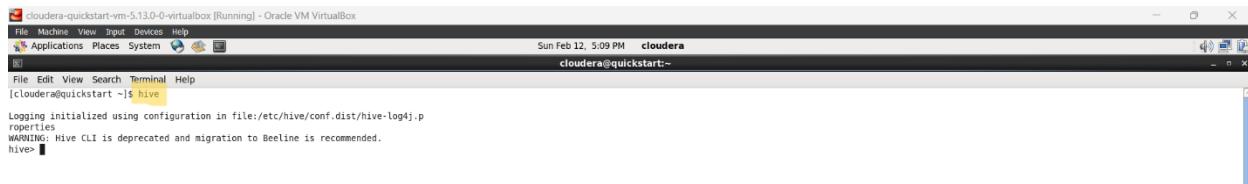


Introduction to Big Data, Data Science

ICE 3

Task 1

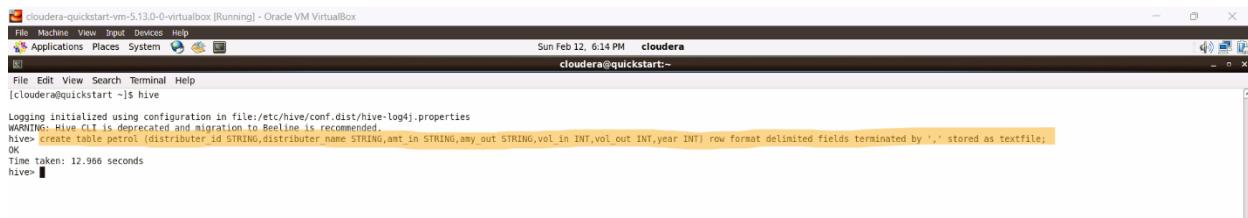
- Open Cloudera virtual machine and open terminal in the virtual machine
- Start hive from terminal by giving ‘hive’ command and hit on enter.



A screenshot of a terminal window titled 'cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox'. The window shows the system tray with icons for Applications, Places, System, and Help. The title bar displays the date 'Sun Feb 12, 5:09 PM' and the user 'cloudera'. The terminal prompt is 'cloudera@quickstart:~'. The user has typed '[cloudera@quickstart ~]\$ hive' and pressed enter. The output shows logging initialization and a warning about Hive CLI being deprecated and migration to Beeline being recommended.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Feb 12, 5:09 PM cloudera
cloudera@quickstart:~
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>
```

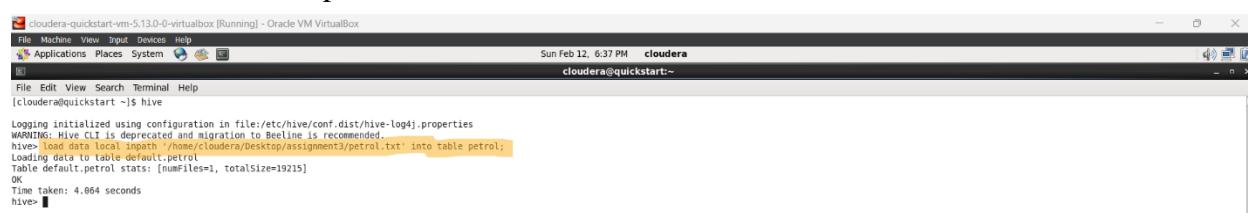
- Create Petrol table by using create command. Syntax for creating the table is provided below.
Create table <table name> (<attribute> <attribute data type>, ...) row format delimiter fields terminated by <column delimiter> lines terminated by <lines delimiter> stored as <file type>;



A screenshot of a terminal window titled 'cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox'. The window shows the system tray with icons for Applications, Places, System, and Help. The title bar displays the date 'Sun Feb 12, 6:14 PM' and the user 'cloudera'. The terminal prompt is 'cloudera@quickstart:~'. The user has typed '[cloudera@quickstart ~]\$ hive' and pressed enter. The output shows logging initialization, a warning about Hive CLI being deprecated and migration to Beeline being recommended, and the execution of the 'create table petrol' command. The command specifies the table name 'petrol', column types (distributor_id STRING, distributor_name STRING, amt_in STRING, amt_out STRING, vol_in INT, vol_out INT, year INT), row format as delimited fields separated by commas, and storage as textfile. The time taken for the operation is 12.966 seconds.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Feb 12, 6:14 PM cloudera
cloudera@quickstart:~
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table petrol (distributor_id STRING,distributor_name STRING,amt_in STRING,amt_out STRING,vol_in INT,vol_out INT,year INT) row format delimited fields terminated by ',' stored as textfile;
OK
Time Taken: 12.966 seconds
hive>
```

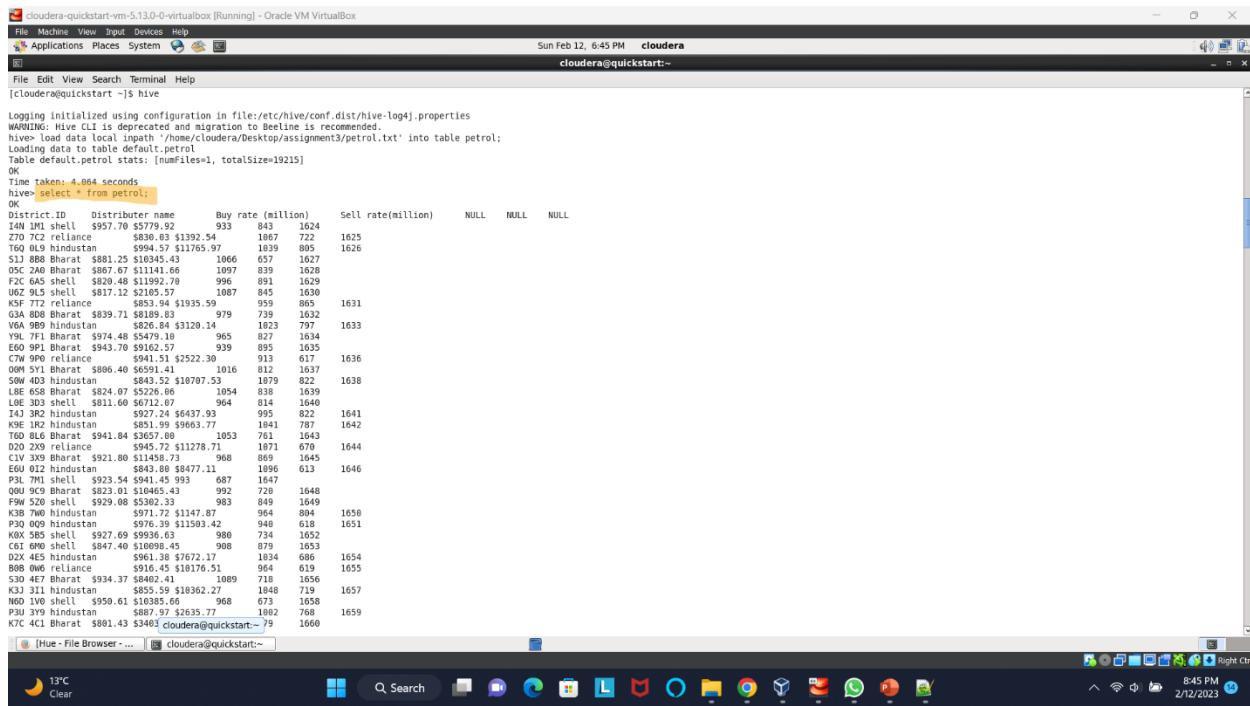
- Load data into petrol table. Syntax for loading data in to petrol table is given below
load data local inpath ‘<file location on local machine>’ into table <table name>;



