CSEE5590 BIG DATA PROGRAMMING

Hadoop Dependent Query Based NoSQL Database Hive

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

Dataset:https://drive.google.com/open?id=0B1QaXx7tpw3SMTBqLUQwX0lOWnM

1. Create table and load data

```
hive> create table petrol (distributer_id STRING, distributer_name STRING, amt_IN STRING, amy_OUT
STRING, vol_IN INT, vol_OUT INT, year INT) row format delimited fields terminated by ',' stored as
textfile;
OK
Time taken: 0.236 seconds
```

```
hive> load data local inpath '/home/cloudera/Desktop/ICP_4/petrol.txt' into table petrol;
Loading data to table default.petrol
Table default.petrol stats: [numFiles=1, totalSize=19215]
OK
Time taken: 1.373 seconds
```

2. Group By

```
hive> SELECT distributer_name,SUM(vol_OUT)  FROM  petrol GROUP  BY distributer_n
ame;
Query ID = root_20200617161414_47e7c16e-e918-4a48-83bd-0931f2c9389b
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_1592431584364_0001, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592431584364_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 16:14:58,846 Stage-1 map = 0%,  reduce = 0%
2020-06-17 16:15:20,119 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 se
c
2020-06-17 16:15:36,609 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.57
sec
MapReduce Total cumulative CPU time: 4 seconds 570 msec
Ended Job = job_1592431584364_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.57 sec   HDFS Read: 27470 H
DFS Write: 76 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 570 msec
OK
Bharat  83662
Distributer name        NULL
hindustan       71767
reliance        76558
shell   69266
Time taken: 73.726 seconds, Fetched: 5 row(s)
```

3. Order By

```
    > select distributer_id,vol_OUT from petrol order by vol_OUT desc limit 10;
Query ID = root_20200617161818_74154136-16cd-495c-8dd7-e07624568249
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_1592431584364_0003, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592431584364_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 16:18:40,453 Stage-1 map = 0%,  reduce = 0%
2020-06-17 16:18:55,482 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.97 se
c
2020-06-17 16:19:09,211 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.09
sec
MapReduce Total cumulative CPU time: 4 seconds 90 msec
Ended Job = job_1592431584364_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.09 sec   HDFS Read: 26481 H
DFS Write: 120 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 90 msec
OK
T1A 9W4 899
S8W 0P4 899
V8U 2T6 898
O8A 6Z5 897
O9P 9S3 897
F6W 6H3 896
N5Q 8E5 895
M6S 1P4 895
E6O 9P1 895
J4M 4G3 895
Time taken: 48.443 seconds, Fetched: 10 row(s)
```

4. Sort By

```
hive> select distributer_id,vol_OUT from petrol SORT BY vol_OUT  limit 10;
Query ID = root_20200617220808_c9077686-76c4-421d-8c54-13f465682667
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_1592431584364_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:08:23,306 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:08:35,416 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.6 sec
2020-06-17 22:08:48,623 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.43 sec
MapReduce Total cumulative CPU time: 3 seconds 430 msec
Ended Job = job_1592431584364_0016
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_1592431584364_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0017
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-06-17 22:09:01,591 Stage-2 map = 0%,  reduce = 0%
2020-06-17 22:09:10,397 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.27 sec
2020-06-17 22:09:22,447 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.22 sec
MapReduce Total cumulative CPU time: 3 seconds 220 msec
Ended Job = job_1592431584364_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.43 sec   HDFS Read: 25609 HDFS Write: 377 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.22 sec   HDFS Read: 5247 HDFS Write: 123 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 650 msec
OK
District.ID     NULL
F4D 6K2 602
H7M 4M4 603
G9F 6U7 607
R3W 2E3 608
O5D 2R6 610
H4P 6A9 610
V0Z 0F6 612
O0D 0L1 612
```

