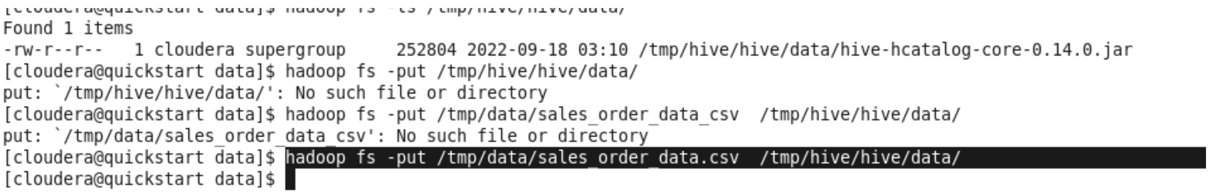
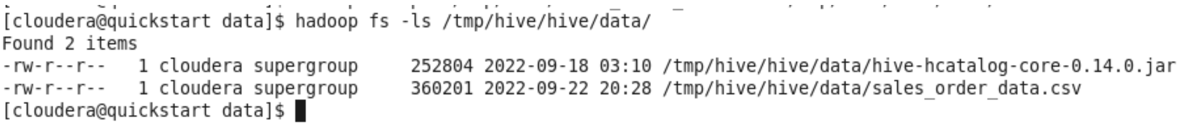
HIVE ASSIGNMENT -1

NAME: ABHIJIT BARIK

2. Store raw data into hdfs location

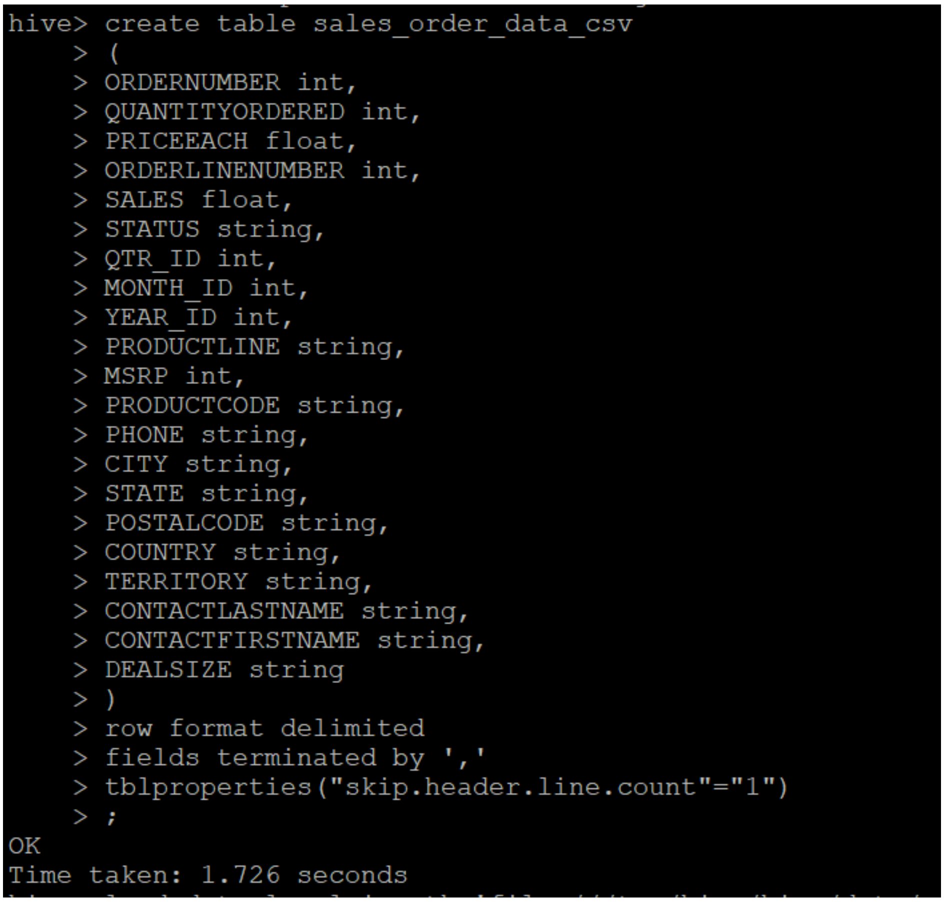


```
[cloudera@quickstart data]$ hadoop fs -ls /tmp/hive/hive/data/
Found 1 items
-rw-r--r--   1 cloudera supergroup     252804 2022-09-18 03:10 /tmp/hive/hive/data/hive-hcatalog-core-0.14.0.jar
[cloudera@quickstart data]$ hadoop fs -put /tmp/hive/hive/data/
put: `/tmp/hive/hive/data/': No such file or directory
[cloudera@quickstart data]$ hadoop fs -put /tmp/data/sales_order_data_csv  /tmp/hive/hive/data/
put: `/tmp/data/sales_order_data_csv': No such file or directory
[cloudera@quickstart data]$ hadoop fs -put /tmp/data/sales_order_data.csv  /tmp/hive/hive/data/
[cloudera@quickstart data]$ 
```



