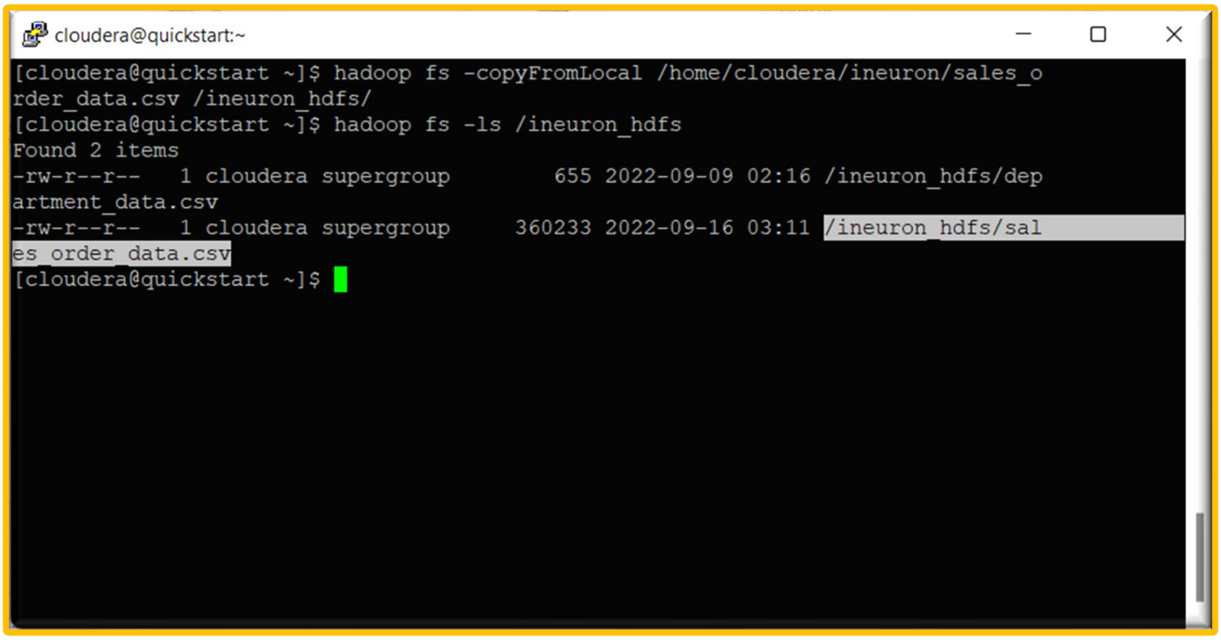
Hive Assignment 1 solutions

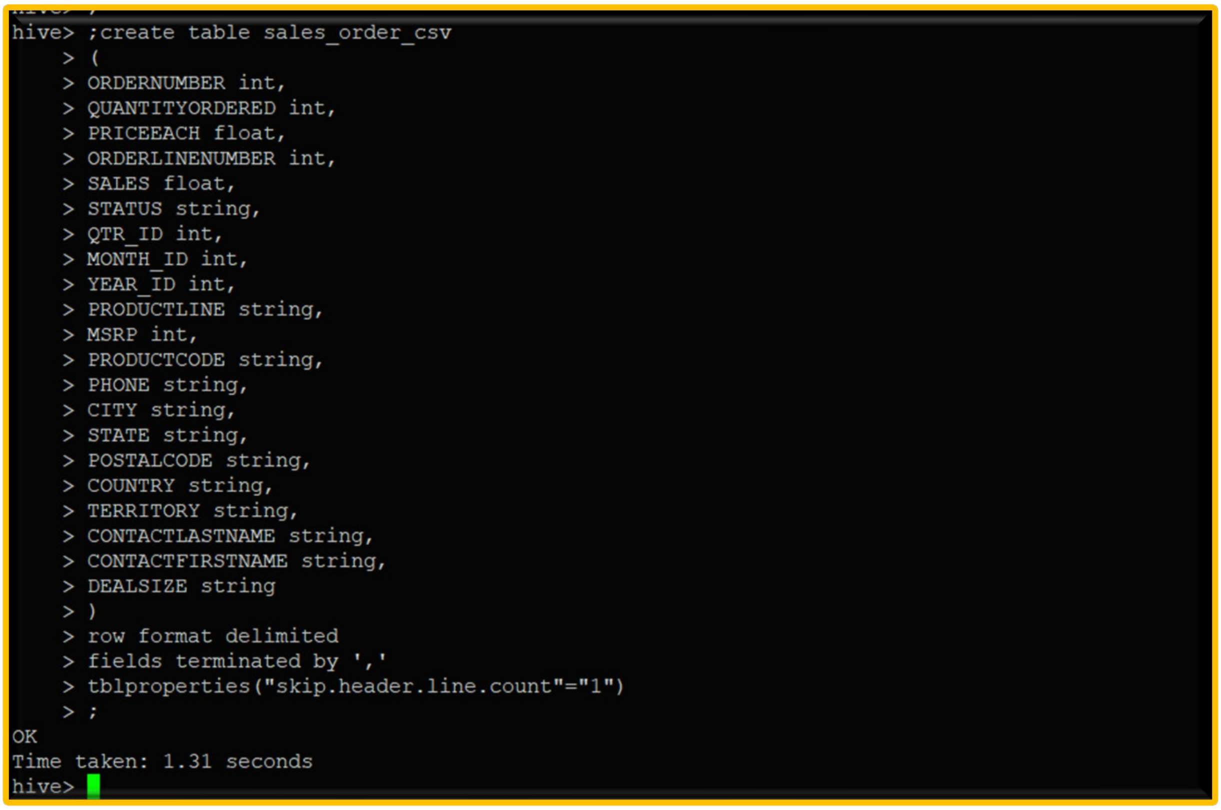
Creating hive orc table and loading data into it:

1. Storing raw data into hdfs location



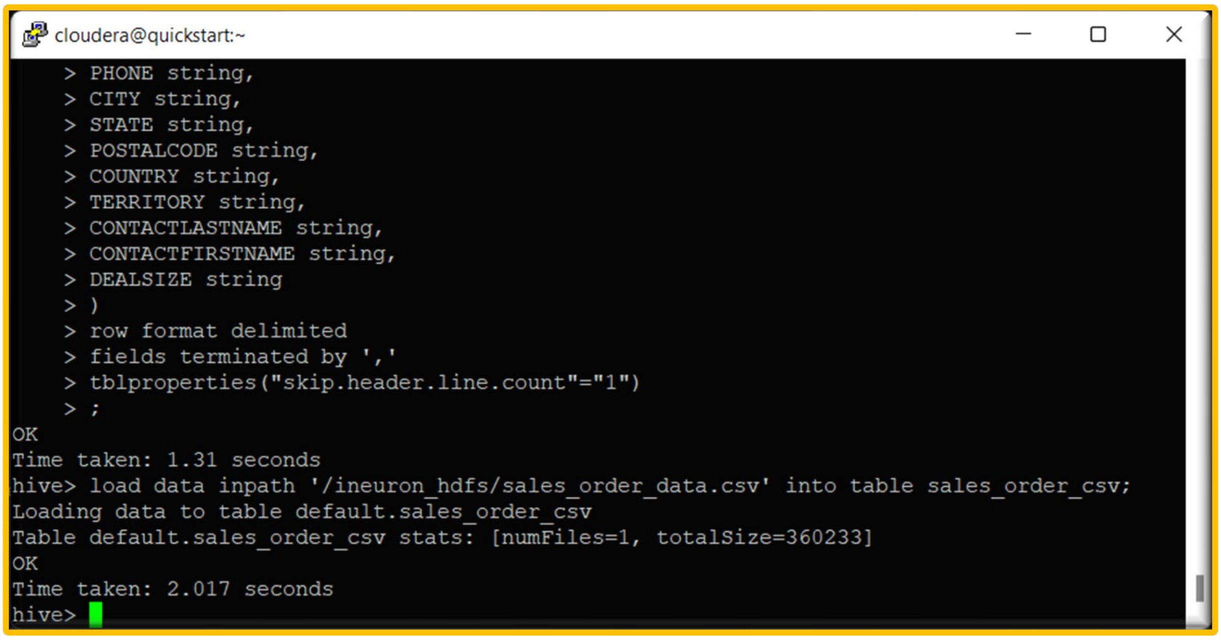
```
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/ineuron/sales_o
rder_data.csv /ineuron_hdfs/
[cloudera@quickstart ~]$ hadoop fs -ls /ineuron_hdfs
Found 2 items
-rw-r--r--   1 cloudera supergroup        655 2022-09-09 02:16 /ineuron_hdfs/dep
artment_data.csv
-rw-r--r--   1 cloudera supergroup     360233 2022-09-16 03:11 /ineuron_hdfs/sal
es_order_data.csv
[cloudera@quickstart ~]$
```