5. Cluster By

```
hive>  select distributer_id,vol_OUT from petrol cluster by vol_OUT  limit 10;
Query ID = root_20200617174343_c7882981-7fca-4aa6-bd7d-6fac22c5ad82
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_1592431584364_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 17:44:05,933 Stage-1 map = 0%,  reduce = 0%
2020-06-17 17:44:22,220 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.07 sec
2020-06-17 17:44:38,493 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.33 sec
MapReduce Total cumulative CPU time: 4 seconds 330 msec
Ended Job = job_1592431584364_0013
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_1592431584364_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-06-17 17:44:59,622 Stage-2 map = 0%,  reduce = 0%
2020-06-17 17:45:15,070 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.68 sec
2020-06-17 17:45:45,682 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.42 sec
MapReduce Total cumulative CPU time: 4 seconds 420 msec
Ended Job = job_1592431584364_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.33 sec   HDFS Read: 25627 HDFS Write: 377 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.42 sec   HDFS Read: 5249 HDFS Write: 123 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 750 msec
OK
District.ID     NULL
F4D 6K2 602
H7M 4M4 603
G9F 6U7 607
R3W 2E3 608
O5D 2R6 610
H4P 6A9 610
V0Z 0F6 612
O0D 0L1 612
```

6. Distribute By

```
    >  select distributer_id,vol_OUT from petrol distribute by vol_OUT  limit 10;
Query ID = cloudera_20200617174545_b9dbbf3c-7eb4-49e7-9613-dc4d4cda3e9f
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_1592431584364_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0015,
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 17:45:52,130 Stage-1 map = 0%,  reduce = 0%
2020-06-17 17:46:05,631 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.9 sec
2020-06-17 17:46:21,639 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.53 sec
MapReduce Total cumulative CPU time: 4 seconds 530 msec
Ended Job = job_1592431584364_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.53 sec   HDFS Read: 26398 HDFS Write: 120 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 530 msec
OK
B4N 8E1 651
E7T 3Q5 709
P6V 7Q2 870
M8G 7Y9 867
J1O 5K1 631
Y1Z 7G3 718
S0I 5M8 844
F5X 8A5 720
I6M 4U3 737
A4U 9B2 731
Time taken: 61.164 seconds, Fetched: 10 row(s)
```

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

Dataset:https://umkc.box.com/s/f918eea7k6mw6h7qiwj4b8im97c6hy84

1. 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_20200617223333_4645fee8-fb01-4420-8959-2c1e17cf7615
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_1592431584364_0018, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592431584364_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0018
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:34:05,365 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:34:16,695 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.31 se
c
2020-06-17 22:34:31,852 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.95
sec
MapReduce Total cumulative CPU time: 4 seconds 950 msec
Ended Job = job_1592431584364_0018
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.95 sec   HDFS Read: 528168
HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 950 msec
OK
Argentina       1
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
```

2. 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_20200617223838_afdea0ac-7a46-428b-b72d-a73e9dc5e966
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_1592431584364_0019, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:39:11,457 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:39:26,461 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.14 sec
2020-06-17 22:39:50,674 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.78 sec
MapReduce Total cumulative CPU time: 5 seconds 780 msec
Ended Job = job_1592431584364_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.78 sec   HDFS Read: 528213 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 780 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 59.717 seconds, Fetched: 4 row(s)
```

3. 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_20200617223939_08a0c33f-1bba-45fe-bd16-be95d8b6b131
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_1592431584364_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:39:58,752 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:40:16,262 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.22 sec
2020-06-17 22:40:35,921 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.52 sec
MapReduce Total cumulative CPU time: 6 seconds 520 msec
Ended Job = job_1592431584364_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.52 sec   HDFS Read: 527094 HDFS Write: 1315 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 520 msec
OK
Afghanistan     2
Algeria 8
Argentina       141
Armenia 10
Australia       609
Austria 91
Azerbaijan      25
Bahamas 24
Bahrain 1
Barbados        1
Belarus 97
Belgium 18
Botswana        1
Brazil  221
Bulgaria        41
Cameroon        20
Canada  370
Chile   22
China   530
Chinese Taipei  20
Colombia        13
Costa Rica      2
Croatia 81
Cuba    188
Cyprus  1
Czech Republic  81
```

4. Find the real-life number of gold medals each country won.