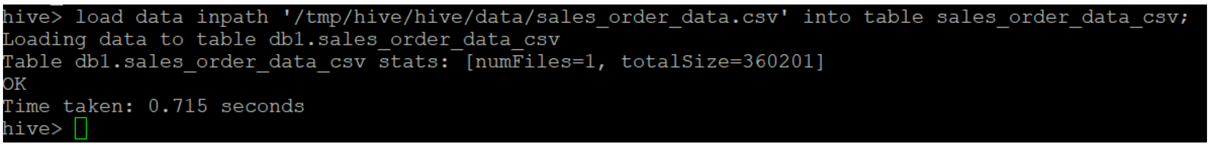
```
[cloudera@quickstart data]$ hadoop fs -ls /tmp/hive/hive/data/
Found 2 items
-rw-r--r--   1 cloudera supergroup     252804 2022-09-18 03:10 /tmp/hive/hive/data/hive-hcatalog-core-0.14.0.jar
-rw-r--r--   1 cloudera supergroup     360201 2022-09-22 20:28 /tmp/hive/hive/data/sales_order_data.csv
[cloudera@quickstart data]$ 
```

3. Create a internal hive table "sales_order_csv" which will store csv data sales_order_csv .. make sure to skip header row while creating table



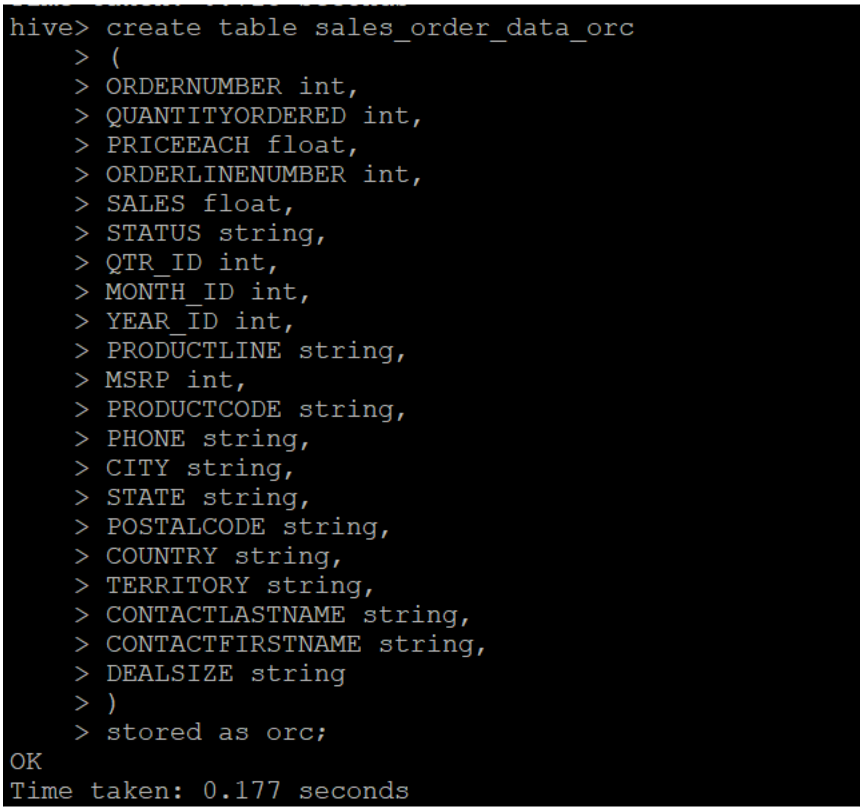
```
hive> create table sales_order_data_csv
    > (
    > ORDERNUMBER int,
    > QUANTITYORDERED int,
    > PRICEEACH float,
    > ORDERLINENUMBER int,
    > SALES float,
    > STATUS string,
    > QTR_ID int,
    > MONTH_ID int,
    > YEAR_ID int,
    > PRODUCTLINE string,
    > MSRP int,
    > PRODUCTCODE string,
    > PHONE string,
    > CITY string,
    > STATE string,
    > POSTALCODE string,
    > COUNTRY string,
    > TERRITORY string,
    > CONTACTLASTNAME string,
    > CONTACTFIRSTNAME string,
    > DEALSIZE string
    > )
    > row format delimited
    > fields terminated by ','
    > tblproperties("skip.header.line.count"="1")
    > ;
OK
Time taken: 1.726 seconds
```