A screenshot of a terminal window titled 'cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox'. The window shows the system tray with icons for Applications, Places, System, and Help. The title bar displays the date 'Sun Feb 12, 6:37 PM' and the user 'cloudera'. The terminal prompt is 'cloudera@quickstart:~'. The user has typed '[cloudera@quickstart ~]\$ hive' and pressed enter. The output shows logging initialization, a warning about Hive CLI being deprecated and migration to Beeline being recommended, and the execution of the 'load data local inpath' command. The command specifies the input path '/home/cloudera/Desktop/assignment3/petrol.txt' and the table name 'petrol'. The table default.stats are shown as [numFiles=1, totalSize=19215]. The time taken for the operation is 4.864 seconds.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Feb 12, 6:37 PM cloudera
cloudera@quickstart:~
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> load data local inpath '/home/cloudera/Desktop/assignment3/petrol.txt' into table petrol;
Loading data to table default.petrol...
Table default.petrol stats: [numFiles=1, totalSize=19215]
OK
Time Taken: 4.864 seconds
hive>
```

- Checking table data after loading data to the table using select command.

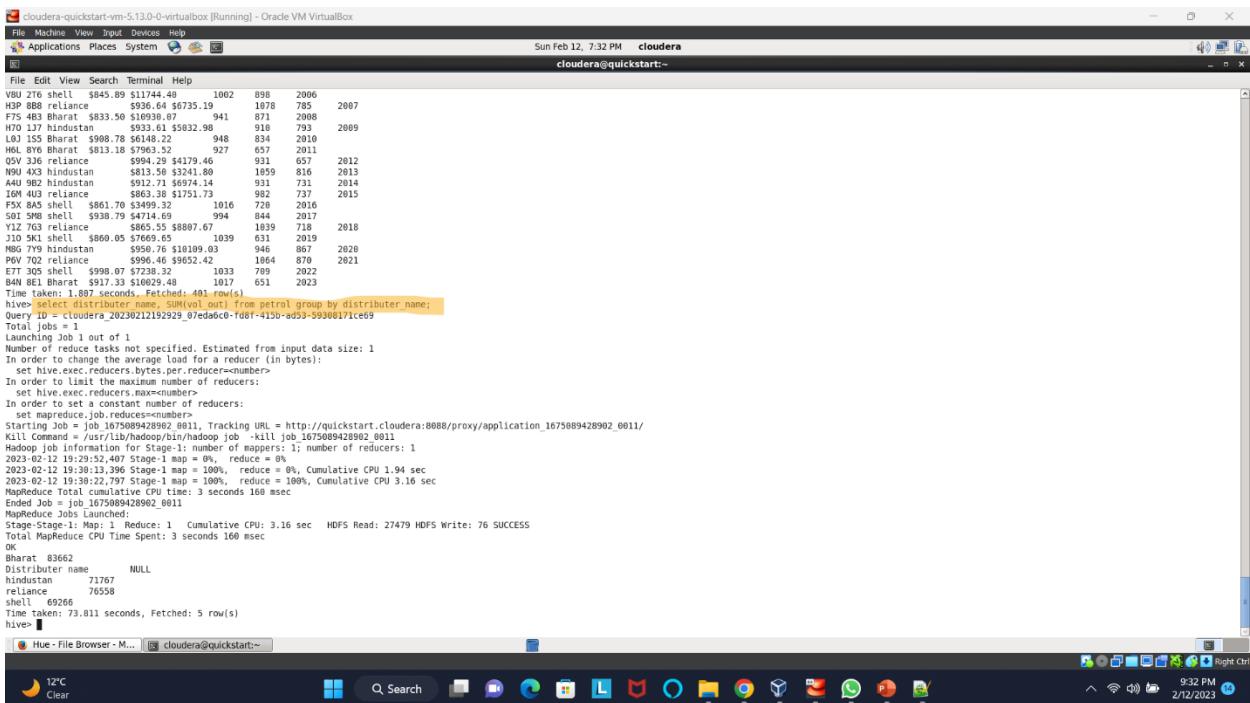


```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System  Sun Feb 12, 6:45 PM cloudera
cloudera@quickstart:-
[cloudera@quickstart ~]$ hive
hive> select * from petrol;
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> load data local inpath '/home/cloudera/Desktop/assignment3/petrol.txt' into table petrol;
Loading data to table default.petrol
Table default.petrol stats: [numFiles=1, totalSize=19215]
OK
Time taken: 4.064 seconds
hive> select * from petrol;
OK
+-----+-----+-----+-----+-----+
| District_ID | Distributor_name | Buy_rate_(million) | Sell_rate(million) | NULL | NULL |
+-----+-----+-----+-----+-----+
| IAN_1M_01 | Srilankan | 933 | 843 | 1624 | | |
| 270_01_reliance | Reliance | 8589.79 | 8139.2 | 1625 |
| T60_09_hindustan | Hindustan | 5994.57 | $11765.97 | 1839 | 805 | 1626 |
| S13_08_Bharat | Bharat | $881.25 | $1045.43 | 1066 | 657 | 1627 |
| 05C_2A_Bharat | Bharat | $867.67 | $11141.66 | 1097 | 839 | 1628 |
| F2C_6A5_shell | Shell | $820.48 | $10392.78 | 996 | 891 | 1629 |
| 002_05_hindustan | Hindustan | 821.57 | 1087 | 939 | 1030 | 1630 |
| KSF_772_reliance | Reliance | 6853.94 | $1935.59 | 959 | 865 | 1631 |
| G3A_8D_Bharat | Bharat | $839.71 | $18189.83 | 797 | 739 | 1632 |
| V6A_9B9_hindustan | Hindustan | $826.64 | $1320.14 | 1823 | 797 | 1633 |
| P9L_7B9_Bharat | Bharat | $974.46 | $879.19 | 808 | 765 | 1634 |
| E5G_01_reliance | Reliance | 943.76 | 8316.57 | 939 | 877 | 1635 |
| C7W_9P0_reliance | Reliance | 941.51 | $2522.30 | 913 | 617 | 1636 |
| 00M_5Y1_Bharat | Bharat | $886.40 | $6591.41 | 1016 | 812 | 1637 |
| S0W_4D5_hindustan | Hindustan | $843.52 | $10797.53 | 1079 | 822 | 1638 |
| L8E_5D9_hindustan | Hindustan | 824.09 | 1054 | 910 | 959 | 1639 |
| L0E_3D3_shell | Shell | $811.60 | $6712.87 | 964 | 814 | 1640 |
| I43_3R2_hindustan | Hindustan | $927.24 | $6437.93 | 995 | 822 | 1641 |
| K9E_1R2_hindustan | Hindustan | $851.99 | $5963.77 | 1043 | 787 | 1642 |
| T6D_8L6_Bharat | Bharat | $941.84 | $3637.80 | 1053 | 791 | 1643 |
| D0D_9F1_reliance | Reliance | $821.45 | $11278.71 | 908 | 873 | 1644 |
| C1V_3X9_Bharat | Bharat | $921.80 | $1458.73 | 968 | 869 | 1645 |
| E6U_02Z_hindustan | Hindustan | $843.88 | $8477.11 | 1096 | 613 | 1646 |
| P3L_7M_hindustan | Hindustan | $923.54 | $941.45 | 993 | 687 | 1647 |
| D0U_9P9_hindustan | Hindustan | $822.01 | $1045.43 | 992 | 720 | 1648 |
| F9w_5Z9_hindustan | Hindustan | 829.08 | 53692.23 | 903 | 881 | 1649 |
| K3B_7W0_hindustan | Hindustan | $971.72 | $1147.87 | 964 | 804 | 1650 |
| P30_009_hindustan | Hindustan | $976.39 | $11593.42 | 940 | 618 | 1651 |
| K0X_5B5_shell | Shell | $927.69 | $5936.63 | 980 | 734 | 1652 |
| C8J_4D7_hindustan | Hindustan | $847.46 | $1045.45 | 908 | 871 | 1653 |
| D2X_4E5_hindustan | Hindustan | $961.38 | $7672.17 | 1034 | 686 | 1654 |
| B0B_0W6_reliance | Reliance | $916.45 | $10767.51 | 964 | 619 | 1655 |
| S30_4E7_Bharat | Bharat | $934.37 | $8462.41 | 1089 | 718 | 1656 |
| K3J_31L_hindustan | Hindustan | $850.51 | $118362.27 | 1048 | 751 | 1657 |
| M0Q_3H9_shell | Shell | $959.61 | $10766.66 | 968 | 671 | 1658 |
| P3U_3V9_hindustan | Hindustan | $887.97 | $2635.77 | 1082 | 768 | 1659 |
| K7C_4C1_Bharat | Bharat | $881.43 | $3403 | cloudera@quickstart:-

```

- Performing group by operation on petrol table data. The table data is grouped based on distributor name and the corresponding total sum of the volume out is displayed right next to the distributor name.

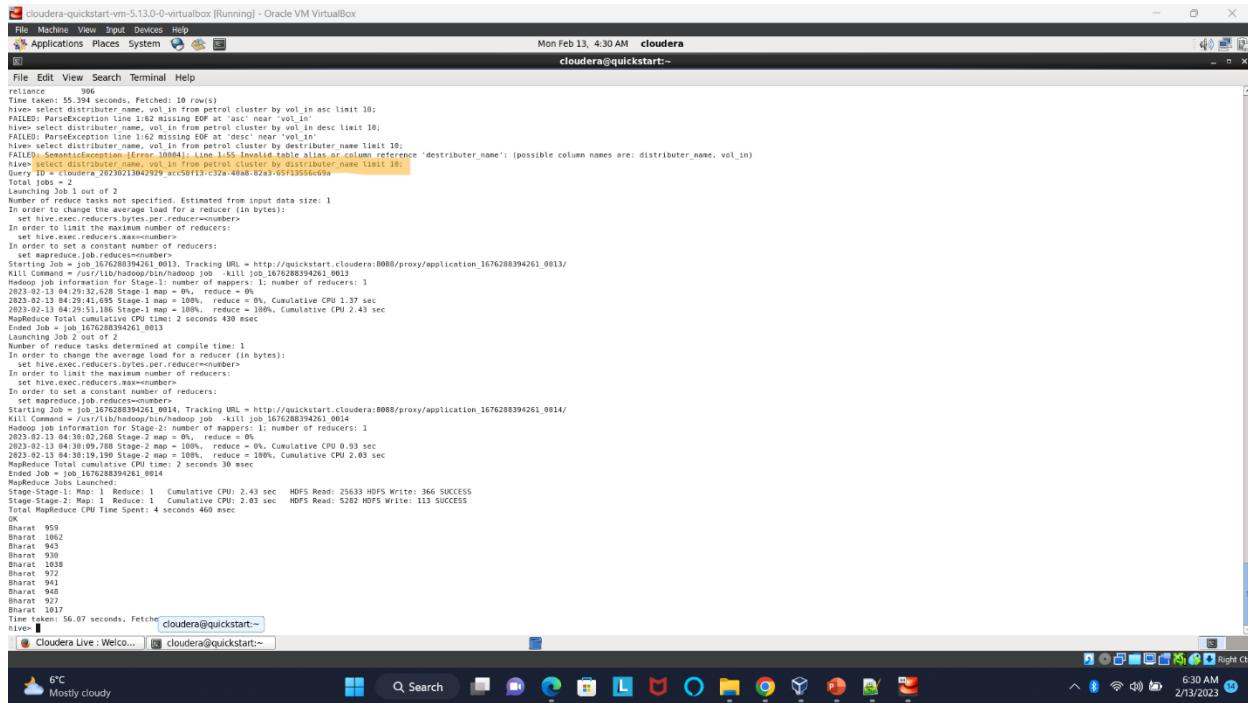