```
hive> select country, SUM(gold) from olympic group by country;
Query ID = cloudera_20200617224040_59b6de3e-53e1-4315-b0cd-5e20de7d7b60
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_1592431584364_0021, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:41:07,495 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:41:21,562 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.88 sec
2020-06-17 22:41:45,646 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.76 sec
MapReduce Total cumulative CPU time: 5 seconds 760 msec
Ended Job = job_1592431584364_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.76 sec   HDFS Read: 527157 HDFS Write: 1276 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 760 msec
OK
Afghanistan     0
Algeria 2
Argentina       49
Armenia 0
Australia       163
Austria 36
Azerbaijan      6
Bahamas 11
Bahrain 0
Barbados        0
Belarus 17
Belgium 2
Botswana        0
Brazil  46
Bulgaria        8
Cameroon        20
Canada  168
Chile   3
China   234
Chinese Taipei  2
Colombia        2
Costa Rica      0
Croatia 35
Cuba    57
Cyprus  0
Czech Republic  14
```

5. 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_20200617224242_125dad96-006d-4c43-b539-f4e66d9ac71e
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_1592431584364_0022, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-17 22:42:20,123 Stage-1 map = 0%,  reduce = 0%
2020-06-17 22:42:32,571 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.84 sec
2020-06-17 22:42:47,075 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.93 sec
MapReduce Total cumulative CPU time: 4 seconds 930 msec
Ended Job = job_1592431584364_0022
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_1592431584364_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1592431584364_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0023
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-06-17 22:43:04,616 Stage-2 map = 0%,  reduce = 0%
2020-06-17 22:43:19,957 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.1 sec
2020-06-17 22:43:36,130 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.35 sec
MapReduce Total cumulative CPU time: 4 seconds 350 msec
Ended Job = job_1592431584364_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.93 sec   HDFS Read: 527003 HDFS Write: 2917 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.35 sec   HDFS Read: 7690 HDFS Write: 1271 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 280 msec
OK
Australia       2000
Azerbaijan      2000
Belarus 2000
Bulgaria        2000
China   2000
Czech Republic  2000
Denmark 2000
Finland 2000
France  2000
```

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

Dataset:https://umkc.box.com/s/m3i7oabkj00boxuiskv5d4aoklh85w3x

1. 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
Money Train (1995)      ["Action","Comedy","Crime","Drama","Thriller"]
Dead Presidents (1995)  ["Action","Crime","Drama"]
White Squall (1996)     ["Action","Adventure","Drama"]
Braveheart (1995)       ["Action","Drama","War"]
Bad Boys (1995) ["Action","Comedy","Crime","Drama","Thriller"]
Rob Roy (1995)  ["Action","Drama","Romance","War"]
First Knight (1995)     ["Action","Drama","Romance"]
Strange Days (1995)     ["Action","Crime","Drama","Mystery","Sci-Fi","Thriller"]
Outbreak (1995) ["Action","Drama","Sci-Fi","Thriller"]
Léon: The Professional (a.k.a. The Professional) (Léon) (1994)  ["Action","Crime","Drama","Thriller"]
Clear and Present Danger (1994) ["Action","Crime","Drama","Thriller"]
Bad Company (1995)      ["Action","Crime","Drama"]
Faster Pussycat! Kill! Kill! (1965)     ["Action","Crime","Drama"]
Menace II Society (1993)        ["Action","Crime","Drama"]
No Escape (1994)        ["Action","Drama","Sci-Fi"]
Rising Sun (1993)       ["Action","Drama","Mystery"]
RoboCop 3 (1993)        ["Action","Crime","Drama","Sci-Fi","Thriller"]
Romper Stomper (1992)   ["Action","Drama"]
Tombstone (1993)        ["Action","Drama","Western"]
Courage Under Fire (1996)       ["Action","Crime","Drama","War"]
Eraser (1996)   ["Action","Drama","Thriller"]
Daylight (1996) ["Action","Adventure","Drama","Thriller"]
Nothing to Lose (1994)  ["Action","Crime","Drama"]
Last Man Standing (1996)        ["Action","Crime","Drama","Thriller"]
Days of Thunder (1990)  ["Action","Drama","Romance"]
Palookaville (1996)     ["Action","Comedy","Drama"]
Apocalypse Now (1979)   ["Action","Drama","War"]
Once Upon a Time in the West (C'era una volta il West) (1968)   ["Action","Drama","Western"]
Henry V (1989)  ["Action","Drama","Romance","War"]
Diva (1981)     ["Action","Drama","Mystery","Romance","Thriller"]
Ben-Hur (1959)  ["Action","Adventure","Drama"]
Marked for Death (1990) ["Action","Drama"]
Under Siege (1992)      ["Action","Drama","Thriller"]
Sneakers (1992) ["Action","Comedy","Crime","Drama","Sci-Fi"]
Metro (1997)    ["Action","Comedy","Crime","Drama","Thriller"]
Rosewood (1997) ["Action","Drama"]
Best Men (1997) ["Action","Comedy","Crime","Drama"]
Volcano (1997)  ["Action","Drama","Thriller"]
Face/Off (1997) ["Action","Crime","Drama","Thriller"]
G.I. Jane (1997)        ["Action","Drama"]
Cop Land (1997) ["Action","Crime","Drama","Thriller"]
Fire Down Below (1997)  ["Action","Drama","Thriller"]
Mad City (1997) ["Action","Drama"]
Mercury Rising (1998)   ["Action","Drama","Thriller"]
All Quiet on the Western Front (1930)   ["Action","Drama","War"]
Lethal Weapon (1987)    ["Action","Comedy","Crime","Drama"]
Lethal Weapon 2 (1989)  ["Action","Comedy","Crime","Drama"]
```

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