4. Load data from hdfs path into "sales_order_csv"



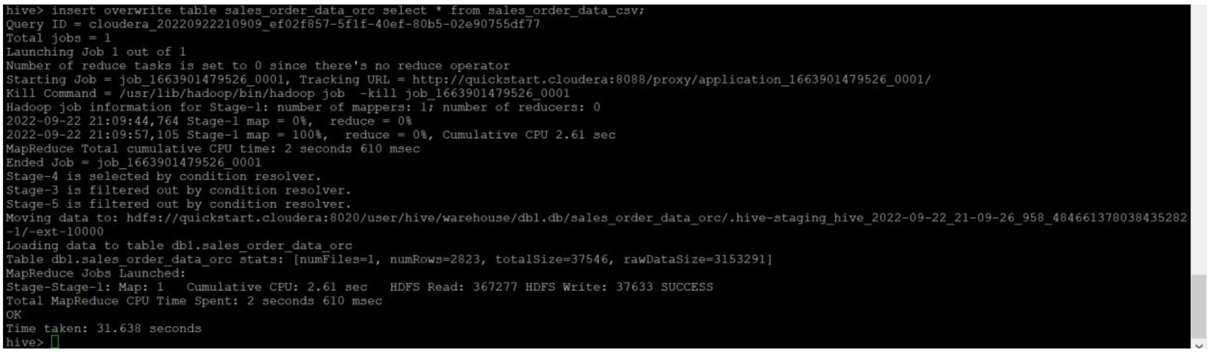
```
hive> load data inpath '/tmp/hive/hive/data/sales_order_data.csv' into table sales_order_data_csv;
Loading data to table db1.sales_order_data_csv
Table db1.sales_order_data_csv stats: [numFiles=1, totalSize=360201]
OK
Time taken: 0.715 seconds
hive> 
```

5. Create an internal hive table which will store data in ORC format "sales_order_orc"



```
hive> create table sales_order_data_orc
    > (
    > ORDERNUMBER int,
    > QUANTITYORDERED int,
    > PRICEEACH float,
    > ORDERLINENUMBER int,
    > SALES float,
    > STATUS string,
    > QTR_ID int,
    > MONTH_ID int,
    > YEAR_ID int,
    > PRODUCTLINE string,
    > MSRP int,
    > PRODUCTCODE string,
    > PHONE string,
    > CITY string,
    > STATE string,
    > POSTALCODE string,
    > COUNTRY string,
    > TERRITORY string,
    > CONTACTLASTNAME string,
    > CONTACTFIRSTNAME string,
    > DEALSIZE string
    > )
    > stored as orc;
OK
Time taken: 0.177 seconds
```

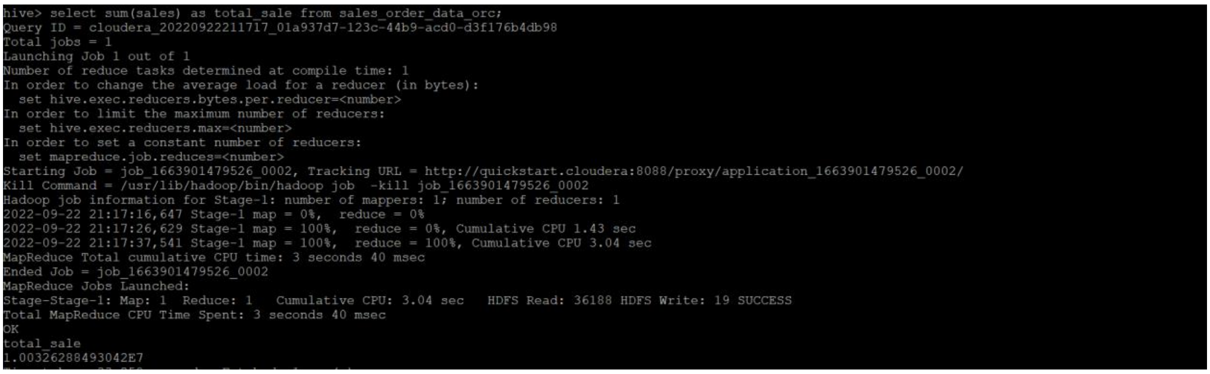
6. Load data from "sales_order_csv" into "sales_order_orc"



```
hive> insert overwrite table sales_order_data_orc select * from sales_order_data_csv;
Query ID = cloudera_20220922210909_ef02f857-5f1f-40ef-80b5-02e90755df77
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1663901479526_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663901479526_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663901479526_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-09-22 21:09:44,764 Stage-1 map = 0%,  reduce = 0%
2022-09-22 21:09:57,105 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.61 sec
MapReduce Total cumulative CPU time: 2 seconds 610 msec
Ended Job = job_1663901479526_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/db1.db/sales_order_data_orc/.hive-staging_hive_2022-09-22_21-09-26_958_484661378038435282
-1/-ext-10000
Loading data to table db1.sales_order_data_orc
Table db1.sales_order_data_orc stats: [numFiles=1, numRows=2823, totalSize=37546, rawDataSize=3153291]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.61 sec   HDFS Read: 367277 HDFS Write: 37633 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 610 msec
OK
Time taken: 31.638 seconds
hive> 
```