```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System  Sun Feb 12, 7:32 PM cloudera
cloudera@quickstart:-
[cloudera@quickstart ~]$ hive
hive> select distributor_name, sum(vol_out) from petrol group by distributor_name;
Query ID = cloudera_20230212192929_07eda6c0-1dbf-415b-ad52-5930e017ice09
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.reducer.bytes.per.reducer=reduces;
In order to limit the 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_1675089428902_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1675089428902_0011/
Kill Command: /usr/lib/hadoop/bin/hadoop job -kill job_1675089428902_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-12 19:29:52,407 Stage-1: map = 100%, reduce = 0%, Cumulative CPU 1.94 sec
2023-02-12 19:30:13,396 Stage-1: map = 100%, reduce = 0%, Cumulative CPU 3.16 sec
MapReduce Total cumulative time: 3 seconds 160 msec
Ended Job = job_1675089428902_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.16 sec HDFS Read: 27479 HDFS Write: 76 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 160 msec
OK
Bharat: 83662
Distributor name: NULL
hindustan: 71767
reliance: 76558
shell: 69266
Time taken: 73.811 seconds, Fetched: 5 row(s)
hive> 

```


- The query used in the figure below demonstrates the use case of ‘cluster by’.
- ‘Cluster by’ in hive is used to cluster similar data in the column and sends that result to different reducers.



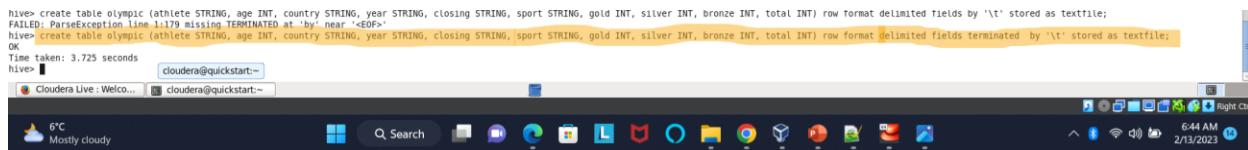
```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
File Applications Places System cloudera
cloudera@quickstart:~ Mon Feb 13, 4:30 AM
File Edit View Search Terminal Help
Time taken: 55.32 seconds. Fetching: 10 rows(s)
hive> select distributor_name, vol_in from petrol_cluster by vol_in asc limit 10;
FAILED: ParseException line 1:62 missing EOF at 'asc' near 'vol_in'
hive> select distributor_name, vol_in from petrol_cluster by distributor_name asc limit 10;
FAILED: ParseException line 1:62 missing EOF at 'desc' near 'vol_in'
hive> select distributor_name, vol_in from petrol_cluster by distributor_name asc limit 10;
FAILED: ParseException line 1:62 missing EOF at 'desc' near 'distributor_name'
hive> select distributor_name, vol_in from petrol_cluster by distributor_name limit 10;
Query ID = cloudera_1676288394261_0013
Total MapReduce Jobs Launched:
Launching Job 1 out of 2
Number of reduce tasks: 1. Specified. Estimated frame input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<numbers>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<numbers>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<numbers>
Start Time: Mon Feb 13 04:29:32 2023 Stage-1: number of mappers: 1; number of reducers: 1
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 04:29:32.628 Stage-1 map = 100%, reduce = 0%
2023-02-13 04:29:41.695 Stage-1 map = 100%, reduce = 0%. Cumulative CPU 1.37 sec
2023-02-13 04:30:19.196 Stage-2 map = 100%, reduce = 0%. Cumulative CPU 2.43 sec
MapReduce Total cumulative CPU Time: 2 seconds 430 msec
Ended Job = job_1676288394261_0013
Landed in HDFS as /user/cloudera/petrol_cluster
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=<numbers>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<numbers>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<numbers>
Start Time: Mon Feb 13 04:29:39 2023 Stage-1: number of mappers: 1; number of reducers: 1
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 04:30:02.780 Stage-2 map = 100%, reduce = 0%. Cumulative CPU 0.93 sec
2023-02-13 04:30:19.788 Stage-2 map = 100%, reduce = 0%. Cumulative CPU 1.37 sec
2023-02-13 04:30:19.196 Stage-2 map = 100%, reduce = 0%. Cumulative CPU 2.03 sec
MapReduce Total cumulative CPU Time: 2 seconds 400 msec
Ended Job = job_1676288394261_0014
MapReduce Jobs Launched:
Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.43 sec HDFS Read: 25633 HDFS Write: 366 SUCCESS
Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.03 sec HDFS Read: 5282 HDFS Write: 113 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 460 msec
OK
Bharat 959
Bharat 1062
Bharat 930
Bharat 930
Bharat 938
Bharat 972
Bharat 941
Bharat 938
Bharat 927
Bharat 1017
Time taken: 56.07 seconds. Fetching: cloudera@quickstart:~
hive> [REDACTED]
Cloudera Live : Welco... cloudera@quickstart:~ [REDACTED]
6°C Mostly cloudy [REDACTED]

```

Task 2

- Created Olympic table with all the provided query in the canvas.



