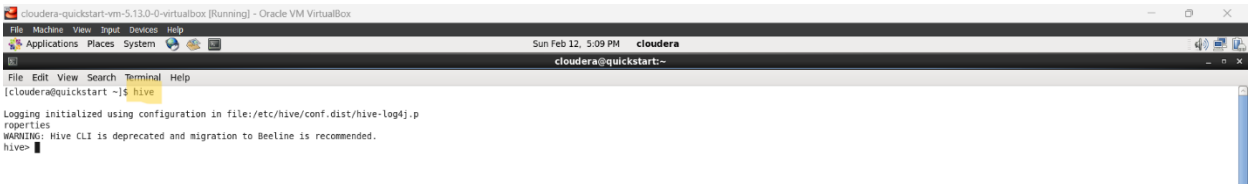
Introduction to Big Data, Data Science

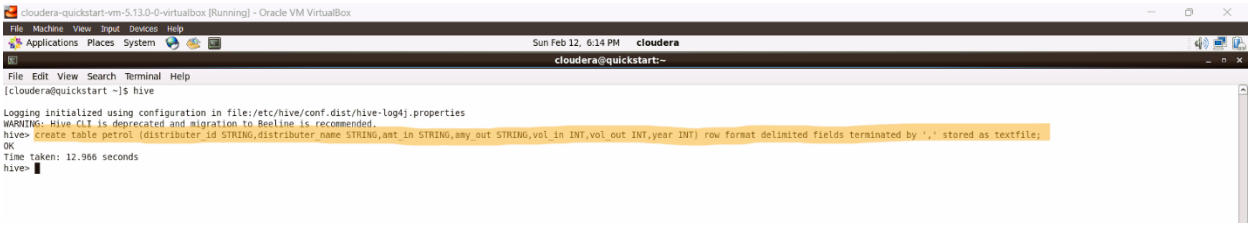
ICE 3

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.

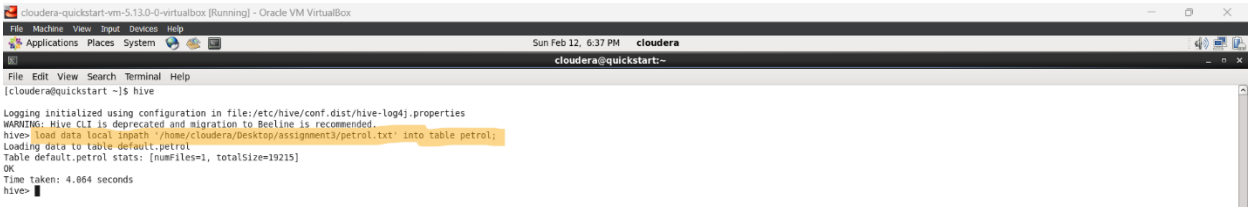


- Create Petrol table by using create command. Syntax for creating the table is provided below.
  Create table <table name> (<attribute> <attribute data type>, …) row format delimiter fields terminated by <column delimiter> lines terminated by <lines delimiter> stored as <file type>;

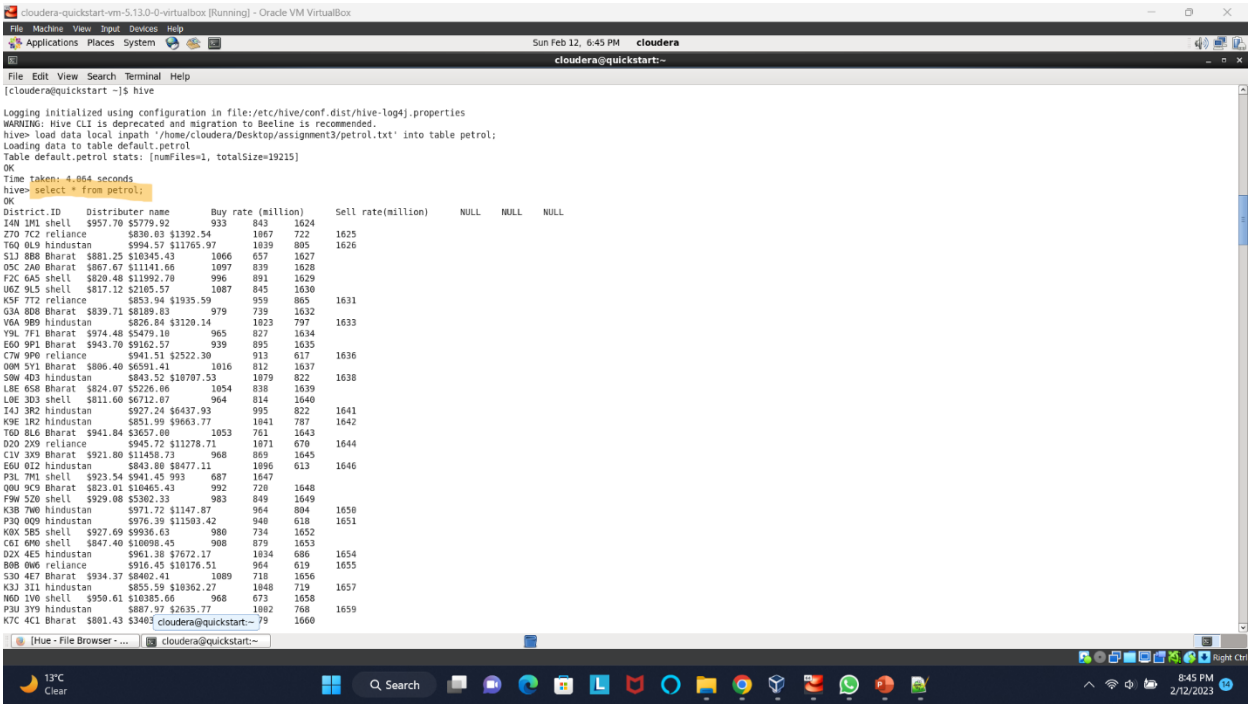


- Load data into petrol table. Syntax for loading data in to petrol table is given below