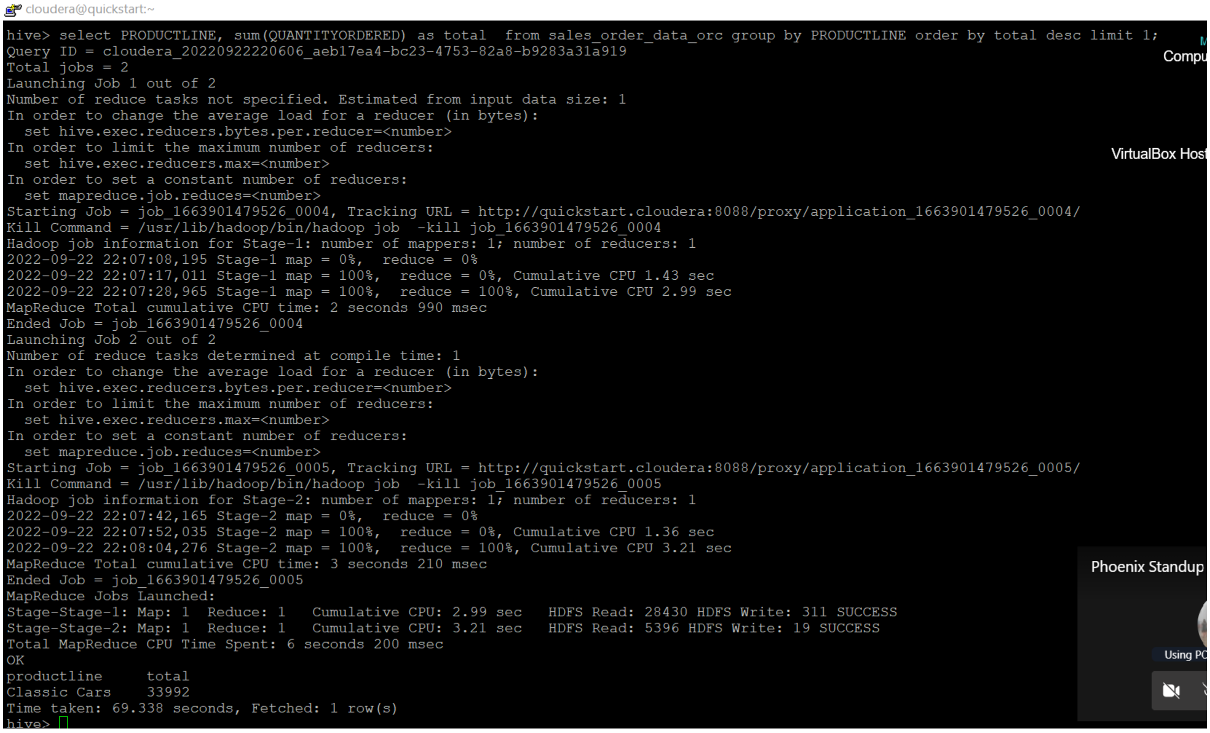
Perform below menioned queries on "sales_order_orc" table :

a. Calculatye total sales per year



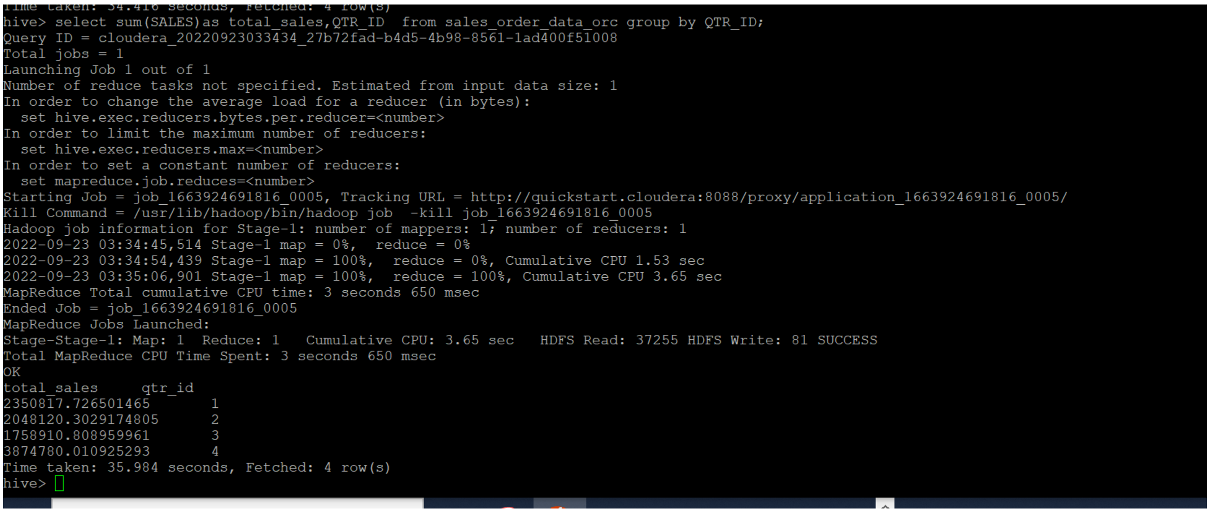
```
hive> select sum(sales) as total_sale from sales_order_data_orc;
Query ID = cloudera_20220922211717_01a937d7-123c-44b9-acd0-d3f176b4db98
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set 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_1663901479526_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663901479526_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663901479526_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-22 21:17:16,647 Stage-1 map = 0%,  reduce = 0%
2022-09-22 21:17:26,629 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.43 sec
2022-09-22 21:17:37,541 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.04 sec
MapReduce Total cumulative CPU time: 3 seconds 40 msec
Ended Job = job_1663901479526_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.04 sec   HDFS Read: 36188 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 40 msec
OK
total_sale
1.00326288493042E7
```

