CSEE5590 Big Data Programming

In Class Programming –4 Report

(Jongkook Son)

Project Overview:

Hive is a data warehousing system to store structured data on Hadoop file system and provides an easy query these data by execution Hadoop MapReduce plans. In this exercise we will learn basics of Hive QL.

Requirements/Task(s):

1. Create Hive Tables and Perform Queries for Use Case based on Petrol or hotel_bookings data. For Petrol, see the slides for details or you may try yourown queries using hotel_bookings data.

2. Create Hive Tables and Perform Queries for Use Case based on Olympics Data. See the Slides for details.

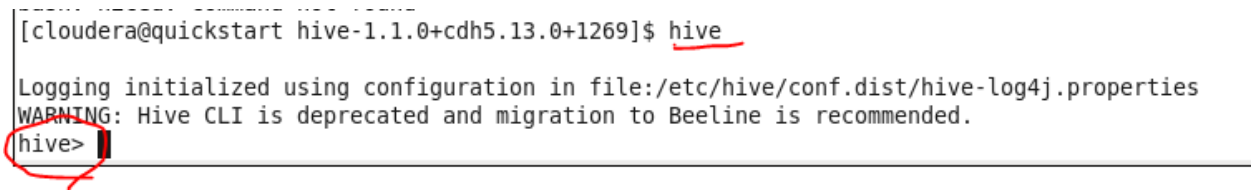
3. Create Hive Tables and Perform Queries for Use Case based on Movielens dataset which has 3 datasets as movies, users and ratings.

What I learned in ICP:

I could have get the further insight process of the Hive. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. It stores schema in a database and processed data into HDFS. It is designed for OLAP. It provides SQL type language for querying called HiveQL or HQL. It is familiar, fast, scalable, and extensible. In this ICP, I used cloudera cli to querying and analyze the datasets that are given. It was helpful to understand How hive works.

ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code

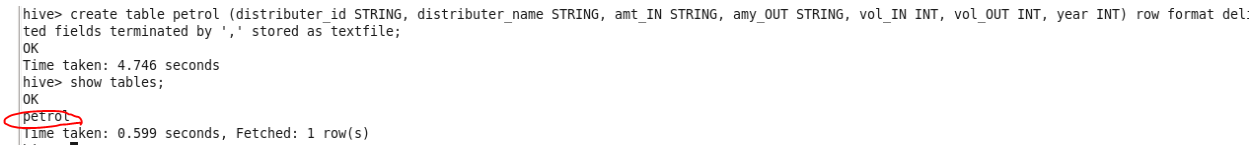
<Starting Hive>

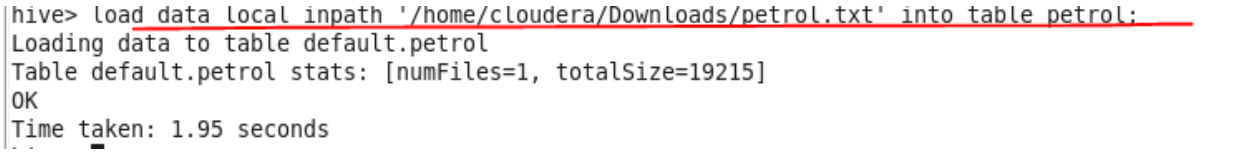


<Task1>

Create Hive Tables and Perform Queries for Use Case based on Petrol or hotel_bookings data. For Petrol, see the slides for details or you may try your own queries using hotel_bookings data.

<Create and load table petrol>





1)In real life what is the total amount of petrol in volume sold by every distributor?

```
hive> select distributer_name, SUM(vol_OUT) FROM petrol GROUP BY distributer_name;
Query ID = cloudera_20210215174949_e5a6d09b-d30b-4aed-be3d-b69acfd1572b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_161343534
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-15 17:49:44,545 Stage-1 map = 0%,  reduce = 0%
2021-02-15 17:49:55,689 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.13 sec
2021-02-15 17:50:07,658 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.39 sec
MapReduce Total cumulative CPU time: 2 seconds 390 msec
Ended Job = job_1613435349938_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.39 sec   HDFS Read: 27542 HDFS Write: 76 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 390 msec
```

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.22 sec   HDFS Read: 27631 HDFS Write: 56 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 220 msec
OK
Bharat  83662
hindustan       71767
reliance        76558
shell   69266
Time taken: 36.601 seconds, Fetched: 4 row(s)
```

2)Which are the top 10 distributors ID's for selling petrol and also display the amount of petrol sold in volume by them individually?