```
hive> select title from movies JOIN ratings ON movies.movieId = ratings.movieId
where array_contains(genres, 'Action') order by rating desc limit 11;
FAILED: SemanticException [Error 10004]: Line 1:122 Invalid table alias or colum
n reference 'rating': (possible column names are: title)
hive> select title, rating from movies JOIN ratings ON movies.movieId = ratings.
movieId where array_contains(genres, 'Action') order by rating desc limit 11;
Query ID = cloudera_20200617225151_4a97f1c7-bfd8-444d-a163-b52387f04c71
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20200617225151_4a97f1c7-bfd8-444d-a163-
b52387f04c71.log
2020-06-17 10:51:17     Starting to launch local task to process map join;     m
aximum memory = 1013645312
2020-06-17 10:51:22     Dump the side-table for tag: 0 with group count: 1499 in
to file: file:/tmp/cloudera/71cb5b68-9107-4451-9d7d-09a69ac4df23/hive_2020-06-17
_22-51-05_118_5212597155394719593-1/-local-10004/HashTable-Stage-2/MapJoin-mapfi
le00--.hashtable
2020-06-17 10:51:22     Uploaded 1 File to: file:/tmp/cloudera/71cb5b68-9107-445
1-9d7d-09a69ac4df23/hive_2020-06-17_22-51-05_118_5212597155394719593-1/-local-10
004/HashTable-Stage-2/MapJoin-mapfile00--.hashtable (80691 bytes)
2020-06-17 10:51:22     End of local task; Time Taken: 4.601 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_1592431584364_0024, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592431584364_0024/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592431584364_0024
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-06-17 22:51:47,116 Stage-2 map = 0%,  reduce = 0%
2020-06-17 22:52:10,252 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 9.11 se
c
2020-06-17 22:52:28,904 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 12.27
 sec
MapReduce Total cumulative CPU time: 12 seconds 270 msec
Ended Job = job_1592431584364_0024
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 12.27 sec   HDFS Read: 249536
7 HDFS Write: 383 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 270 msec
OK
Deadpool 2 (2018)       5
Indiana Jones and the Kingdom of the Crystal Skull (2008)       5
Executive Decision (1996)       5
```

3. Find top 11 average rated "Action" movies with descending order of rating.

```
48516   5
50872   5
59315   5
60069   5
68954   5
76093   5
79132   5
109487  5
134853  5
166024  5
170705  5
14      5
39      5
45      5
47      5
50      5
110     5
165     5
260     5
296     5
300     5
318     5
337     5
367     5
373     5
457     5
474     5
475     5
527     5
589     5
593     5
608     5
21      5
28      5
39      5
53      5
85      5
99      5
111     5
162     5
163     5
175     5
194     5
199     5
215     5
232     5
296     5
326     5
334     5
```