```

hive> create table olympic (athlete STRING, age INT, country STRING, year STRING, closing STRING, sport STRING, gold INT, silver INT, bronze INT, total INT) row format delimited fields by '\t' stored as textfile;
FAILED: ParseException line 1:179 missing TERMINATED at 'by' near '<EOF>'

hive> create table olympic (athlete STRING, age INT, country STRING, year STRING, closing STRING, sport STRING, gold INT, silver INT, bronze INT, total INT) row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 1.725 seconds
hive> [REDACTED]
cloudera@quickstart:~ [REDACTED]
Cloudera Live : Welco... cloudera@quickstart:~ [REDACTED]
6°C Mostly cloudy [REDACTED]

```

- Loaded Olympic data to Olympic table using load data command.

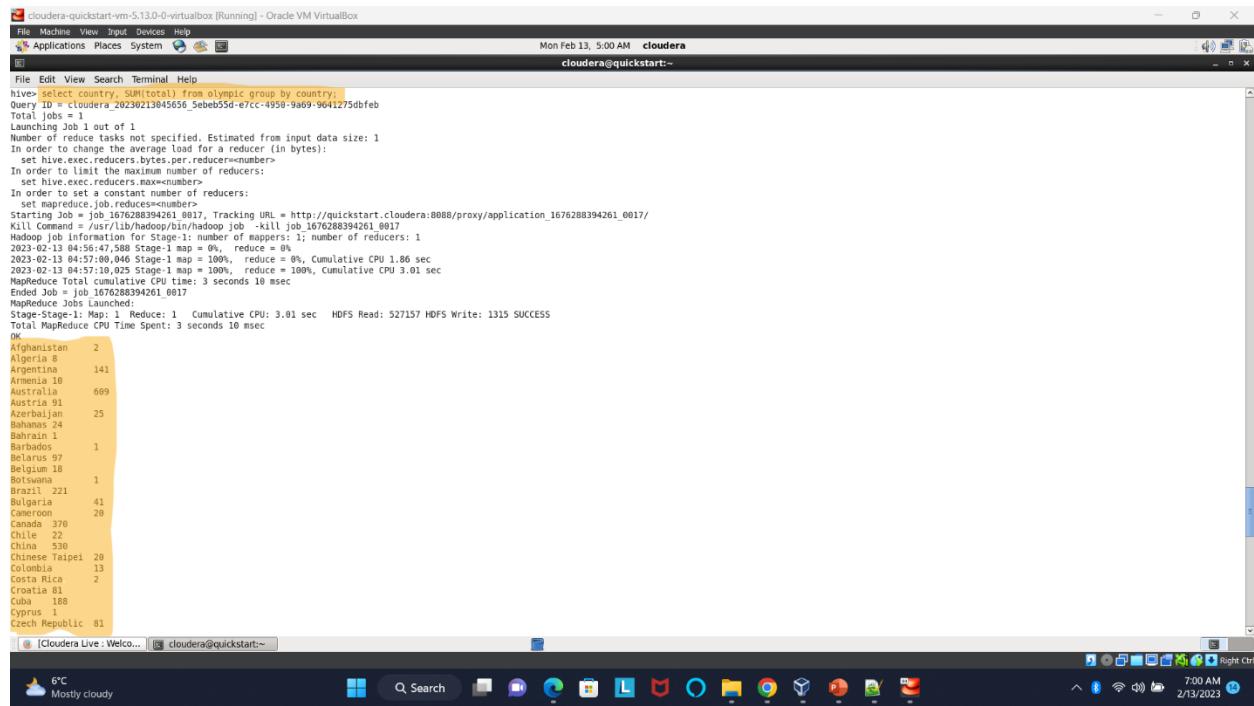


```

OK
Time taken: 3.772 seconds
hive> load data local input '/home/cloudera/Desktop/assignment3/olympic_data.csv' into table olympic;
Loading data to table default.olympic
Table default.olympic stats: [numFiles=1, totalSize=518669]
OK
Time taken: 1.306 seconds
hive> [REDACTED]
cloudera@quickstart:~ [REDACTED]
Cloudera Live : Welco... cloudera@quickstart:~ [REDACTED]
6°C Mostly cloudy [REDACTED]

```

- Found the total number of medals used by each country. All the countries are grouped by using group by query and the total number of medals in the group are aggregated using sum function.



cloudera-quickstart-vm-5.13.0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~
Mon Feb 13, 5:00 AM cloudera@quickstart:~

```
hive> select country, SUM(total) from olympic group by country;
Query ID = cloudera_20230213045656_5ebdb55d-ercc-4959-9ab9-9041275dbfeb
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.reducer.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-1288394261_0017. Tracking URL = http://quickstart.cloudera:8080/proxy/application_1676288394261_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 04:56:47,508 Stage-1 map = 0%, reduce = 0%
2023-02-13 04:57:05,025 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.86 sec
2023-02-13 04:57:10,025 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.01 sec
MapReduce Total cumulative CPU time: 3 seconds 10 msec
Ended Job = job_1676288394261_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.01 sec HDFS Read: 527157 HDFS Write: 1315 SUCCESS
OK
Afghanistan 2
Algeria 8
Argentina 141
Armenia 10
Australia 699
Austria 91
Azerbaijan 25
Bahamas 24
Bahrain 1
Barbados 1
Belarus 97
Belgium 18
Botswana 1
Brazil 221
Bulgaria 41
Cameroon 28
Canada 370
Chile 22
China 530
Chinese Taipei 20
Croatia 13
Costa Rica 2
Cuba 188
Cyprus 1
Czech Republic 81
```

Cloudera Live: Welcome cloudera@quickstart:~

8°C Mostly cloudy

7:00 AM 2/13/2023

- The medals won for swimming by all the countries are found by getting all the records whose sport is ‘Swimming’ and the result is grouped based on country name. The sum of the medals won by country in swimming are aggregated and displayed beside the country name.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera
cloudera@quickstart:~ Mon Feb 13, 5:14 AM
File Edit View Search Terminal Help
hive> select country, sum(total) as total_medals from olympic where sport='Swimming' group by country order by total_medals;
Query ID = cloudera_20230213059999_e9f4308c-6d31-434d-b53e-5bla52ale964
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_1676288394261_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0020
Hadoop job information for stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:09:27,195 Stage-1 map = 0%, reduce = 0%
2023-02-13 05:09:34,719 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.46 sec
2023-02-13 05:09:46,772 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.33 sec
MapReduce Total cumulative CPU time: 2 seconds 338 msec
Ended Job = job_1676288394261_0020
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_1676288394261_0021, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0021
Hadoop job information for stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 05:10:02,743 Stage-2 map = 0%, reduce = 0%, Cumulative CPU 0.9 sec
2023-02-13 05:10:10,100 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.9 sec
2023-02-13 05:10:19,685 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.85 sec
MapReduce Total cumulative CPU time: 1 seconds 850 msec
Ended Job = job_1676288394261_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.33 sec HDFS Read: 527323 HDFS Write: 1016 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 1.85 sec HDFS Read: 5958 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 180 msec
OK
Argentines 1
Trinidad and Tobago 1
Slovenia 1
Serbia 1
Lithuania 1
Denmark 1
Croatia 1
Norway 2
Costa Rica 2
2023-02-13 05:14:27,195 Stage-1 map = 0%, reduce = 0%
2023-02-13 05:14:34,719 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.46 sec
2023-02-13 05:14:46,772 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.33 sec
MapReduce Total cumulative CPU time: 2 seconds 338 msec
Ended Job = job_1676288394261_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.33 sec HDFS Read: 527323 HDFS Write: 1016 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 1.85 sec HDFS Read: 5958 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 180 msec
OK
Argentines 1
Trinidad and Tobago 1
Slovenia 1
Serbia 1
Lithuania 1
Denmark 1
Croatia 1
Norway 2
Costa Rica 2
Slovakia 2
Belarus 2
Tunisia 3
Spain 3
Australia 3
Poland 3
South Korea 4
Canada 5
Romania 6
Zimbabwe 7
Ukraine 7
Brazil 8
Hungary 9
Sweden 9
Great Britain 11
South Africa 11
Italy 16
Russia 30
Germany 32
China 35
France 39
Japan 43
Netherlands 46
Australia 163
United States 267
Time taken: 63.155 seconds, Fetched: 34 row(s)
hive>
```