```
hive>  select distributer_id,vol_OUT from petrol order by vol_OUT desc limit 10;
Query ID = cloudera_20210215175555_992838d2-0d26-4126-b474-74255caa57d7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
```

```
OK
S8W 0P4 899
T1A 9W4 899
V8U 2T6 898
O8A 6Z5 897
O9P 9S3 897
F6W 6H3 896
E6O 9P1 895
N5Q 8E5 895
M6S 1P4 895
J4M 4G3 895
Time taken: 34.322 seconds, Fetched: 10 row(s)
```

3)Find real life 10 distributor name who sold petrol in the least amount.

```
hive> select distributer_id,vol_OUT from petrol order by vol_OUT limit 10;
Query ID = cloudera_20210215175858_059a0b17-f37b-4b26-876b-3fbfeee59b9f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-15 17:58:47,955 Stage-1 map = 0%,  reduce = 0%
2021-02-15 17:58:57,727 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.98 sec
2021-02-15 17:59:10,011 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.17 sec
Total MapReduce CPU Time Spent: 2 seconds 170 msec
OK
F4D 6K2 602
H7M 4M4 603
G9F 6U7 607
R3W 2E3 608
H4P 6A9 610
O5D 2R6 610
W0M 8R7 612
V0Z 0F6 612
O0D 0L1 612
L9H 1K6 613
Time taken: 34.679 seconds, Fetched: 10 row(s)
```

4) List all distributors who have this difference, along with the year and the difference which they have in that year.
Hint: (vol_IN-vol_OUT)>500

```
hive> SELECT year, distributer_name, (vol_IN-vol_OUT) FROM petrol where (vol_IN-vol_OUT)>500 ORDER BY distributer_name;
Query ID = cloudera_20210216232121_02a4ec8b-9c19-485e-8235-302498ad58c1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 23:21:46,746 Stage-1 map = 0%,  reduce = 0%
2021-02-16 23:21:59,200 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.44 sec
2021-02-16 23:22:09,937 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.47 sec
MapReduce Total cumulative CPU time: 2 seconds 470 msec
Ended Job = job_1613435349938_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.47 sec   HDFS Read: 28773 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 470 msec
OK
Time taken: 37.06 seconds
```

<Task2>

Create Hive Tables and Perform Queries for Use Case based on Olympics Data.

See the Slides for details.

1) Creation of Table in Hive and Loading of data

```
hive> create table olympic (athlete STRING, age INT, country STRING, year STRING, closing STRING, sport STRING, gold INT, silver INT, bronze INT, total INT)
ow format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 0.824 seconds
hive> load data local inpath '/home/cloudera/Downloads/olympic_data.csv' into table olympic;
Loading data to table default.olympic
Table default.olympic stats: [numFiles=1, totalSize=518669]
OK
Time taken: 1.971 seconds
hive>
```

2) Using the dataset list the total number of medals won by each country in swimming

```
hive> select country, SUM(total) from olympic where sport = "Swimming" group by country;
Query ID = cloudera_20210216214444_85cce5c8-fa5c-4c71-a861-3ca4794d089a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 21:44:51,031 Stage-1 map = 0%,  reduce = 0%
2021-02-16 21:45:04,909 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.83 sec
2021-02-16 21:45:19,579 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.5 sec
MapReduce Total cumulative CPU time: 3 seconds 500 msec
Ended Job = job_1613435349938_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.5 sec   HDFS Read: 528168 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 500 msec
OK
Australia       163
Austria 3
Belarus 2
Brazil  8
Canada  5
China   35
Costa Rica      2
Croatia 1
Denmark 1
France  39
Germany 32
Great Britain   11
Hungary 9
Italy   16
Japan   43
Lithuania       1
Netherlands     46
Norway  2
Poland  3
Romania 6
Russia  20
Serbia  1
Slovakia        2
Slovenia        1
South Africa    11
South Korea     4
Spain   3
Sweden  9
Trinidad and Tobago     1
Tunisia 3
Ukraine 7
United States   267
```

3) Display real life number of medals India won year wise.