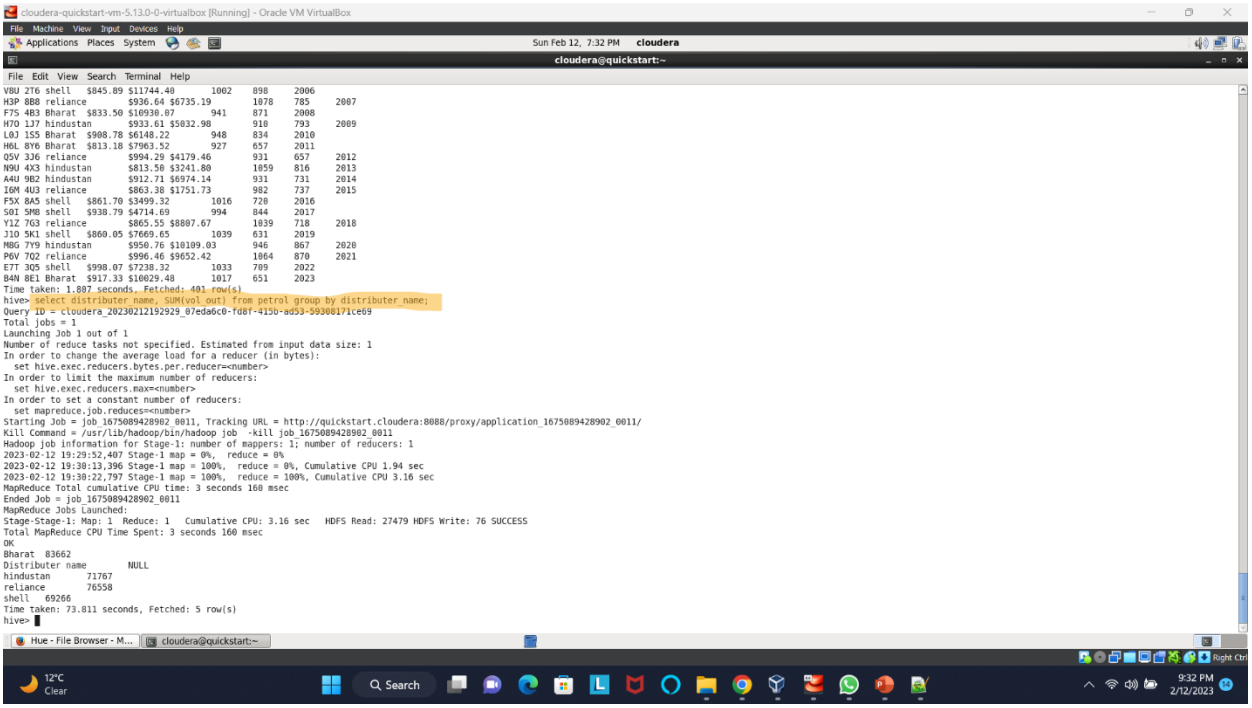
  load data local inpath ‘<file location on local machine>’ into table <table name>;

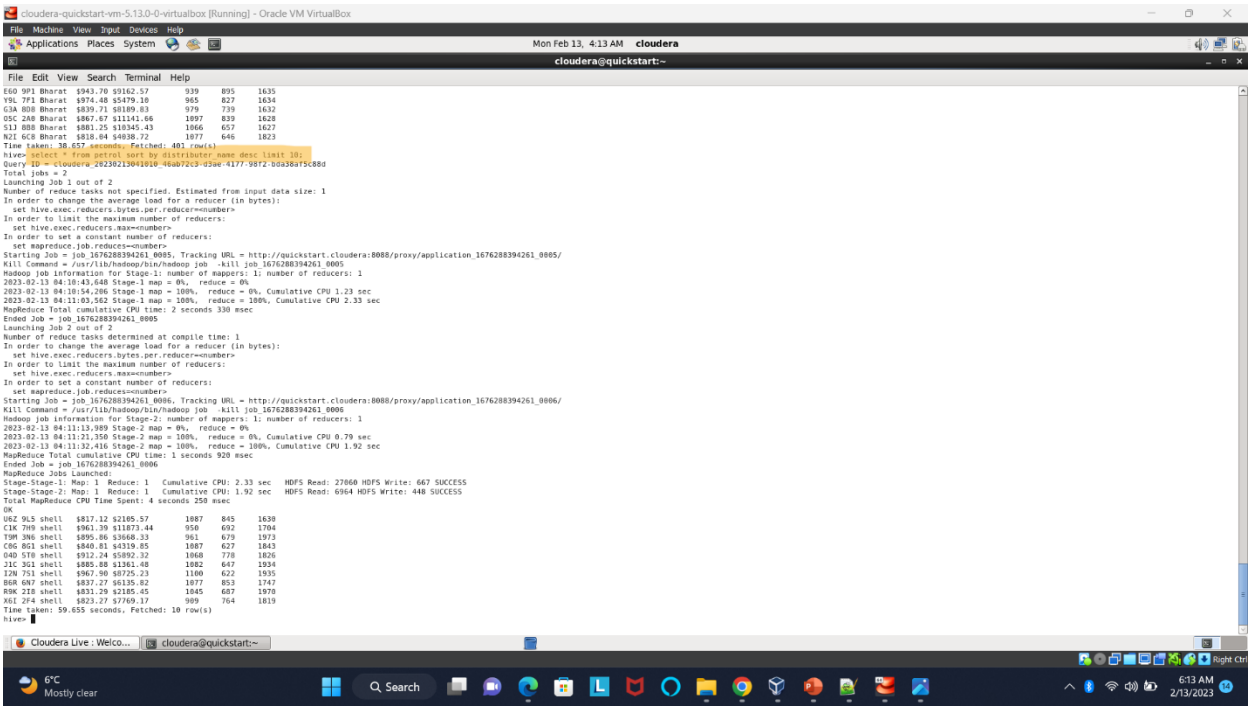


- Checking table data after loading data to the table using select command.



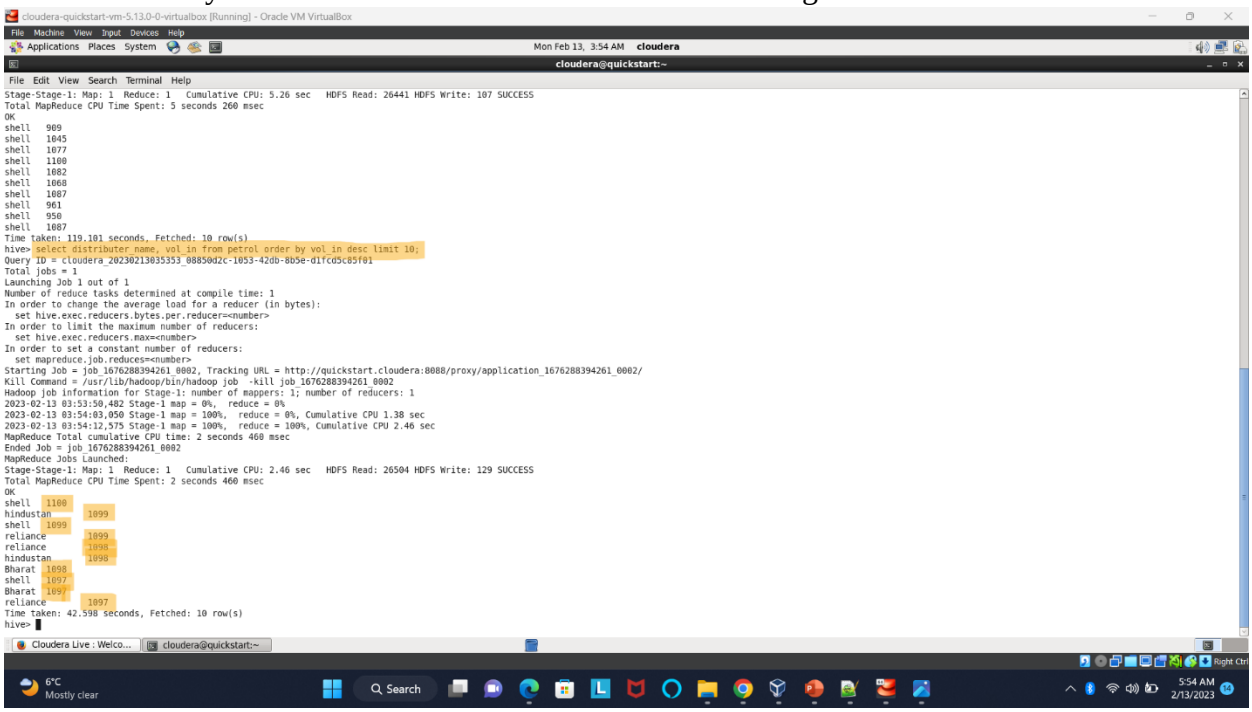
- Performing group by operation on petrol table data. The table data is grouped based on distributor name and the corresponding total sum of the volume out is displayed right next the distributer name.