b. Find a product for which maximum orders were placed



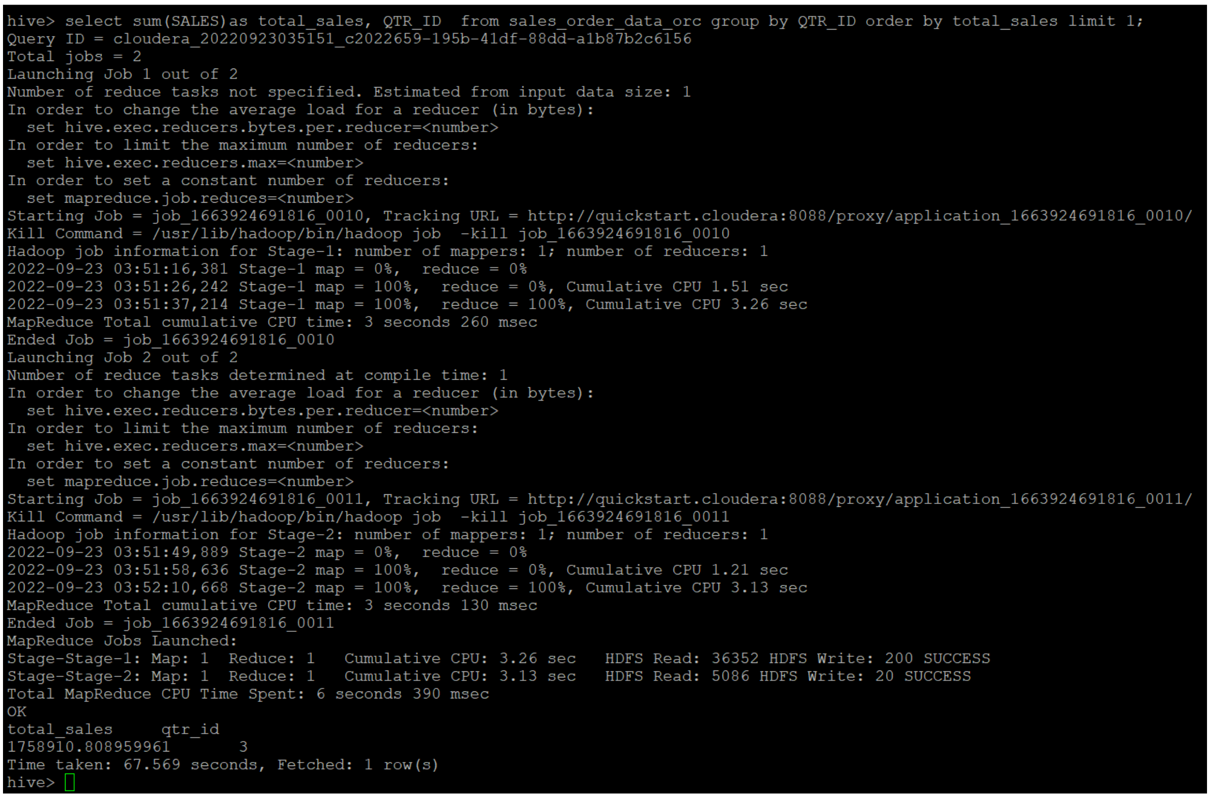
```
hive> select PRODUCTLINE, sum(QUANTITYORDERED) as total  from sales_order_data_orc group by PRODUCTLINE order by total desc limit 1;
Query ID = cloudera_20220922220606_aeb17ea4-bc23-4753-82a8-b9283a31a919
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from 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_1663901479526_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663901479526_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663901479526_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-22 22:07:08,195 Stage-1 map = 0%,  reduce = 0%
2022-09-22 22:07:17,011 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.43 sec
2022-09-22 22:07:28,965 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.99 sec
MapReduce Total cumulative CPU time: 2 seconds 990 msec
Ended Job = job_1663901479526_0004
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_1663901479526_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663901479526_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663901479526_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-09-22 22:07:42,165 Stage-2 map = 0%,  reduce = 0%
2022-09-22 22:07:52,035 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.36 sec
2022-09-22 22:08:04,276 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.21 sec
MapReduce Total cumulative CPU time: 3 seconds 210 msec
Ended Job = job_1663901479526_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.99 sec   HDFS Read: 28430 HDFS Write: 311 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.21 sec   HDFS Read: 5396 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 200 msec
OK
productline     total
Classic Cars    33992
Time taken: 69.338 seconds, Fetched: 1 row(s)
hive>
```