```
hive> select year, SUM(total) from olympic where country = "India" group by year;
Query ID = cloudera_20210216214949_24c4e181-c6fa-45b1-bde2-e3a95f95216c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 21:49:18,818 Stage-1 map = 0%,  reduce = 0%
2021-02-16 21:49:30,935 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.5 sec
2021-02-16 21:49:41,659 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.65 sec
MapReduce Total cumulative CPU time: 2 seconds 650 msec
Ended Job = job_1613435349938_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.65 sec   HDFS Read: 528213 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 650 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 36.731 seconds, Fetched: 4 row(s)
```

4) Find the total number of medals each country won display the name along with total medals.

```
hive> select country, SUM(total) from olympic group by country;
Query ID = cloudera_20210216220606_94c7b226-e69c-47bb-9527-c2a2925bdd06
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_000
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 22:06:17,393 Stage-1 map = 0%,  reduce = 0%
2021-02-16 22:06:29,870 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.27 sec
2021-02-16 22:06:39,573 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.41 sec
MapReduce Total cumulative CPU time: 2 seconds 410 msec
Ended Job = job_1613435349938_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.41 sec   HDFS Read: 527157 HDFS Write: 1315 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 410 msec
OK
```

```
Russia  768
Saudi Arabia    6
Serbia  31
Serbia and Montenegro   38
Singapore       7
Slovakia        35
Slovenia        25
South Africa    25
South Korea     308
Spain   205
Sri Lanka       1
Sudan   1
Sweden  181
Switzerland     93
Syria   1
Tajikistan      3
Thailand        18
Togo    1
Trinidad and Tobago     19
Tunisia 4
Turkey  28
Uganda  1
Ukraine 143
United Arab Emirates    1
United States   1312
Uruguay 1
Uzbekistan      19
Venezuela       4
Vietnam 2
Zimbabwe        7
Time taken: 34.76 seconds, Fetched: 110 row(s)
```

5) Find the real life number of gold medals each country won.

```
hive> select country, SUM(gold) from olympic group by country;
Query ID = cloudera_20210216221212_bdebb1d1-52f7-44ae-8502-b8acd1cc053e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_00
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 22:12:33,827 Stage-1 map = 0%,  reduce = 0%
2021-02-16 22:12:43,557 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.2 sec
2021-02-16 22:12:54,264 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.33 sec
MapReduce Total cumulative CPU time: 2 seconds 330 msec
Ended Job = job_1613435349938_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.33 sec   HDFS Read: 527157 HDFS Write: 1276 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 330 msec
OK
```

```
Russia  234
Saudi Arabia    0
Serbia  1
Serbia and Montenegro   11
Singapore       0
Slovakia        10
Slovenia        5
South Africa    10
South Korea     110
Spain   19
Sri Lanka       0
Sudan   0
Sweden  57
Switzerland     21
Syria   0
Tajikistan      0
Thailand        6
Togo    0
Trinidad and Tobago     1
Tunisia 2
Turkey  9
Uganda  1
Ukraine 31
United Arab Emirates    1
United States   552
Uruguay 0
Uzbekistan      5
Venezuela       1
Vietnam 0
Zimbabwe        2
Time taken: 32.235 seconds, Fetched: 110 row(s)
```

6) Which country got medals for Shooting, year wise classification?

```
hive> select distinct(country), year from olympic where sport="Shooting" order by year, country;
Query ID = cloudera_20210216221717_1dca81ec-2701-41a8-aad5-c83ae0dc5e93
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-02-16 22:17:52,494 Stage-1 map = 0%,  reduce = 0%
2021-02-16 22:18:02,254 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.38 sec
2021-02-16 22:18:13,174 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.4 sec
MapReduce Total cumulative CPU time: 2 seconds 400 msec
Ended Job = job_1613435349938_0010
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0011
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-02-16 22:18:24,811 Stage-2 map = 0%,  reduce = 0%
2021-02-16 22:18:33,574 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.8 sec
```