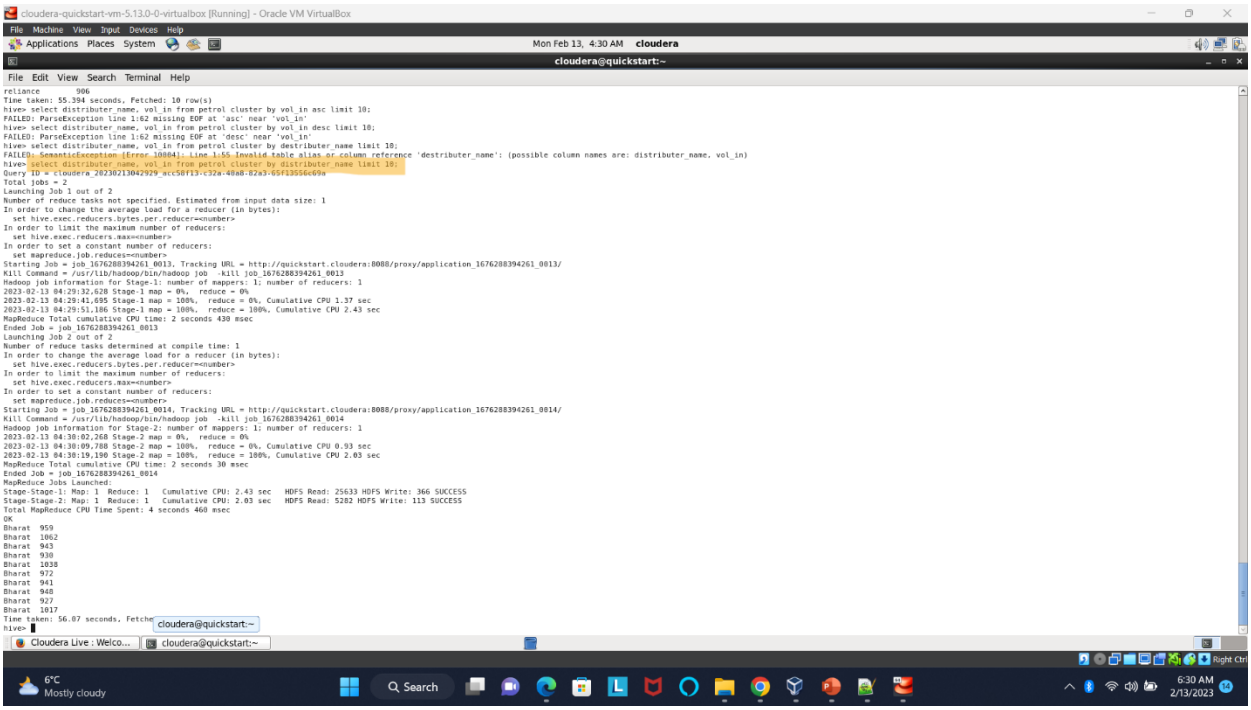


- In the below figure the order by operation is performed on the table data of petrol. Order by is used to sort the data in an order (ascending or descending order). This operation in hive performed after completing hadoop's map reduce operation.
- The data is sorted based on volume_in in descending order.

- The query used in below figure demonstrates about the use case of 'sort by'.
- 'Sort by' in hive is used to sort the data before sending data to the reducer.

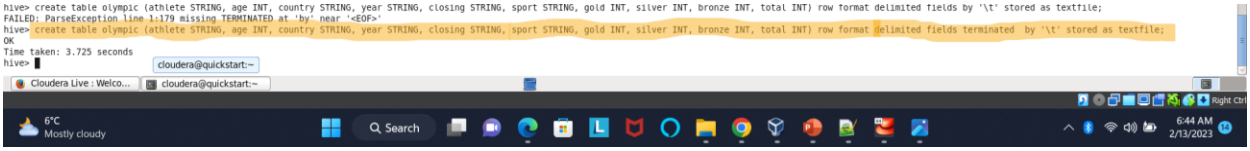


- The query used in the figure below demonstrates the use case of ‘cluster by’.
- ‘Cluster by’ in hive is used to cluster similar data in the column and sends that result to different reducers.