c. Calculate the total sales for each quarter



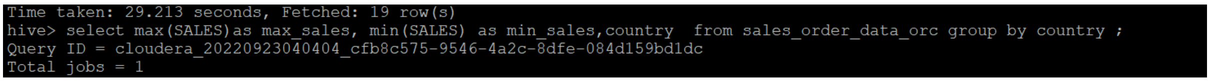
```
Time taken: 34.416 seconds, Fetched: 4 row(s)
hive> select sum(SALES)as total_sales,QTR_ID  from sales_order_data_orc group by QTR_ID;
Query ID = cloudera_20220923033434_27b72fad-b4d5-4b98-8561-1ad400f51008
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from 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_1663924691816_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663924691816_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663924691816_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-23 03:34:45,514 Stage-1 map = 0%,  reduce = 0%
2022-09-23 03:34:54,439 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.53 sec
2022-09-23 03:35:06,901 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.65 sec
MapReduce Total cumulative CPU time: 3 seconds 650 msec
Ended Job = job_1663924691816_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.65 sec   HDFS Read: 37255 HDFS Write: 81 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 650 msec
OK
total_sales     qtr_id
2350817.726501465       1
2048120.3029174805      2
1758910.808959961       3
3874780.010925293       4
Time taken: 35.984 seconds, Fetched: 4 row(s)
hive>
```

d. In which quarter sales was minimum

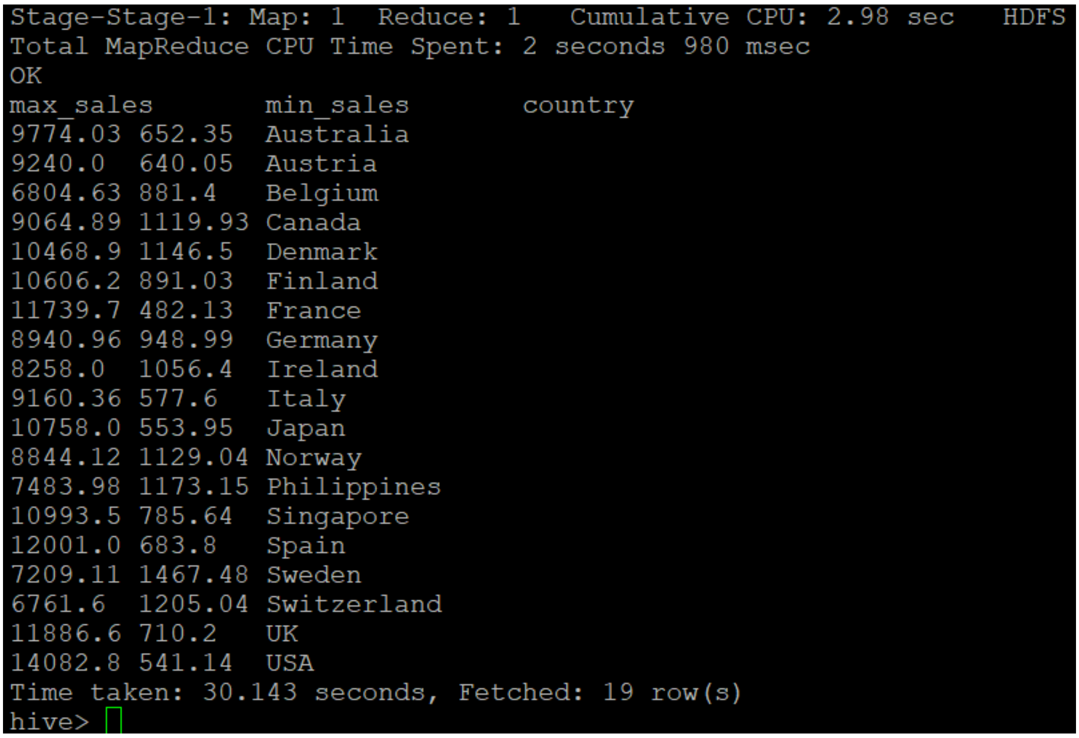


```
hive> select sum(SALES)as total_sales, QTR_ID  from sales_order_data_orc group by QTR_ID order by total_sales limit 1;
Query ID = cloudera_20220923035151_c2022659-195b-41df-88dd-a1b87b2c6156
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1663924691816_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663924691816_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663924691816_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-23 03:51:16,381 Stage-1 map = 0%,  reduce = 0%
2022-09-23 03:51:26,242 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.51 sec
2022-09-23 03:51:37,214 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.26 sec
MapReduce Total cumulative CPU time: 3 seconds 260 msec
Ended Job = job_1663924691816_0010
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1663924691816_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663924691816_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663924691816_0011
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-09-23 03:51:49,889 Stage-2 map = 0%,  reduce = 0%
2022-09-23 03:51:58,636 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.21 sec
2022-09-23 03:52:10,668 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.13 sec
MapReduce Total cumulative CPU time: 3 seconds 130 msec
Ended Job = job_1663924691816_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.26 sec   HDFS Read: 36352 HDFS Write: 200 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.13 sec   HDFS Read: 5086 HDFS Write: 20 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 390 msec
OK
total_sales     qtr_id
1758910.808959961       3
Time taken: 67.569 seconds, Fetched: 1 row(s)
hive> 
```

e. In which country sales was maximum and in which country sales was minimum

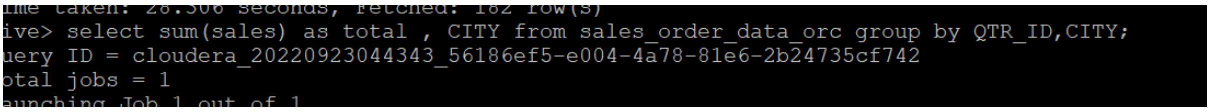


```
Time taken: 29.213 seconds, Fetched: 19 row(s)
hive> select max(SALES)as max_sales, min(SALES) as min_sales,country  from sales_order_data_orc group by country ;
Query ID = cloudera_20220923040404_cfb8c575-9546-4a2c-8dfe-084d159bd1dc
Total jobs = 1
```