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera
cloudera@quickstart:~ Mon Feb 13, 5:14 AM
File Edit View Search Terminal Help
hive> select country, sum(total) as total_medals from olympic where sport='Swimming' group by country order by total_medals;
Query ID = cloudera_20230213059999_e9f4308c-6d31-434d-b53e-5bla52ale964
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_1676288394261_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0020
Hadoop job information for stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:10:02,743 Stage-2 map = 0%, reduce = 0%, Cumulative CPU 0.9 sec
2023-02-13 05:10:11,169 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.9 sec
2023-02-13 05:10:19,685 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.85 sec
MapReduce Total cumulative CPU time: 1 seconds 850 msec
Ended Job = job_1676288394261_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.33 sec HDFS Read: 527323 HDFS Write: 1016 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 1.85 sec HDFS Read: 5958 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 180 msec
OK
Argentines 1
Trinidad and Tobago 1
Slovenia 1
Serbia 1
Lithuania 1
Denmark 1
Croatia 1
Norway 2
Costa Rica 2
Slovakia 2
Belarus 2
Tunisia 3
Spain 3
Australia 3
Poland 3
South Korea 4
Canada 5
Romania 6
Zimbabwe 7
Ukraine 7
Brazil 8
Hungary 9
Sweden 9
Great Britain 11
South Africa 11
Italy 16
Russia 30
Germany 32
China 35
France 39
Japan 43
Netherlands 46
Australia 163
United States 267
Time taken: 63.155 seconds, Fetched: 34 row(s)
hive>
```

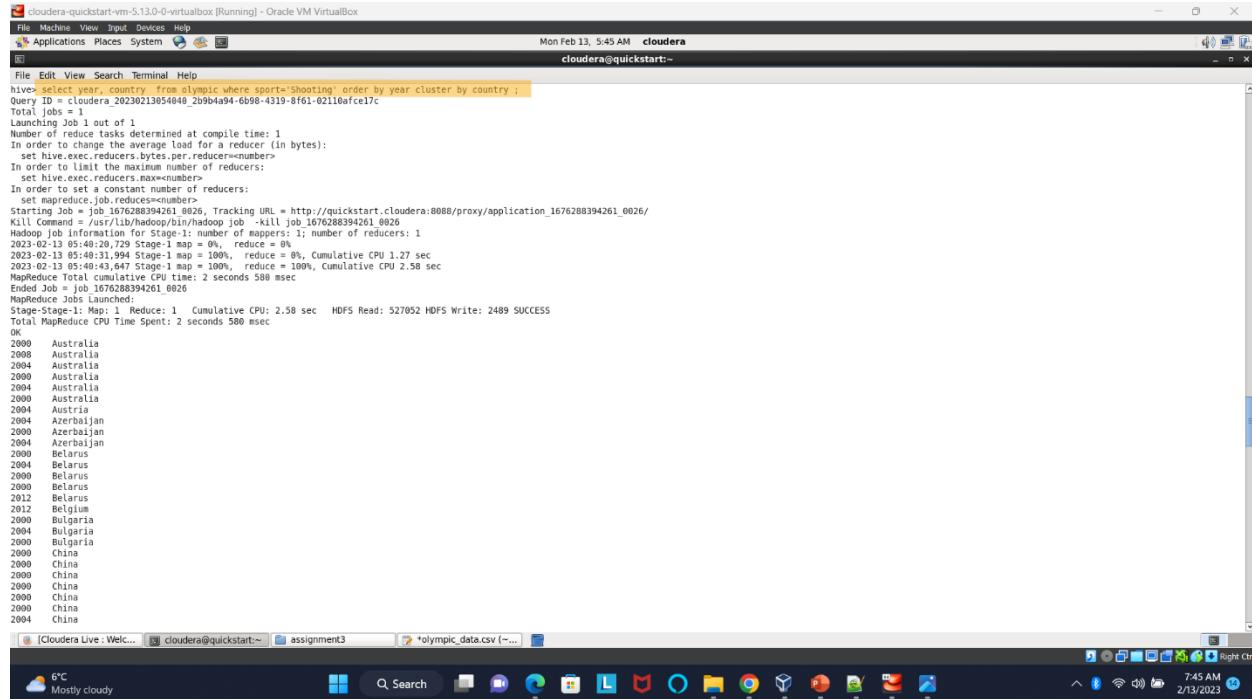
- Found all the medals that are won by countries for ‘Athletics’. All the countries are group using group by operation and the sum of the medals for athletics for each country are found using sum method and are displayed beside the country.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 5:24 AM cloudera
cloudera@quickstart:~>

File Edit View Search Terminal Help
hive> select country, SUM(total) as total_medals from olympic where sport='Athletics' group by country order by total_medals;
Query ID = cloudera_20230213052121_a19f346a-732e-4fc8-bf49-dfc37c4c02a9
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_1676288394261_0022, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:21:26,279 Stage-1 map = 0%, reduce = 0%
2023-02-13 05:21:38,789 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.83 sec
2023-02-13 05:21:48,238 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.88 sec
MapReduce Total cumulative CPU time: 2 seconds 888 msec
Ended Job = job_1676288394261_0022
Launching Job 2 out of 2
Number of reduce tasks not specified 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 see a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1676288394261_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0023
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 05:21:57,738 Stage-2 map = 0%, reduce = 0%
2023-02-13 05:22:16,706 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 0.99 sec
2023-02-13 05:22:16,706 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.99 sec
MapReduce Total cumulative CPU time: 1 seconds 998 msec
Ended Job = job_1676288394261_0023
MapReduce Jobs Completed=2
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.88 sec HDFS Read: 527327 HDFS Write: 1914 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 1.99 sec HDFS Read: 6856 HDFS Write: 748 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 870 msec
OK
Eritrea 1
Australia 1
Tunisia 1
Bahrain 1
Sudan 1
Sri Lanka 1
Barbados 1
Saudi Arabia 1
Botswana 1
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 5:24 AM cloudera
cloudera@quickstart:~>

File Edit View Search Terminal Help
hive> select country, SUM(total) as total_medals from olympic where sport='Athletics' group by country order by total_medals;
Query ID = cloudera_20230213052121_a19f346a-732e-4fc8-bf49-dfc37c4c02a9
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_1676288394261_0022, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:21:26,279 Stage-1 map = 0%, reduce = 0%
2023-02-13 05:21:38,789 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.83 sec
2023-02-13 05:21:48,238 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.88 sec
MapReduce Total cumulative CPU time: 2 seconds 888 msec
Ended Job = job_1676288394261_0022
Launching Job 2 out of 2
Number of reduce tasks not specified 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 see a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1676288394261_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0023
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 05:21:57,738 Stage-2 map = 0%, reduce = 0%
2023-02-13 05:22:16,706 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 0.99 sec
2023-02-13 05:22:16,706 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 1.99 sec
MapReduce Total cumulative CPU time: 1 seconds 998 msec
OK
Grenada 1
Cameroon 2
Croatia 2
Congo 2
Hungary 2
Denmark 3
New Zealand 3
Spain 3
Dominican Republic 3
Slovenia 3
Mexico 3
Latvia 3
Finland 3
Kazakhstan 3
Estonia 4
Sweden 4
Lithuania 4
Portugal 4
Algeria 5
Turkey 5
Norway 5
Belgium 5
South Africa 7
Brazil 7
Czech Republic 7
Italy 8
Japan 8
France 9
Romania 9
Greece 10
Poland 10
Nortocco 10
China 11
Belarus 15
Germany 16
Australia 16
Ukraine 17
Trinidad and Tobago 18
Cuba 21
Nigeria 21
Great Britain 23
Bahamas 24
Ethiopia 29
Kenya 39
Jamaica 60
Russia 98
United States 147
Time taken: 62.056 seconds, Fetched: 68 row(s)
hive>
```