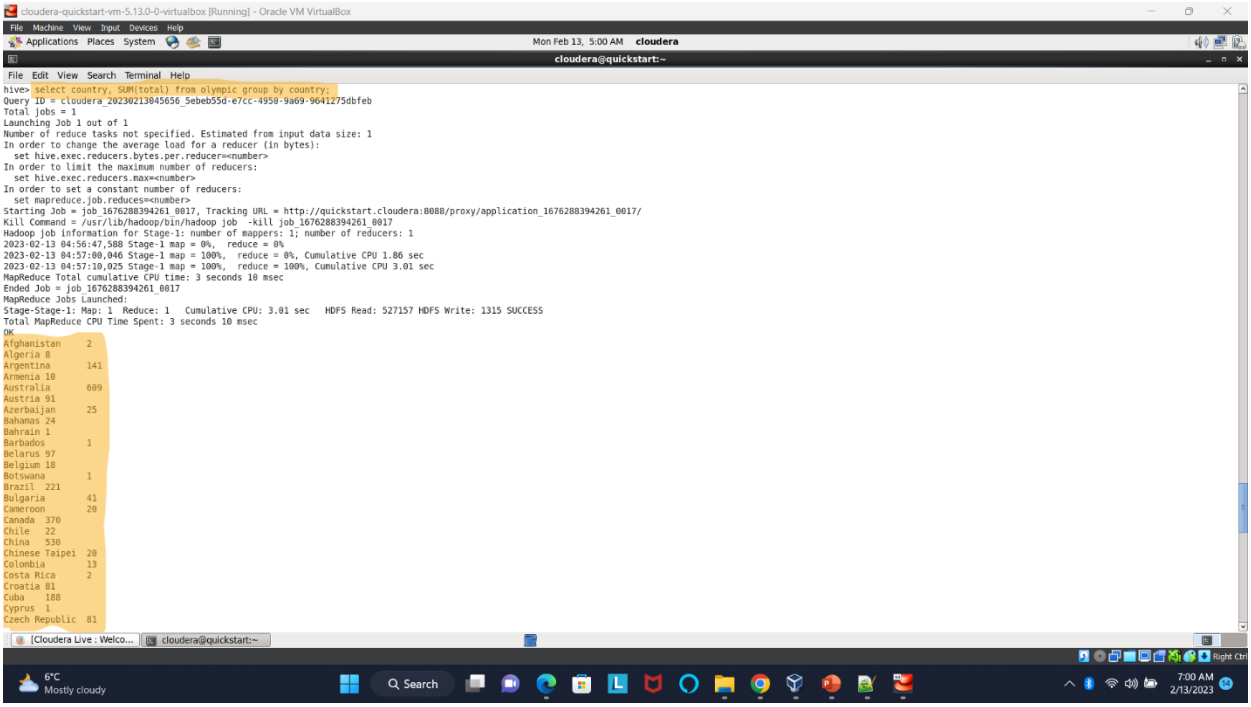
Task 2

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

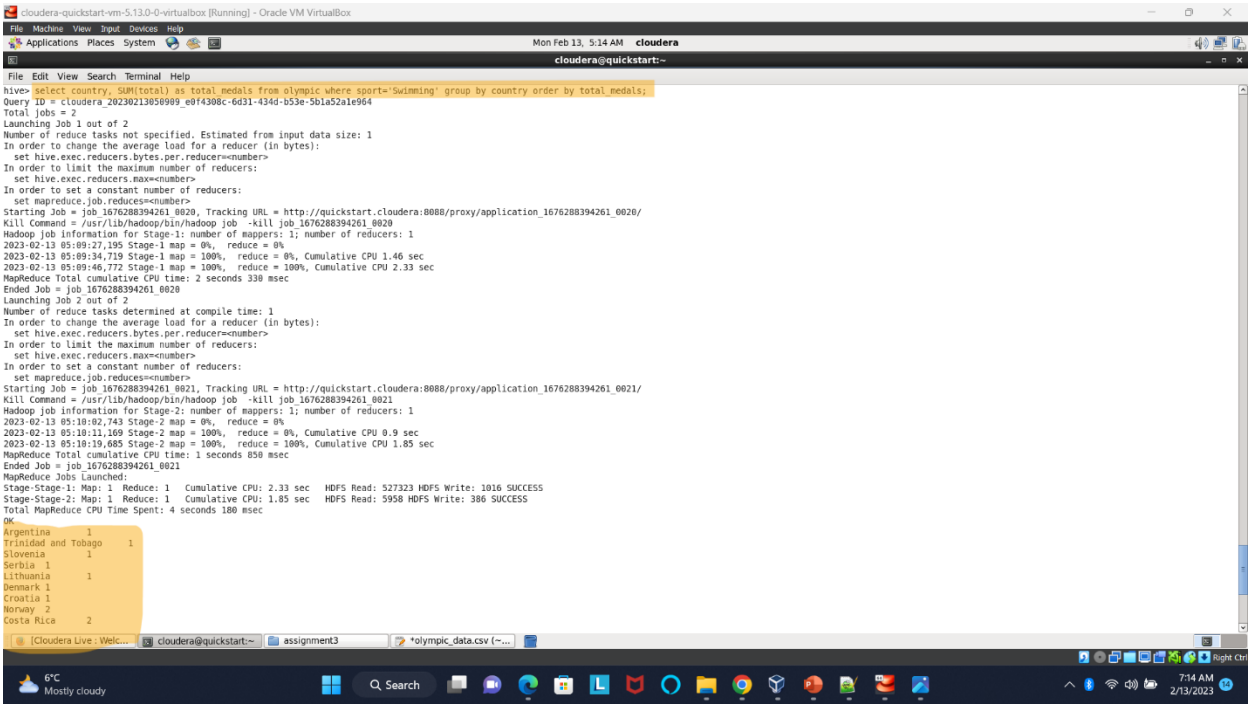


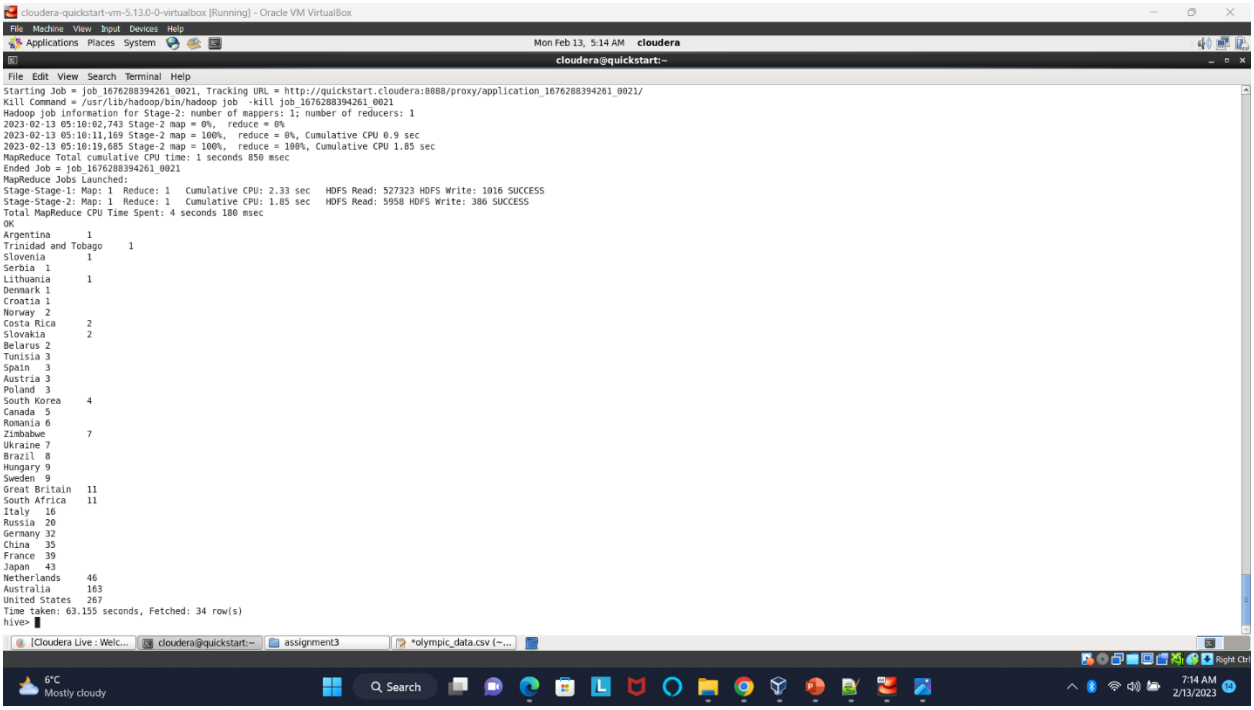
- Loaded Olympic data to Olympic table using load data command.

- 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.