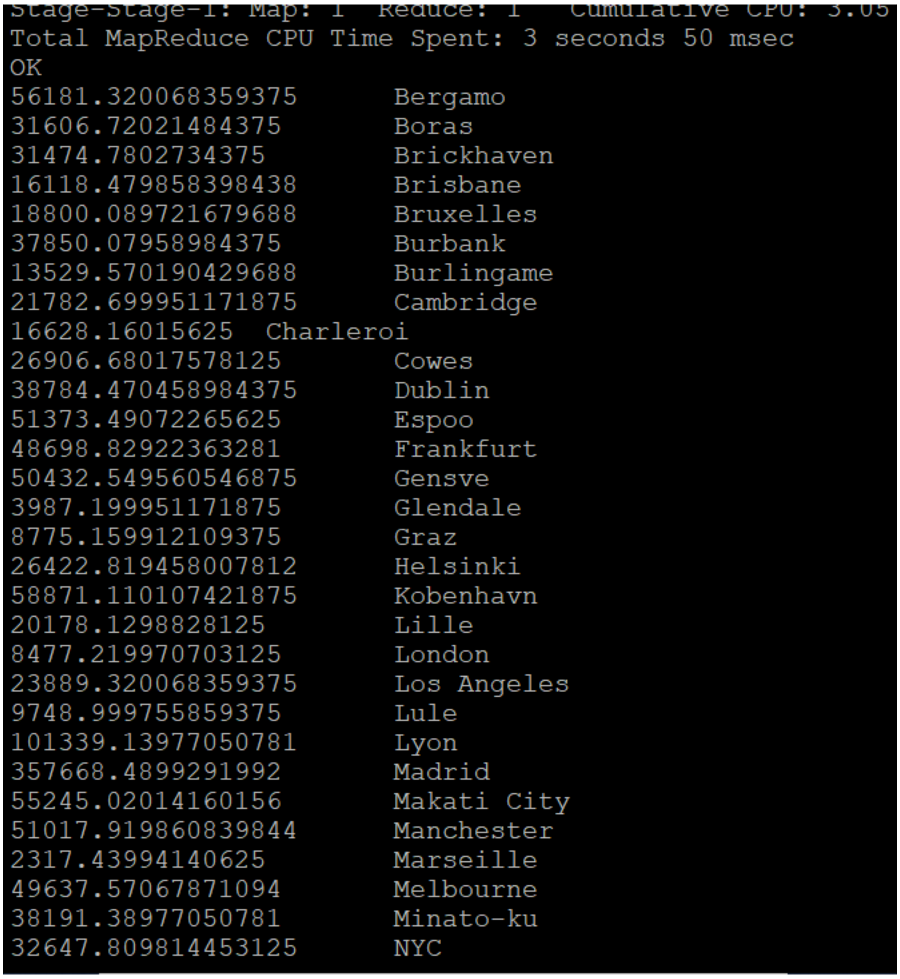


```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.98 sec   HDFS
Total MapReduce CPU Time Spent: 2 seconds 980 msec
OK
max_sales       min_sales       country
9774.03 652.35  Australia
9240.0  640.05  Austria
6804.63 881.4   Belgium
9064.89 1119.93 Canada
10468.9 1146.5  Denmark
10606.2 891.03  Finland
11739.7 482.13  France
8940.96 948.99  Germany
8258.0  1056.4  Ireland
9160.36 577.6   Italy
10758.0 553.95  Japan
8844.12 1129.04 Norway
7483.98 1173.15 Philippines
10993.5 785.64  Singapore
12001.0 683.8   Spain
7209.11 1467.48 Sweden
6761.6  1205.04 Switzerland
11886.6 710.2   UK
14082.8 541.14  USA
Time taken: 30.143 seconds, Fetched: 19 row(s)
hive> 
```

f. Calculate quartelry sales for each city

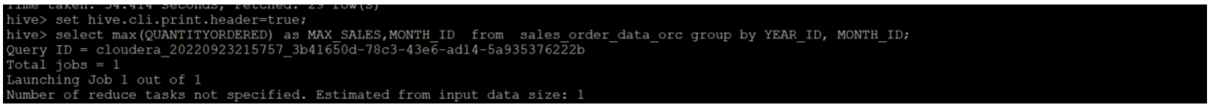


```
ime taken: 28.306 seconds, Fetched: 182 row(s)
ive> select sum(sales) as total , CITY from sales_order_data_orc group by QTR_ID,CITY;
uery ID = cloudera_20220923044343_56186ef5-e004-4a78-81e6-2b24735cf742
otal jobs = 1
aunching Job 1 out of 1
```