```
Norway  2008
Russia  2008
Slovakia        2008
Slovenia        2008
South Korea     2008
Ukraine 2008
United States   2008
Belarus 2012
Belgium 2012
China   2012
Croatia 2012
Cuba    2012
Czech Republic  2012
Denmark 2012
France  2012
Great Britain   2012
India   2012
Italy   2012
Kuwait  2012
Poland  2012
Qatar   2012
Romania 2012
Russia  2012
Serbia  2012
Slovakia        2012
Slovenia        2012
South Korea     2012
Sweden  2012
Ukraine 2012
United States   2012
Time taken: 63.262 seconds, Fetched: 90 row(s)
```

<Task3>

Create Hive Tables and Perform Queries for Use Case based on Movielens dataset which has 3 datasets as movies, users and ratings.

1) Create 3 tables called movies, ratings and users. Load the data into tables.

<Create and load movies table>

```
hive> create table movies (movieId STRING, title STRING, genres ARRAY<STRING>) row format delimited fields terminated by ',' collection items terminated by '|
' stored as textfile;
OK
Time taken: 0.178 seconds
hive> load data local inpath '/home/cloudera/Downloads/movies.csv' into table movies;
Loading data to table default.movies
Table default.movies stats: [numFiles=1, totalSize=494431]
OK
Time taken: 0.476 seconds
```

<Create and load ratings table>

```
hive> create table ratings (userId STRING, movieId INT, rating DECIMAL(2,1), timestamp STRING) row format delimited fields terminated by ',' stored as textfi
e;
OK
Time taken: 0.064 seconds
hive> load data local inpath '/home/cloudera/Downloads/ratings.csv' into table ratings;
Loading data to table default.ratings
Table default.ratings stats: [numFiles=1, totalSize=2483723]
OK
Time taken: 0.341 seconds
```

<Create and load user table>

```
hive> create table users (userId INT, gender STRING, occupation INT,zipcode INT) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 0.09 seconds
hive> load data local inpath '/home/cloudera/Downloads/users.txt' into table users;
Loading data to table default.users
Table default.users stats: [numFiles=1, totalSize=116282]
OK
Time taken: 0.262 seconds
```

2) For movies table: List all movies with genre of movie is “Action” and “Drama”

```
hive> select title, genres from movies where array_contains(genres, 'Action') and array_contains(genres,'Drama');
OK
Tokyo Tribe (2014)      ["Action","Crime","Drama","Sci-Fi"]
Afro Samurai (2007)     ["Action","Adventure","Animation","Drama","Fantasy"]
Southpaw (2015) ["Action","Drama"]
High Rise (2015)        ["Action","Drama","Sci-Fi"]
Unforgiven (2013)       ["Action","Crime","Drama"]
Lost in the Sun (2015)  ["Action","Drama","Thriller"]
Swelter (2014)  ["Action","Drama","Thriller"]
Sherlock: The Abominable Bride (2016)   ["Action","Crime","Drama","Mystery","Thriller"]
The Huntsman Winter's War (2016)        ["Action","Adventure","Drama","Fantasy"]
I Am Wrath (2016)       ["Action","Crime","Drama","Thriller"]
Karate Bullfighter (1975)       ["Action","Drama"]
Sympathy for the Underdog (1971)        ["Action","Crime","Drama"]
The Adderall Diaries (2015)     ["Action","Drama","Thriller"]
Hazard (2005)   ["Action","Drama","Thriller"]
The Grandmother (1970)  ["Action","Drama"]
Batman: The Killing Joke (2016) ["Action","Animation","Crime","Drama"]
Kingsglaive: Final Fantasy XV (2016)    ["Action","Adventure","Animation","Drama","Fantasy","Sci-Fi"]
Jack Reacher: Never Go Back (2016)      ["Action","Crime","Drama","Mystery","Thriller"]
Mercury Plains (2016)   ["Action","Adventure","Drama"]
Ghost in the Shell (2017)       ["Action","Drama","Sci-Fi"]
Free Fire (2017)        ["Action","Crime","Drama"]
CHiPS (2017)    ["Action","Comedy","Drama"]
Band of Brothers (2001) ["Action","Drama","War"]
The Fate of the Furious (2017)  ["Action","Crime","Drama","Thriller"]
Okja (2017)     ["Action","Adventure","Drama","Sci-Fi"]
War for the Planet of the Apes (2017)   ["Action","Adventure","Drama","Sci-Fi"]
Dunkirk (2017)  ["Action","Drama","Thriller","War"]
Shot Caller (2017)      ["Action","Crime","Drama","Thriller"]
Death Wish (2018)       ["Action","Crime","Drama","Thriller"]
Jurassic World: Fallen Kingdom (2018)   ["Action","Adventure","Drama","Sci-Fi","Thriller"]
Time taken: 0.095 seconds, Fetched: 427 row(s)
```

3) For Ratings table: List movie ids of all movies with rating equal to 5.