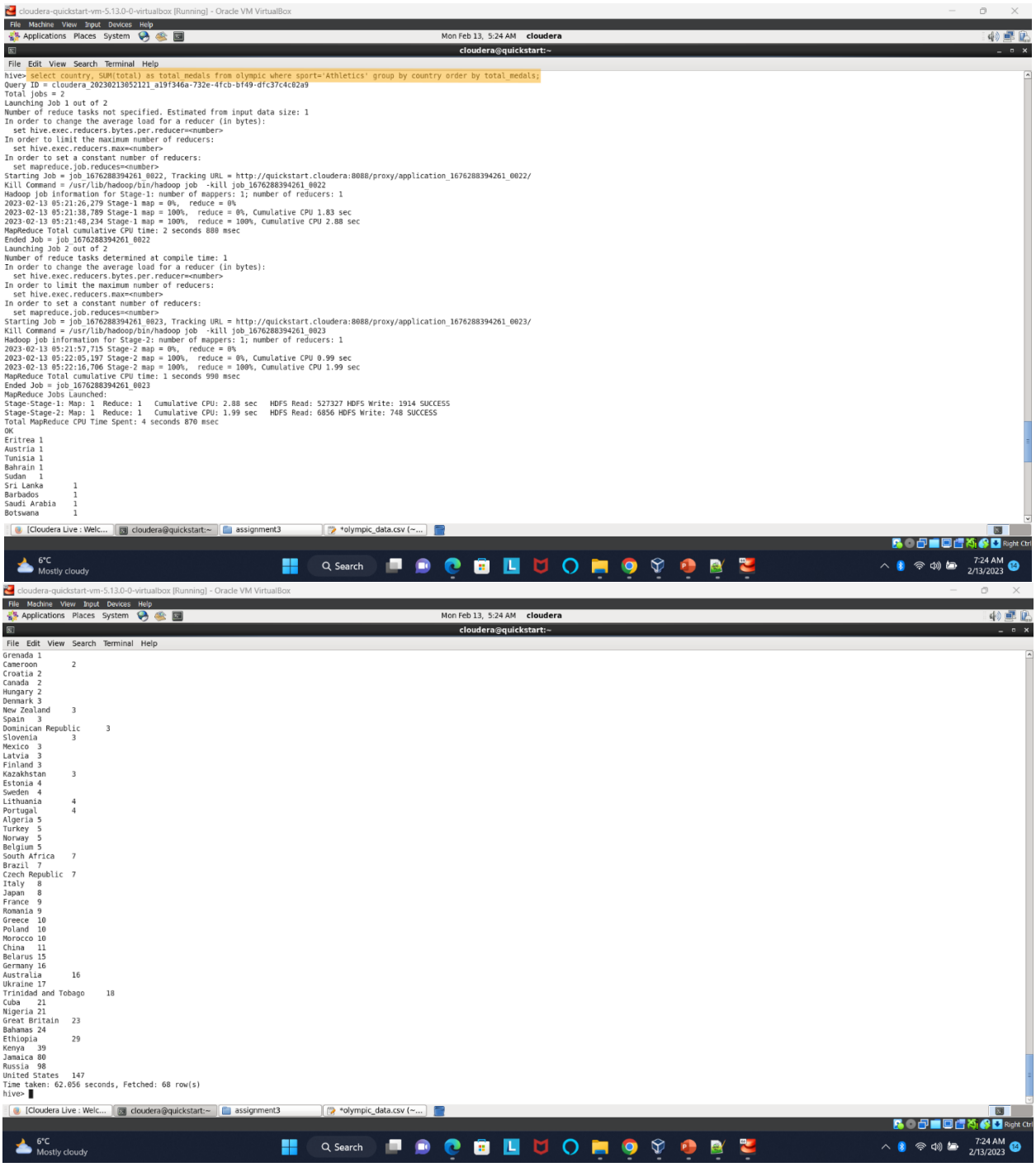


- The medals won for swimming by all the countries are found by getting all the records whose sport is ‘Swimming’ and the result is grouped based on country name. The sum of the medals won by country in swimming are aggregated and displayed beside the country name.