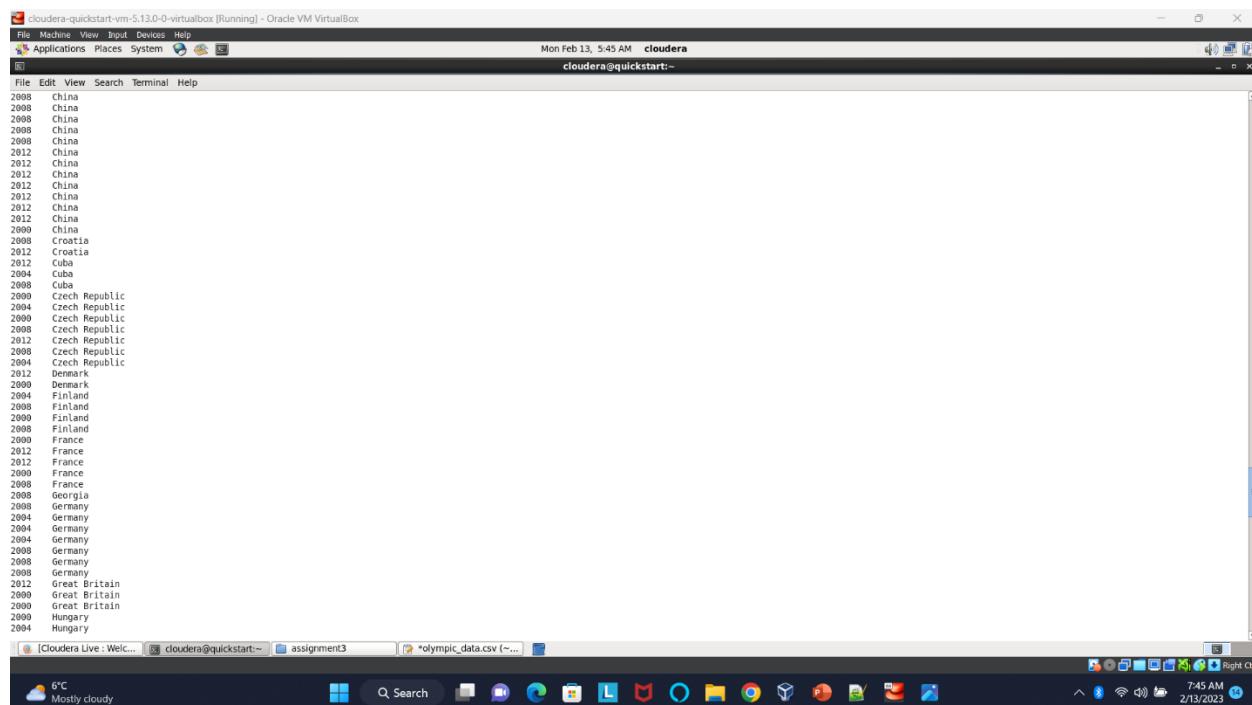
- Year wise classification of the countries who secured medals in shooting are extracted by clustering based on country name and sorted based on year.



```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera
cloudera@quickstart:~ Mon Feb 13, 5:45 AM
File Edit View Search Terminal Help
hive> select year, country from olympic where sport='Shooting' order by year cluster by country ;
Query ID = cloudera_20230213054040_20b4a94-6b98-4219-8f61-02110afce17c
Total jobs = 1
Launching Job 1 out of 1
Number of reducers last 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 mapred.reduce.tasks=<number>
Starting Job = job_1676288394261_0026, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0026
Hadoop job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:40:31,738 Stage-1: map = 100%, reduce = 0%, Cumulative CPU 1.27 sec
2023-02-13 05:40:31,994 Stage-1: map = 100%, reduce = 100%, Cumulative CPU 2.58 sec
MapReduce Total cumulative CPU time: 2 seconds 588 msec
Ended Job = job_1676288394261_0026
MapReduce Jobs Done
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.58 sec HDFS Read: 527052 HDFS Write: 2489 SUCCESS
OK
2000 Australia
2008 Australia
2004 Australia
2000 Australia
2004 Australia
2000 Austria
2004 Austria
2004 Azerbaijan
2000 Azerbaijan
2004 Azerbaijan
2000 Belarus
2004 Belarus
2000 Belarus
2012 Belarus
2000 Belgium
2000 Bulgaria
2004 Bulgaria
2000 Bulgaria
2000 China
2000 China
2000 China
2000 China
2004 China
2000 Croatia
2012 Croatia
2012 Cuba
2000 Cuba
2008 Cuba
2000 Czech Republic
2004 Czech Republic
2000 Czech Republic
2000 Czech Republic
2012 Czech Republic
2008 Czech Republic
2004 Czech Republic
2012 Denmark
2000 Finland
2004 Finland
2008 Finland
2000 Finland
2000 France
2012 France
2012 France
2000 France
2008 France
2008 Georgia
2008 Germany
2004 Germany
2004 Germany
2004 Germany
2008 Germany
2008 Germany
2008 Great Britain
2000 Great Britain
2000 Great Britain
2000 Hungary
2004 Hungary

```



```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera
cloudera@quickstart:~ Mon Feb 13, 5:45 AM
File Edit View Search Terminal Help
hive> select year, country from olympic where sport='Shooting' order by year cluster by country ;
Query ID = cloudera_20230213054040_20b4a94-6b98-4219-8f61-02110afce17c
Total jobs = 1
Launching Job 1 out of 1
Number of reducers last 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 mapred.reduce.tasks=<number>
Starting Job = job_1676288394261_0026, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676288394261_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676288394261_0026
Hadoop job Information for Stage-1: number of mappers: 1; number of reducers: 1
2023-02-13 05:40:31,738 Stage-1: map = 100%, reduce = 0%, Cumulative CPU 1.27 sec
2023-02-13 05:40:31,994 Stage-1: map = 100%, reduce = 100%, Cumulative CPU 2.58 sec
MapReduce Total cumulative CPU time: 2 seconds 588 msec
Ended Job = job_1676288394261_0026
MapReduce Jobs Done
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.58 sec HDFS Read: 527052 HDFS Write: 2489 SUCCESS
OK
2008 China
2008 China
2008 China
2008 China
2008 China
2012 China
2012 China
2012 China
2012 China
2012 China
2000 China
2008 Croatia
2012 Croatia
2012 Cuba
2000 Cuba
2008 Cuba
2000 Czech Republic
2004 Czech Republic
2000 Czech Republic
2000 Czech Republic
2012 Czech Republic
2008 Czech Republic
2004 Czech Republic
2012 Denmark
2000 Finland
2004 Finland
2008 Finland
2000 Finland
2000 France
2012 France
2012 France
2000 France
2008 France
2008 Georgia
2008 Germany
2004 Germany
2004 Germany
2004 Germany
2008 Germany
2008 Germany
2008 Great Britain
2000 Great Britain
2000 Great Britain
2000 Hungary
2004 Hungary

```

TASK 3

- Created movies table using create table query.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 8:52 PM cloudera
cloudera@quickstart:~$ 
File Edit View Search Terminal Help
190183 Sci-Fi\Thriller
190207 Drama\Romance
190209 Comedy
190213 Drama
190214 Drama
190219 Animation
190221 Documentary
191005 Action\Adventure\Comedy\Sci-Fi
193505 Action\Animation\Comedy\Sci-Fi
193507 Action\Thriller\Drama
193571 Comedy\Drama
193573 Animation
193579 Documentary
193580 Action\Animation\Comedy\Fantasy
193582 Action\Comedy\Fantasy
193585 Drama
193587 Action\Animation
193609 Comedy
Time taken: 1.31 seconds, Fetched: 9743 row(s)
hive> create table explodable as select movie_id, explode_genres from movies lateral view explode(genres) genretab as explode_genres;
FAILED: UDFArgumentException explode() takes an array or a map as a parameter
hive> clear
NoValueException[260]
    at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1828)
    at org.apache.hadoop.hive.ql.parse.ParsedDriver.parse(ParsedDriver.java:281)
    at org.apache.hadoop.hive.ql.parse.ParsedDriver.parse(ParsedDriver.java:166)
    at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:522)
    at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1356)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1473)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1285)
    at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1272)
    at org.apache.hadoop.hive.cli.CLI Driver.main(CLI Driver.java:226)
    at org.apache.hadoop.hive.cli.CLI Driver.processCmd(CLI Driver.java:175)
    at org.apache.hadoop.hive.cli.CLI Driver.processLine(CLI Driver.java:389)
    at org.apache.hadoop.hive.cli.CLI Driver.executeDriver(CLI Driver.java:781)
    at org.apache.hadoop.hive.cli.CLI Driver.run(CLI Driver.java:698)
    at org.apache.hadoop.hive.cli.CLI Driver.main(CLI Driver.java:634)
    at sun.reflect.NativeMethodAccessorImpl.invoke(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:494)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:138)
FAILED: ParseException line 1:0 cannot recognize input near 'clear' '<EOF>' '<EOF>'
hive> create table movies(movie_id int, title string, genres array<string>) row format delimited fields terminated by ',' collection items terminated by ']' stored as textfile;
OK
Time taken: 8.285 seconds
hive> 
```

The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox". The terminal is running on a Cloudera VM. The user has run several commands to create a table named "movies" and then attempt to use a UDF named "explode" which failed due to a type mismatch. The terminal also shows the creation of a new table named "explodable" using a lateral view explode function. The system status bar at the bottom indicates it's 16°C cloudy and the date is 2/13/2023.

- Loaded data into movies table using load query.

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 9:10 AM cloudera
cloudera@quickstart:~$ 
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table movies(movie_id INT, title STRING, genres STRING) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 14.059 seconds
hive> load data local inpath '/home/cloudera/Desktop/assignment3/movies.csv' into table movies
> ;
Loading data to table default.movies
Table default.movies stats: [numFiles=1, totalSize=494431]
OK
Time taken: 5.896 seconds
hive> 
```

The screenshot shows a terminal window titled "cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox". The user has run the "hive" command and then used the "load data local" command to load data from a CSV file named "movies.csv" into the "movies" table. The table has been created with columns "movie_id", "title", and "genres". The system status bar at the bottom indicates it's 13°C partly sunny and the date is 2/13/2023.