```
hive> select movieID, rating from ratings where rating = 5;
72142   5
73569   5
76091   5
76093   5
78499   5
78836   5
81834   5
86142   5
86898   5
88129   5
89745   5
92420   5
93838   5
93840   5
96610   5
96832   5
100906  5
102125  5
103341  5
106920  5
107406  5
107771  5
109968  5
110501  5
112175  5
112183  5
112290  5
115149  5
115727  5
121231  5
122882  5
122920  5
138632  5
156371  5
158238  5
164179  5
168248  5
168250  5
168252  5
Time taken: 0.078 seconds, Fetched: 13211 row(s)
```

4) Find top 11 average rated "Action" movies with descending order of rating. (Hint: Need to perform join operation on Movies and Ratings table)

```
hive> select title, rating, genres from movies JOIN ratings ON movies.movieId = ratings.movieId where array_contains(genres, 'Action') order by rating desc
mit 11;
Query ID = cloudera_20210216225656_cd7d7e9e-fd17-4168-ad8f-fc867eee7055
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20210216225656_cd7d7e9e-fd17-4168-ad8f-fc867eee7055.log
2021-02-16 10:56:13     Starting to launch local task to process map join;      maximum memory = 1013645312
2021-02-16 10:56:17     Dump the side-table for tag: 0 with group count: 1499 into file: file:/tmp/cloudera/8358e0ea-8de7-4c5c-a598-c573842769d1/hive_2021-
16_22-56-04_382_5495026417637068969-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile00--.hashtable
2021-02-16 10:56:17     Uploaded 1 File to: file:/tmp/cloudera/8358e0ea-8de7-4c5c-a598-c573842769d1/hive_2021-02-16_22-56-04_382_5495026417637068969-1/-loc
10004/HashTable-Stage-2/MapJoin-mapfile00--.hashtable (126054 bytes)
2021-02-16 10:56:17     End of local task; Time Taken: 4.269 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1613435349938_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1613435349938_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1613435349938_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-02-16 22:56:30,479 Stage-2 map = 0%,  reduce = 0%
2021-02-16 22:56:44,821 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2021-02-16 22:56:55,504 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.55 sec
MapReduce Total cumulative CPU time: 4 seconds 550 msec
Ended Job = job_1613435349938_0012
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.55 sec   HDFS Read: 2496446 HDFS Write: 632 SUCCESS
```

```
Deadpool 2 (2018)       5       ["Action","Comedy","Sci-Fi"]
Indiana Jones and the Kingdom of the Crystal Skull (2008)       5       ["Action","Adventure","Comedy","Sci-Fi"]
Executive Decision (1996)       5       ["Action","Adventure","Thriller"]
Blade Runner (1982)     5       ["Action","Sci-Fi","Thriller"]
300 (2007)      5       ["Action","Fantasy","War","IMAX"]
Terminator 2: Judgment Day (1991)       5       ["Action","Sci-Fi"]
First Knight (1995)     5       ["Action","Drama","Romance"]
Saving Private Ryan (1998)      5       ["Action","Drama","War"]
Star Wars: Episode V - The Empire Strikes Back (1980)   5       ["Action","Adventure","Sci-Fi"]
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)  5       ["Action","Adventure"]
Aliens (1986)   5       ["Action","Adventure","Horror","Sci-Fi"]
Time taken: 52.234 seconds, Fetched: 11 row(s)
```