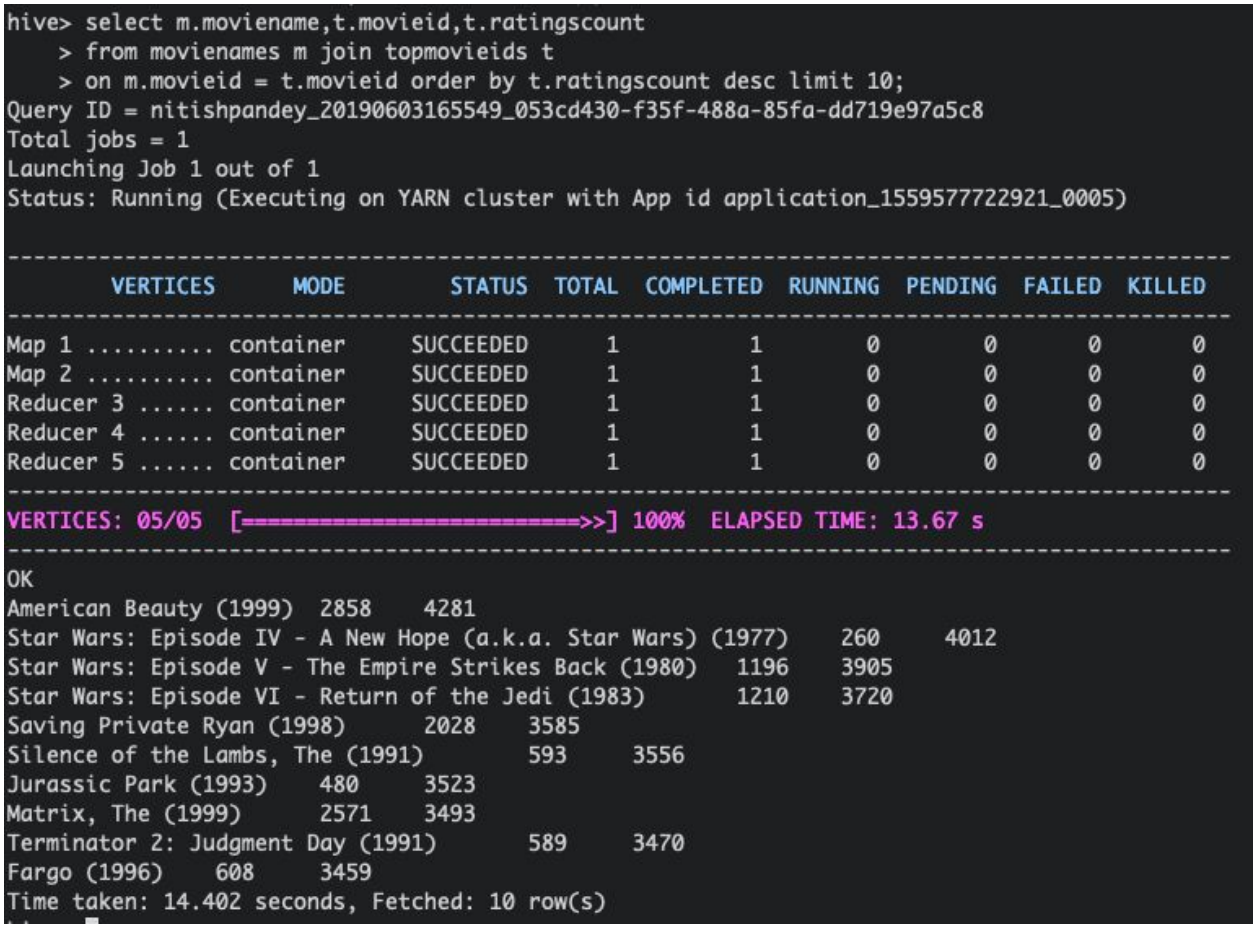
Showing top 10 movies

```
hive> select m.moviename,t.movieid,t.ratingscount
    > from movienames m join topmovieids t
    > on m.movieid = t.movieid order by t.ratingscount desc limit 10;
```

moviename	movieid	rating
American Beauty (1999)	2858	4281
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	260	4012
Star Wars: Episode V - The Empire Strikes Back (1980)	1196	3905
Star Wars: Episode VI - Return of the Jedi (1983)	1210	3720
Saving Private Ryan (1998)	2028	3585
Silence of the Lambs, The (1991)	593	3556
Jurassic Park (1993)	480	3523
Matrix, The (1999)	2571	3493
Terminator 2: Judgment Day (1991)	589	3470

Fargo (1996) 608 3459

(see screenshot in next page....)



```
hive> select m.moviename,t.movieid,t.ratingscount
    > from movienames m join topmovieids t
    > on m.movieid = t.movieid order by t.ratingscount desc limit 10;
Query ID = nitishpandey_20190603165549_053cd430-f35f-488a-85fa-dd719e97a5c8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1559577722921_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 13.67 s
--------------------------------------------------------------------------------
OK
American Beauty (1999)  2858    4281
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)    260     4012
Star Wars: Episode V - The Empire Strikes Back (1980)   1196    3905
Star Wars: Episode VI - Return of the Jedi (1983)       1210    3720
Saving Private Ryan (1998)      2028    3585
Silence of the Lambs, The (1991)        593     3556
Jurassic Park (1993)    480     3523
Matrix, The (1999)      2571    3493
Terminator 2: Judgment Day (1991)       589     3470
Fargo (1996)    608     3459
Time taken: 14.402 seconds, Fetched: 10 row(s)
```