- Created Users table using create query for users.



```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 9:13 AM cloudera
cloudera@quickstart:~$ 
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table movies(movie_id INT, title STRING, genres STRING) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 14.059 seconds
hive> load data local inpath '/home/cloudera/Desktop/assignment3/movies.csv' into table movies
> ;
Loading data to table default.movies
Table default.movies stats: [numFiles=1, totalSize=494431]
OK
Time taken: 5.896 seconds
hive> create Table users(user_id INT, gender STRING, age int, occupation int, zipcode string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 1.107 seconds
hive> 

```

- Loaded data into users table using load data query.



```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Mon Feb 13, 9:14 AM cloudera
cloudera@quickstart:~$ 
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table movies(movie_id INT, title STRING, genres STRING) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 14.059 seconds
hive> load data local inpath '/home/cloudera/Desktop/assignment3/movies.csv' into table movies
> ;
Loading data to table default.movies
Table default.movies stats: [numFiles=1, totalSize=494431]
OK
Time taken: 5.896 seconds
hive> create table users(user_id INT, gender STRING, age int, occupation int, zipcode string) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 1.107 seconds
hive> load data local inpath '/home/cloudera/Desktop/assignment3/users.txt' into table users;
FAILED: ParseException line 1:10 extraneous input 'local' expecting INPATH near '<EOF>'
hive> load data local inpath '/home/cloudera/Desktop/assignment3/users.txt' into table users;
>Loading data to table default.users
Table default.users stats: [numFiles=1, totalSize=116282]
OK
Time taken: 4.705 seconds
hive> 

```

- Created ratings table using create table query.

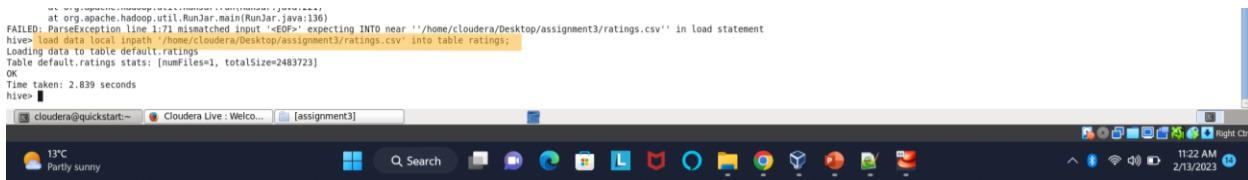


```

Time taken: 0.579 seconds
hive> create table ratings(user_id int, movie_id int, rating float, timestamp string) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.962 seconds
hive> 

```

- Loaded data into ratings table using load data query.



```

at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:71 mismatched input '<EOF>' expecting INTO near ''/home/cloudera/Desktop/assignment3/ratings.csv'' in load statement
hive> load data local inpath '/home/cloudera/Desktop/assignment3/ratings.csv' into table ratings;
>Loading data to table default.ratings
Table default.ratings stats: [numFiles=1, totalSize=2483723]
OK
Time taken: 2.839 seconds
hive> 

```

- Converted timestamp in string format to date format by typecasting the string to big int and converting to unix timestamp and then ‘to_date’.

```
OK
Time taken: 1.061 seconds
hive> select user_id, movie_id, rating, to_date(from_unixtime(timestamp DIV 1000)) as date from ratings limit 20;
FAILED: SemanticException Line 0:-1 Wrong arguments `1000': No matching method for class org.apache.hadoop.hive.ql.udf.UDFOPLongDivide with (string, int). Possible choices: _FUNC_(bigint, bigint)
hive> select user_id, movie_id, rating, to_date(from_unixtime(cast(timestamp as bigint) DIV 1000)) as date from ratings limit 20;
OK
NULL    NULL    NULL    NULL
1      1      4.0    1970-01-11
1      3      4.0    1970-01-11
1      6      4.0    1970-01-11
1      17     5.0    1970-01-11
1      50     5.0    1970-01-11
1      70     3.0    1970-01-11
1      101    5.0    1970-01-11
1      148    3.0    1970-01-11
1      151    5.0    1970-01-11
1      157    5.0    1970-01-11
1      163    5.0    1970-01-11
1      216    5.0    1970-01-11
1      224    3.0    1970-01-11
1      231    5.0    1970-01-11
1      235    4.0    1970-01-11
1      266    5.0    1970-01-11
1      296    3.0    1970-01-11
1      316    3.0    1970-01-11
1      333    5.0    1970-01-11
Time taken: 0.352 seconds, Fetched: 20 row(s)
hive>
```



- The movie table is joined with ratings table using inner join based on movie id of ratings table and movie id in movies table. Based on movie id the title of the movie is taken from movies table and the ratings are taken from ratings table and are displayed right next to the movie title.

```
cloudera@quickstart-vm:~/Desktop$ ./assignment3
Mon Feb 13 12:56 PM cloudera
cloudera@quickstart:~
```

The screenshot shows a terminal window titled 'cloudera@quickstart:~' with the following output:

```
File Edit View Search Terminal Help
File Machine View Input Devices Help
Applications Places System
Mon Feb 13 12:56 PM cloudera
cloudera@quickstart:~
```

```
File Edit View Search Terminal Help
FAILED: SemanticException [Error 10004]: Line 1:43 Invalid table alias or column reference 'ratings': (possible column names are: movie_id, title, genres)
hive> select movies.title, ratings.rating from ratings inner join movies on ratings where movies.movie_id == ratings.movie_id limit 20;
FAILED: SemanticException [Error 10004]: Line 1:70 Invalid table alias or column reference 'ratings': (possible column names are: movie_id, title, genres)
hive> select movies.title, ratings.rating from ratings inner join movies on movies.movie_id == ratings.movie_id limit 20;
OK
Query ID = cloudera_20230213125059_710db095-62de-42ab-bf88-c4ac9598479d
Total jobs: 1
Execution log at: /tmp/cloudera/cloudera_20230213125059_710db095-62de-42ab-bf88-c4ac9598479d.log
2023-02-13 12:58:36 Starting to launch local task to process map join; maximum memory = 932108464
2023-02-13 12:58:44 Dump the side-table for tag: 1 with group count: 9742 into file: file:/tmp/cloudera/e7bf992d-9a1a-4225-8ae7-a1b22f5c8a1/hive_2023-02-13 12-58-10_774_27454878227840276-1/-local-10003/HashTable-Stage-3/MapJoin-mapf
1 job(s) launched
2023-02-13 12:58:44 Uploaded 1 file to: file:/tmp/cloudera/e7bf992d-9a1a-4225-8ae7-a1b22f5c8a1/hive_2023-02-13 12-58-10_774_27454878227840276-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile01-.hashtable (439750 bytes)
2023-02-13 12:58:44 End of local task; Time Taken: 8.849 sec.
Execution completed successfully
MapReduceLocal tasks succeeded
Last MapReduce job: 1 job(s) launched
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1676306114521_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676306114521_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676306114521_0001
Hadoop job Information for Stage-3: number of mappers: 1; number of reducers: 0
2023-02-13 12:58:45 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 2.57 sec
MapReduce Total cumulative CPU time: 2 seconds 570 msec
Ended Job = job_1676306114521_0001
MapReduce Jobs Launched
Stage-3: 1 job(s) Launched
Cumulative CPU: 2.57 sec HDFS Read: 10711 HDFS Write: 513 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 570 msec
OK
Toy Story (1995)          4.0
Grumpier Old Men (1995)  4.0
Heat (1995)               4.0
Seven (a.k.a. Seven) (1995) 5.0
*Usual Suspects 5.0
From Dusk Till Dawn (1996) 3.0
Born on the Fourth of July (1996) 5.0
Braveheart (1995)        4.0
Rob Roy (1995)            5.0
Canadian Bacon (1995)    5.0
Berkeley Square (1995)    5.0
Billy Madison (1995)     5.0
Clerks (1994)              3.0
Dumb & Dumber (Dumb and Dumber) (1994) 5.0
Ed Wood (1994)             4.0
Star Wars: Episode IV - A New Hope (1977)      5.0
Pulp Fiction (1994)       3.0
Stargate (1994)            3.0
Tommy Boy (1995)           5.0
Clear and Present Danger (1994) 4.0
Time taken: 89.678 seconds, Fetched: 20 row(s)
hive>
```

The desktop environment includes a taskbar with various application icons and system status indicators like battery level (19°C Cloudy), network, and time (2:56 PM 2/13/2023).