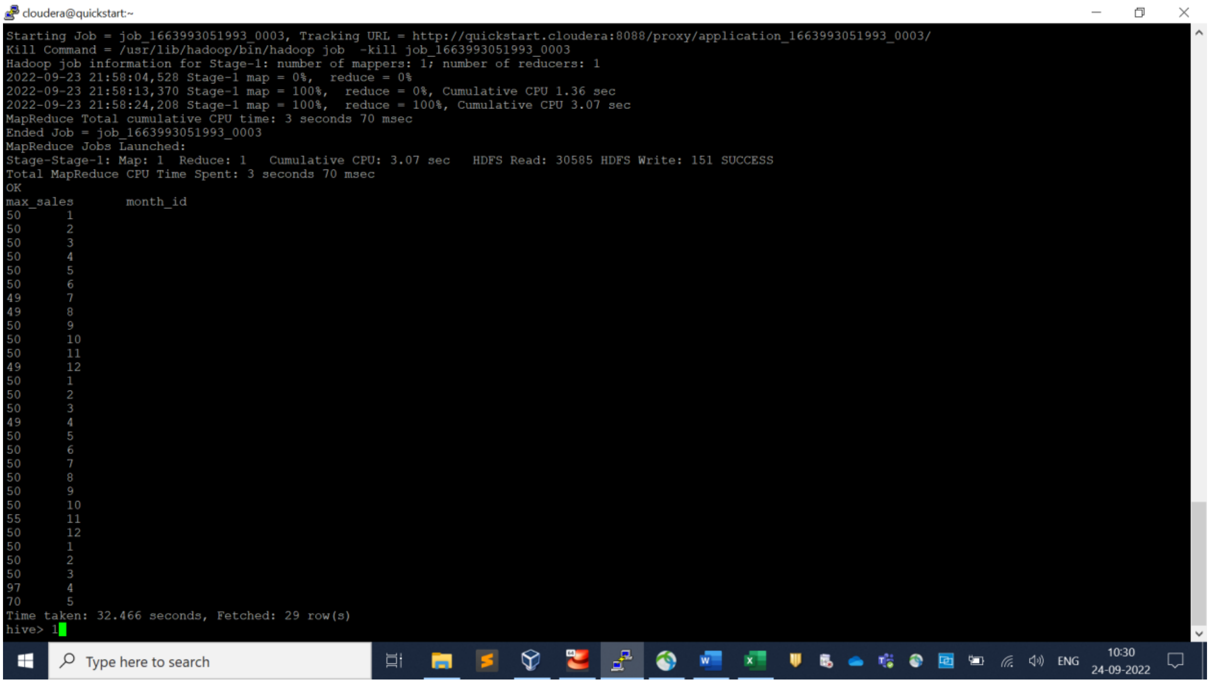


```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.05
Total MapReduce CPU Time Spent: 3 seconds 50 msec
OK
56181.320068359375      Bergamo
31606.72021484375       Boras
31474.7802734375        Brickhaven
16118.479858398438      Brisbane
18800.089721679688      Bruxelles
37850.07958984375       Burbank
13529.570190429688      Burlingame
21782.699951171875      Cambridge
16628.16015625  Charleroi
26906.68017578125       Cowes
38784.470458984375      Dublin
51373.49072265625       Espoo
48698.82922363281       Frankfurt
50432.549560546875      Gensve
3987.199951171875       Glendale
8775.159912109375       Graz
26422.819458007812      Helsinki
58871.110107421875      Kobenhavn
20178.1298828125        Lille
8477.219970703125       London
23889.320068359375      Los Angeles
9748.999755859375       Lule
101339.13977050781      Lyon
357668.4899291992       Madrid
55245.02014160156       Makati City
51017.919860839844      Manchester
2317.43994140625        Marseille
49637.57067871094       Melbourne
38191.38977050781       Minato-ku
32647.809814453125      NYC
```

h. Find a month for each year in which maximum number of quantities were sold



```
Time taken: 34.414 seconds, Fetched: 29 row(s)
hive> set hive.cli.print.header=true;
hive> select max(QUANTITYORDERED) as MAX_SALES,MONTH_ID  from  sales_order_data_orc group by YEAR_ID, MONTH_ID;
Query ID = cloudera_20220923215757_3b41650d-78c3-43e6-ad14-5a935376222b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```



```
Starting Job = job_1663993051993_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663993051993_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663993051993_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-23 21:58:04,528 Stage-1 map = 0%,  reduce = 0%
2022-09-23 21:58:13,370 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.36 sec
2022-09-23 21:58:24,208 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.07 sec
MapReduce Total cumulative CPU time: 3 seconds 70 msec
Ended Job = job_1663993051993_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.07 sec   HDFS Read: 30585 HDFS Write: 151 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 70 msec
OK
max_sales       month_id
50      1
50      2
50      3
50      4
50      5
50      6
49      7
49      8
50      9
50      10
50      11
49      12
50      1
50      2
50      3
49      4
50      5
50      6
50      7
50      8
50      9
50      10
55      11
50      12
50      1
50      2
50      3
97      4
70      5
Time taken: 32.466 seconds, Fetched: 29 row(s)
hive> 1
```