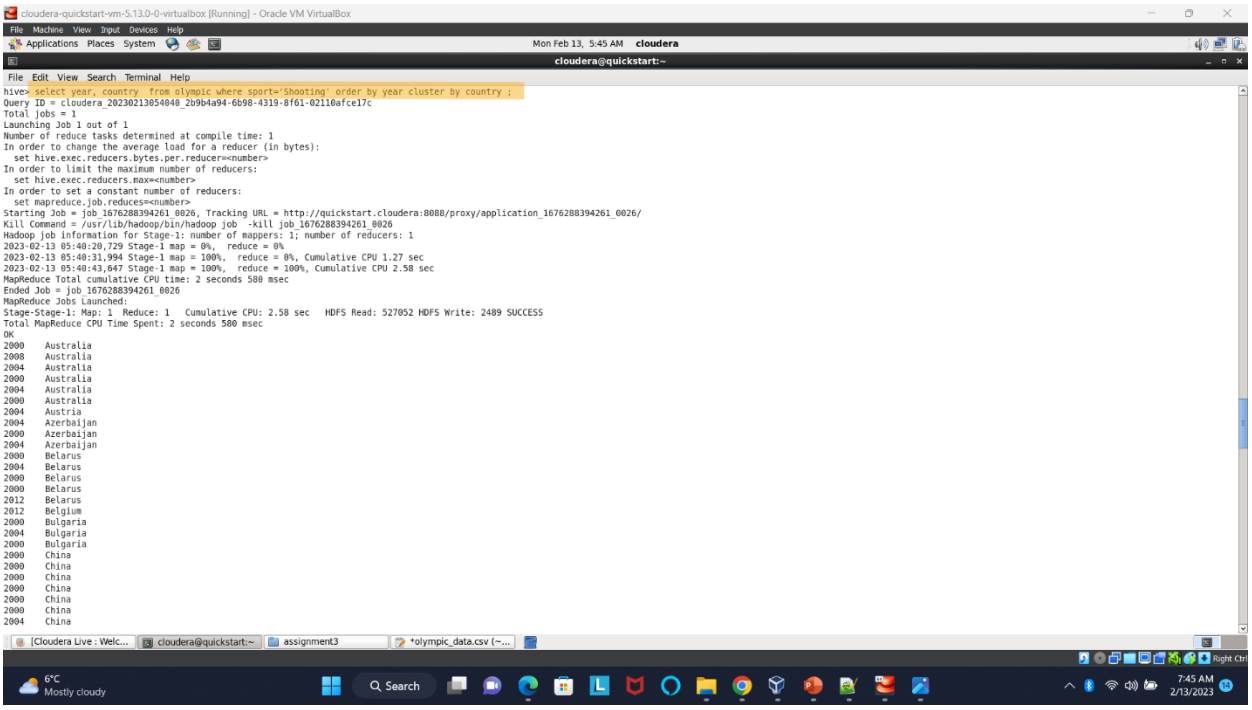


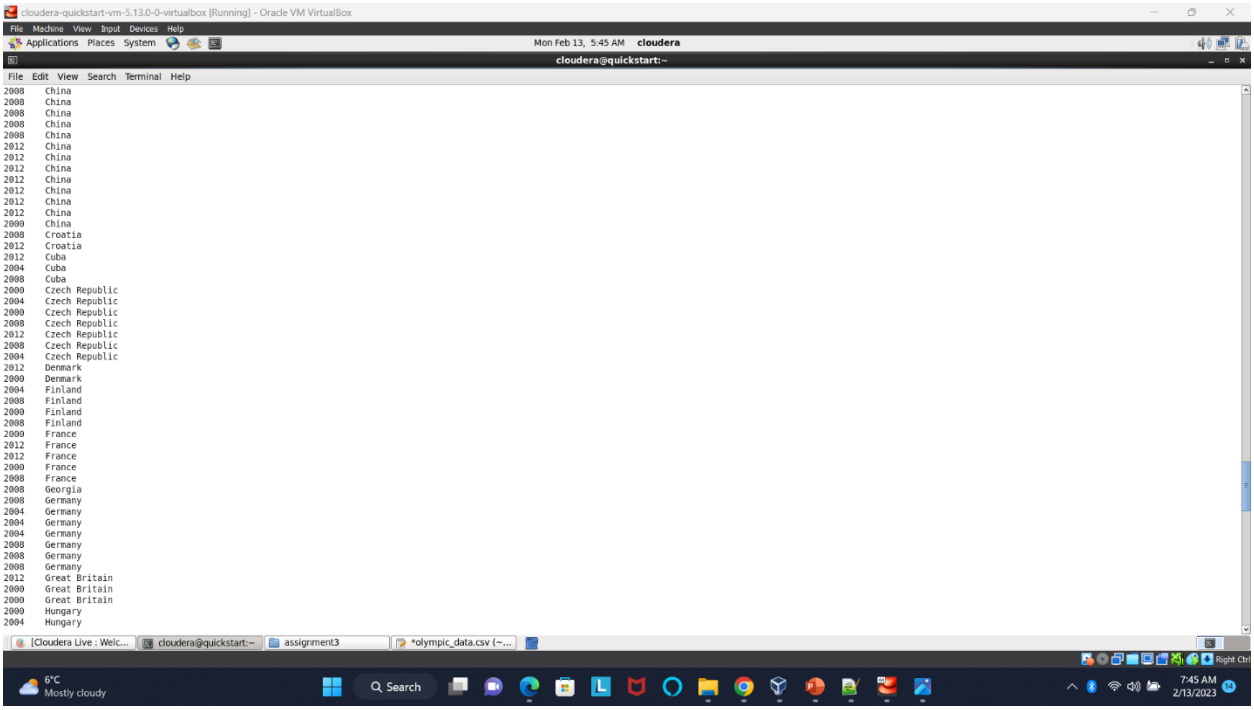


- Found all the medals that are won by countries for ‘Athletics’. All the countries are group using group by operation and the sum of the medals for athletics for each country are found using sum method and are displayed beside the country.



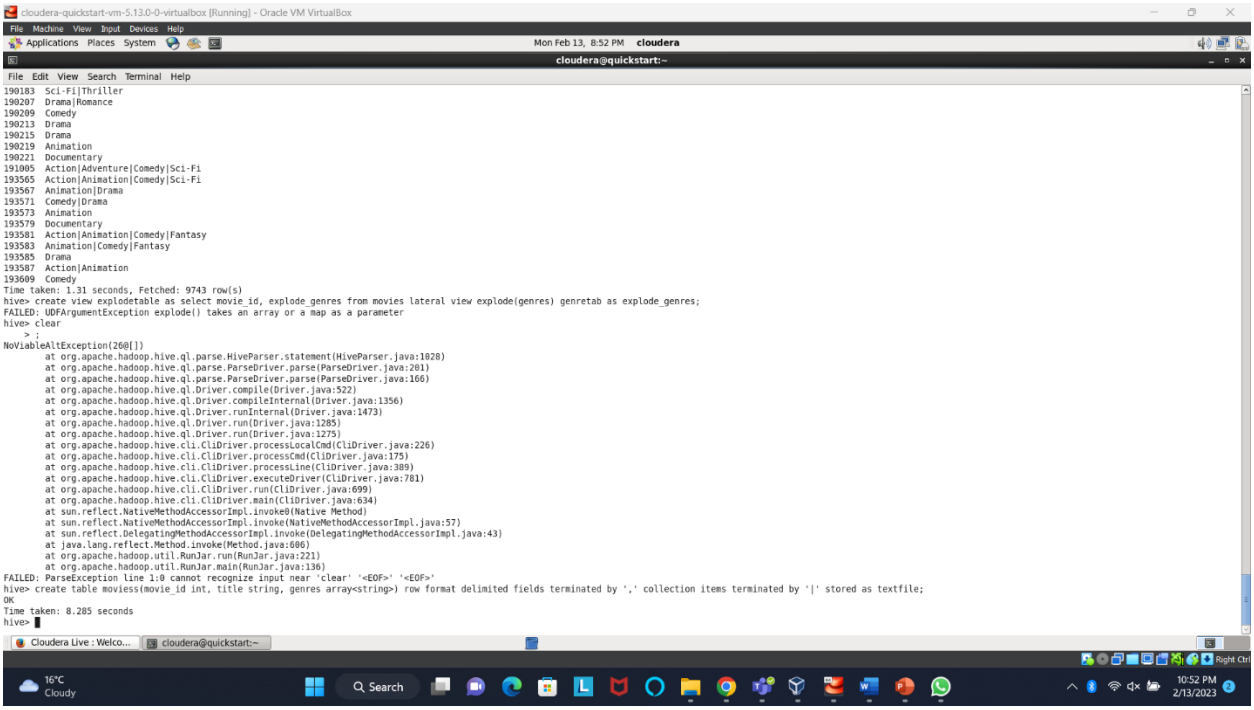
- Year wise classification of the countries who secured medals in shooting are extracted by clustering based on country name and sorted based on year.