TASK 4

- Created explodable table view for movie id and genere.

```
table default.movies stats: [numFiles=1, totalSize=694431]
OK
Time taken: 1.4 seconds
hive> create view explodable as select movie_id, explode_genres from movies lateral view explode(genres) genretab as explode_genres;
FAILED: SemanticException [Error 10004]: Line 1:45 Invalid table alias or column reference 'explode_genres': (possible column names are: movies.movie_id, movies.title, movies.genres, genretab.explode_genres)
hive> create view explodable as select movie_id, explode_genres from movies lateral view explode(genres) genretab as explode_genres;
OK
Time taken: 0.497 seconds
hive> 
```



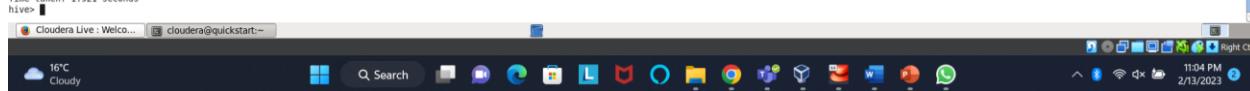
- Did select all for explodable table to view movie genere and rating of the genere and viewed on the screen.

```
Time taken: 1.4 seconds
hive> create view explodable as select movie_id, explode_genres from movies lateral view explode(genres) genretab as explode_genres;
FAILED: SemanticException [Error 10004]: Line 1:45 Invalid table alias or column reference 'explode_genres': (possible column names are: movies.movie_id, movies.title, movies.genres, genretab.explode_genres)
hive> create view explodable as select movie_id, explode_genres from movies lateral view explode(genres) genretab as explode_genres;
OK
Time taken: 0.497 seconds
hive> select * from explodable limit 20;
OK
NULL    genres
1       Adventure
1       Animation
1       Children
1       Comedy
1       Fantasy
2       Adventure
2       Children
2       Fantasy
3       Comedy
3       Romance
4       Comedy
4       Drama
4       Romance
5       Comedy
6       Action
6       Crime
6       Thriller
7       Comedy
7       Romance
Time taken: 1.323 seconds, Fetched: 20 row(s)
hive> 
```



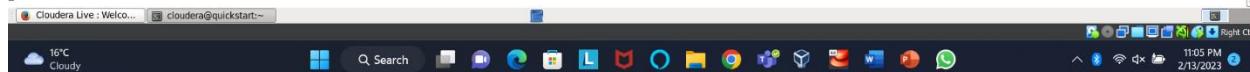
- Created another user table ‘user2’ to compute rank for the genre of movie for each user

```
FAILED: SemanticException [Error 10004]: Line 1:35 Invalid column reference 'movie_rating'. (possible column names are: e.movie_id, e.movie_name, e.rating, e.release_date, e.release_year, r.movie_id, r.rating, r.release_date)
hive> create view user2 as select user_id, explode_genres, count(rating) as rating from explodable e join ratings r on e.movie_id = r.movie_id group by user_id, explode_genres;
OK
Time taken: 1.921 seconds
hive> 
```



- Extracting and viewing the data for user2 table which is about the rank of the genere

```
OK
Time taken: 1.921 seconds
hive> select * from user2 order by user_id desc limit 10;
Query ID = cloudera_20230213210505_b8d3301c-245e-49b2-a706-99398d8476ed
Total jobs = 2
hive> 
```



```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
File Applications Places System
Mon Feb 13, 9:13 PM cloudera
cloudera@quickstart:-
File Edit View Search Terminal Help
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_1676342955407_0001, Tracking URL: http://quickstart.cloudera:8088/proxy/application_1676342955407_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676342955407_0001
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 21:06:02,481 Stage-2 map = 0%, reduce = 0%
2023-02-13 21:06:22,481 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 4.05 sec
2023-02-13 21:06:54,548 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 7.29 sec
MapReduce Total cumulative CPU time: 7 seconds 290 msec
Ended Job = job_1676342955407_0001
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_1676342955407_0002, Tracking URL: http://quickstart.cloudera:8088/proxy/application_1676342955407_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676342955407_0002
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2023-02-13 21:07:01,479 Stage-3 map = 0%, reduce = 0%, Cumulative CPU 2.68 sec
2023-02-13 21:07:15,337 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 4.38 sec
2023-02-13 21:07:15,337 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 4.38 sec
MapReduce Total cumulative CPU time: 4 seconds 380 msec
Ended Job = job_1676342955407_0002
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.29 sec HDFS Read: 2494822 HDFS Write: 746898 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 4.38 sec HDFS Read: 752564 HDFS Write: 149 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 670 msec
OK
1      Musical 18
1      IMAX 68
1      Horror 241
1      Film-Noir 9
1      Comedy 116
1      War 48
1      Drama 376
1      Thriller 398
1      Western 25
1      Documentary 4
Time taken: 122.726 seconds, Fetched: 10 row(s)

```

- created temporary view for top 3 generes for each user.

```

hive> select users.ID,explode_genres,movies.ratings from(select users.ID,explode_genres,movies.ratings,ROW_NUMBER() over(partition by users.ID order by movies.ratings DESC)
> as rank from explodable e join movieratings r on e.movie_ID=r.movies_ID) t where rank<=3 limit 10;
Query ID = cloudera_20230213193232_db38edc9-led3-4785-9a43-01fe5c85f5b0
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20230213193232_db38edc9-led3-4785-9a43-01fe5c85f5b0.log
2023-02-13 07:33:13 Starting to launch local task to process map join; maximum memory = 932184064
2023-02-13 07:33:19 Dump the side-table for tag: 0 with group count: 9742 into file: file:/tmp/cloudera/38d01a94-8ff5-4ebe-a500-e9bd337b34b8/hive_2023-02-13_19-32-40_641_6547829233:file30_-.hashtable
2023-02-13 07:33:19 Uploaded 1 File to: file:/tmp/cloudera/38d01a94-8ff5-4ebe-a500-e9bd337b34b8/hive_2023-02-13_19-32-40_641_654782923372189621-1/-local-10004/HashTable-Stage-2/Map
2023-02-13 07:33:19 End of local task; Time Taken: 5.67 sec.
Execution completed successfully
MapredLocal task succeeded
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>

```

```

Starting Job = job_1676138365403_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1676138365403_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1676138365403_0020
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-02-13 19:33:46,039 Stage-2 map = 0%, reduce = 0%
2023-02-13 19:34:47,669 Stage-2 map = 0%, reduce = 0%
2023-02-13 19:35:37,506 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 19.22 sec
2023-02-13 19:35:57,585 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 23.54 sec
MapReduce Total cumulative CPU time: 23 seconds 540 msec
Ended Job = job_1676138365403_0020
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 23.54 sec HDFS Read: 2497507 HDFS Write: 117 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 540 msec
OK
1      Romance 5
1      Comedy 5
1      War 5
2      Comedy 5
2      Documentary 5
2      Drama 5
3      Action 5
3      Thriller 5
3      Sci-Fi 5
4      Adventure 5
Time taken: 200.753 seconds, Fetched: 10 row(s)
hive>

```