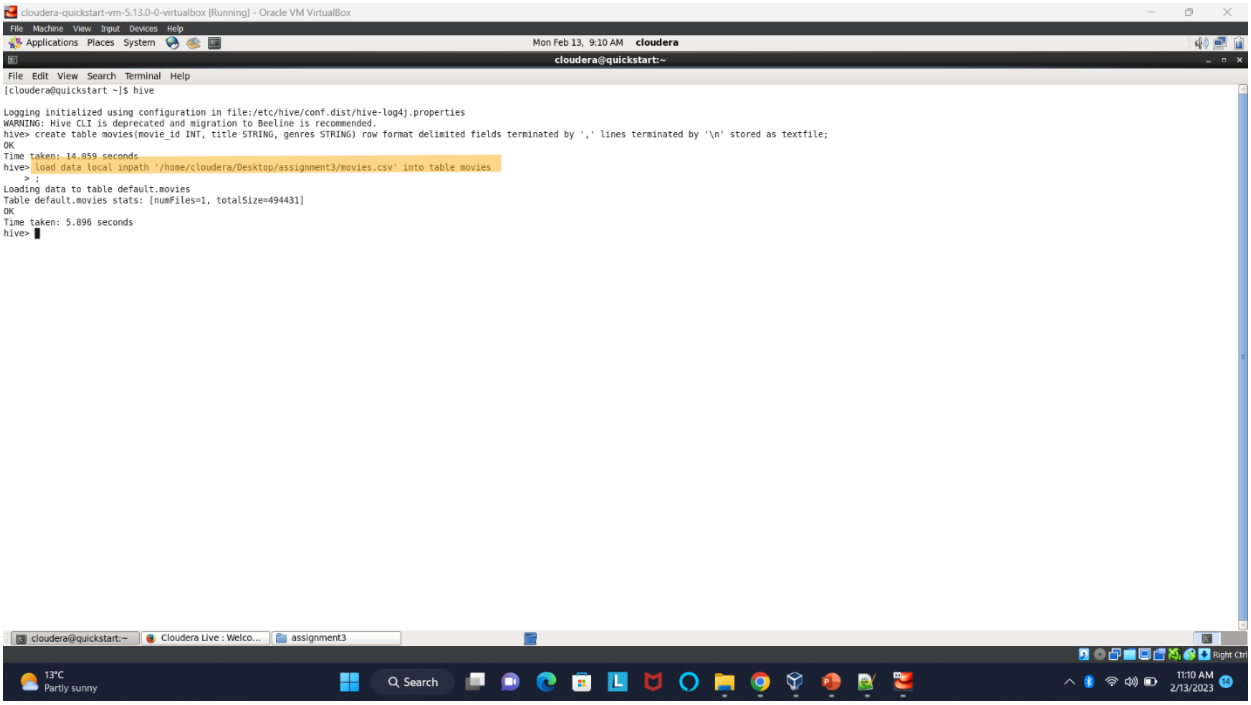


TASK 3

- Created movies table using create table query.



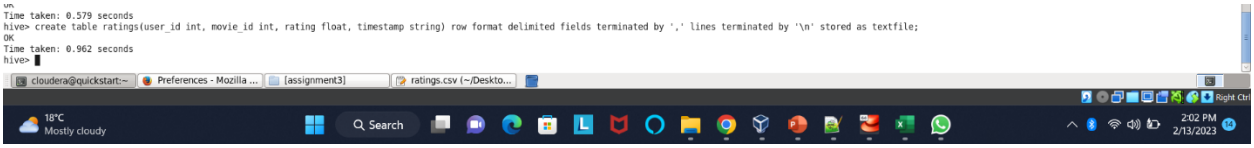
- Loaded data into movies table using load query.



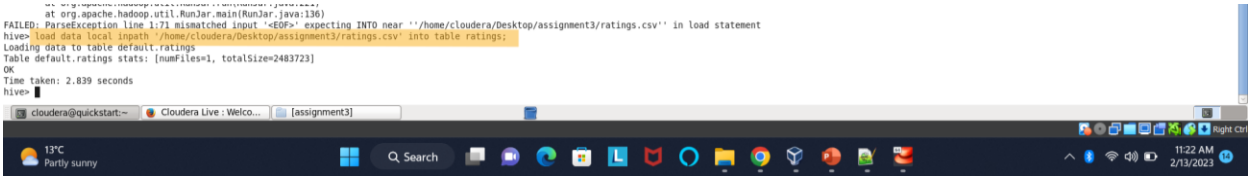
- Created Users table using create query for users.

- Loaded data into users table using load data query.

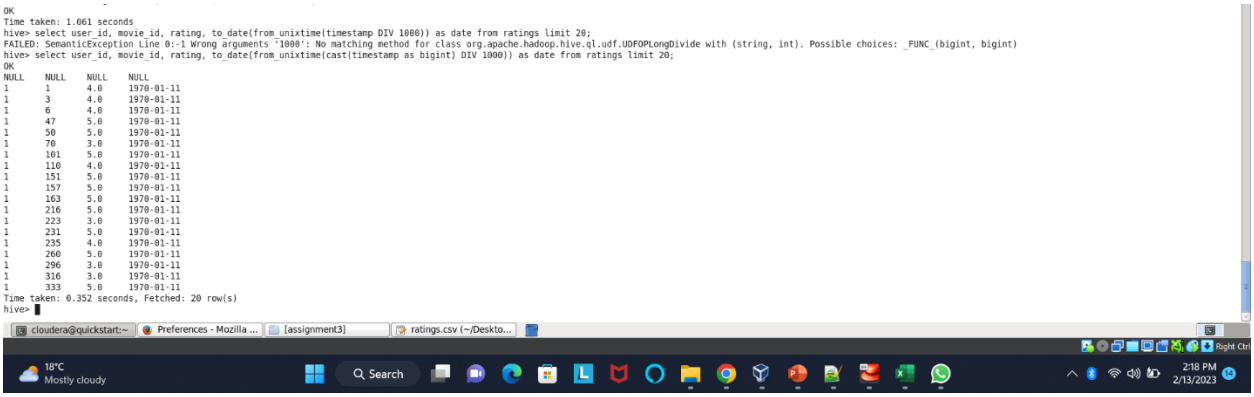
- Created ratings table using create table query.



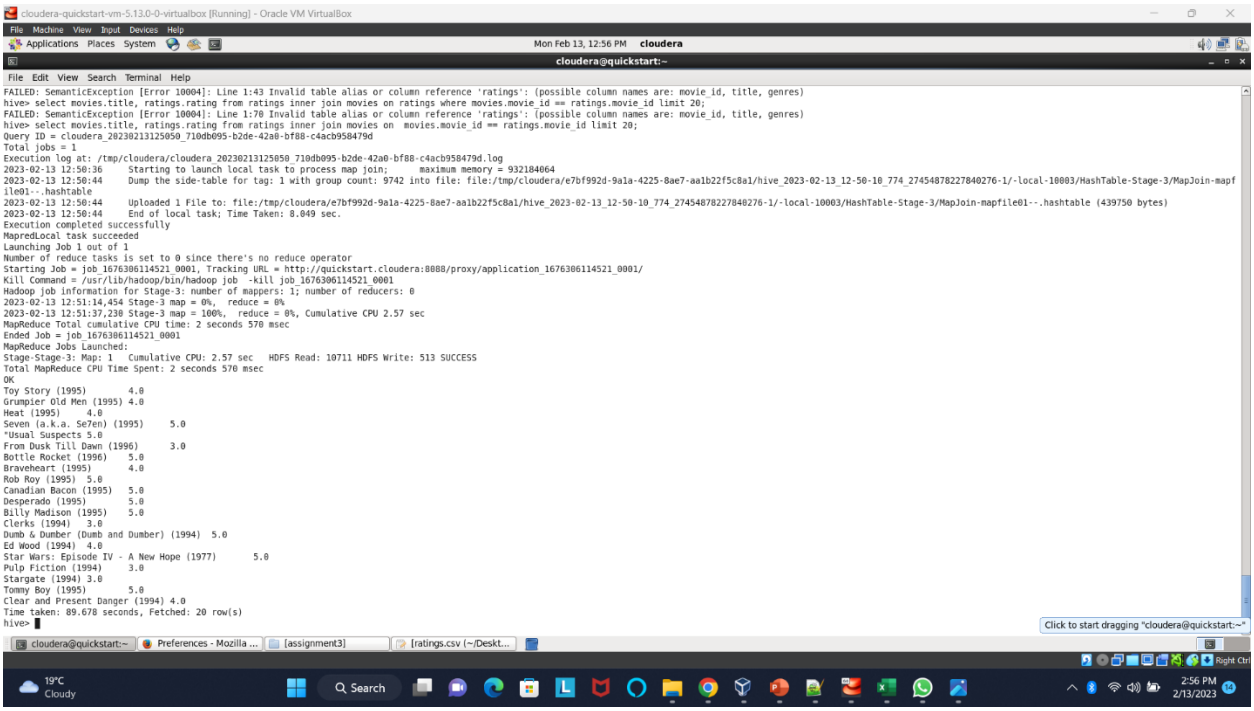
- Loaded data into ratings table using load data query.



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

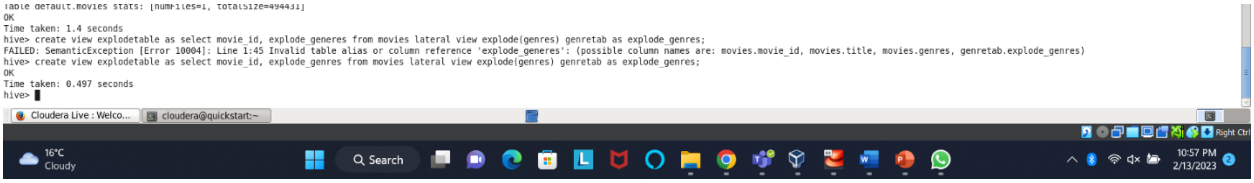


- The movie table is joined with ratings table using inner join based on movie id of ratings table and movie id in movies table. Based on movie id the title of the movie is taken from movies table and the ratings are taken from ratings table and are displayed right next to the movie title.

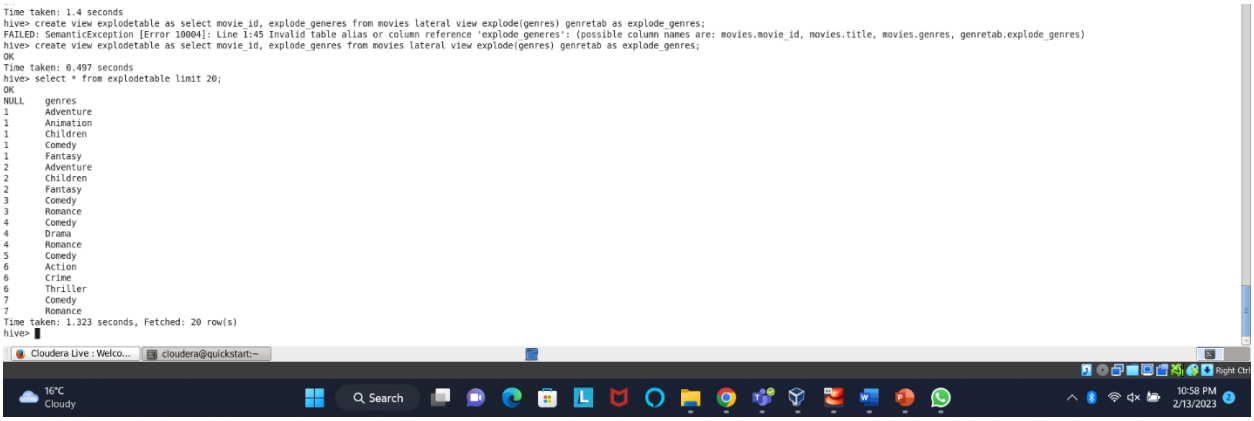


TASK 4

- Created explodetable table view for movie id and genere.

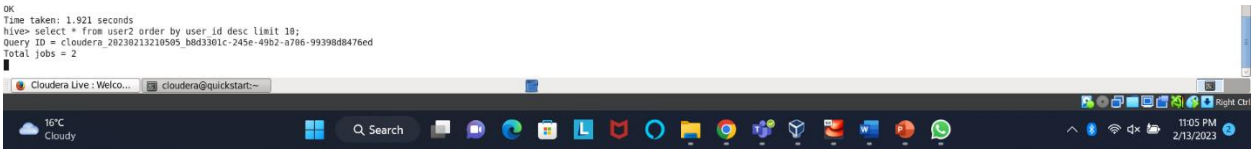


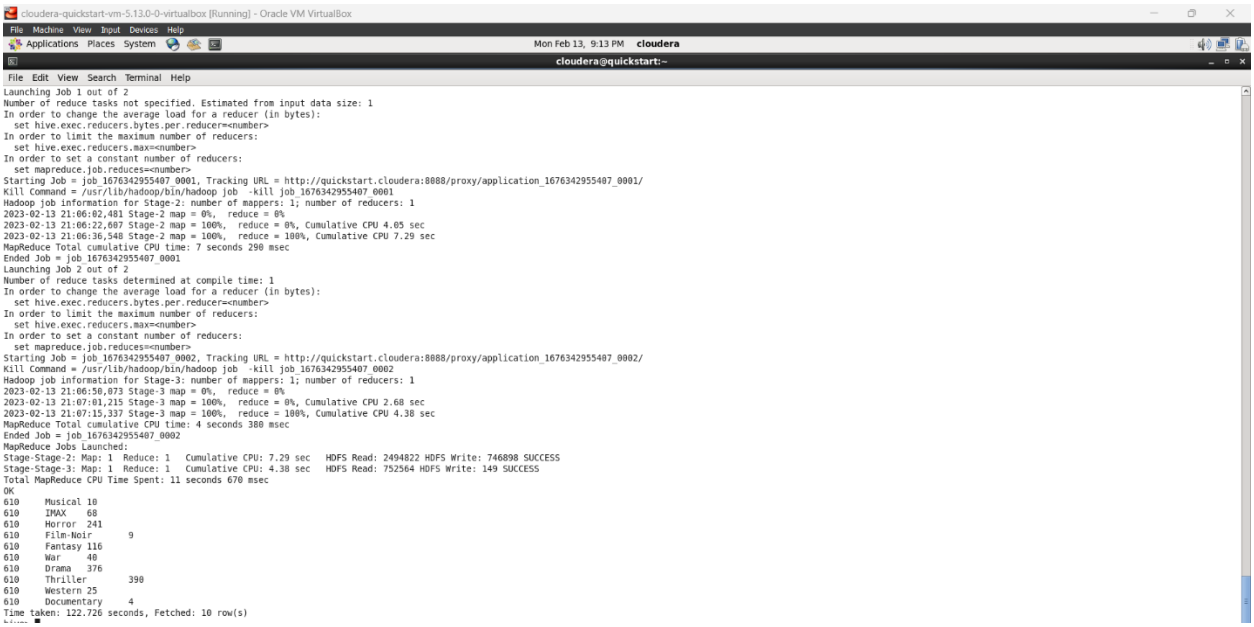
- Did select all for explodetable table to view movie genere and rating of the genere and viewed on the screen.



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

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





- 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 explodetable e join moviesrating r on e.movie_ID=r.movies_ID) t where rank<=3 limit 10;
Query ID = cloudera_20230213193232_db38edc9-1ed3-4785-9a43-01fe5c85f5b0
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20230213193232_db38edc9-1ed3-4785-9a43-01fe5c85f5b0.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_6547829233721896 21-1/-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>
```