2.Creating internal hive table "sales_order_csv" which will store sales_order_data.csv



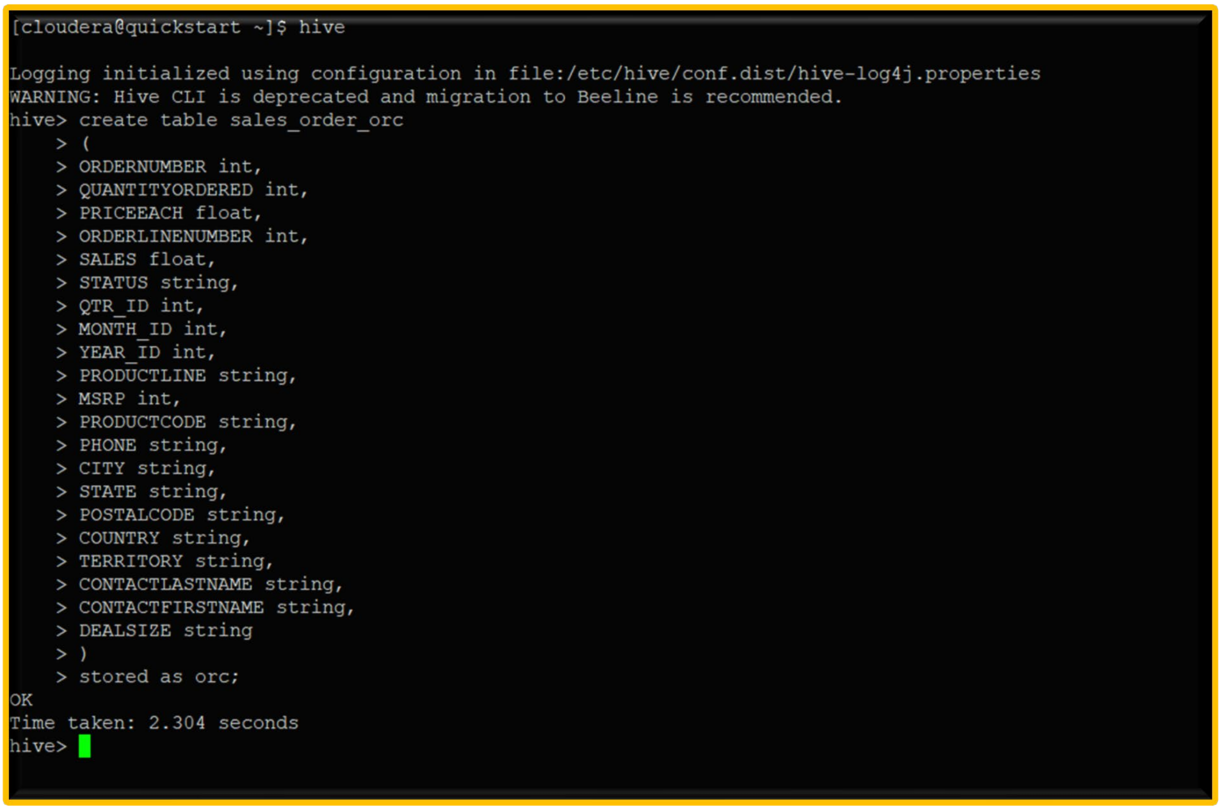
```
hive> ;
hive> ;create table sales_order_csv
    > (
    > ORDERNUMBER int,
    > QUANTITYORDERED int,
    > PRICEEACH float,
    > ORDERLINENUMBER int,
    > SALES float,
    > STATUS string,
    > QTR_ID int,
    > MONTH_ID int,
    > YEAR_ID int,
    > PRODUCTLINE string,
    > MSRP int,
    > 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: 1.31 seconds
hive>
```

3. Loading data from sales_order_data.csv which is in hdfs into "sales_order_csv" table.



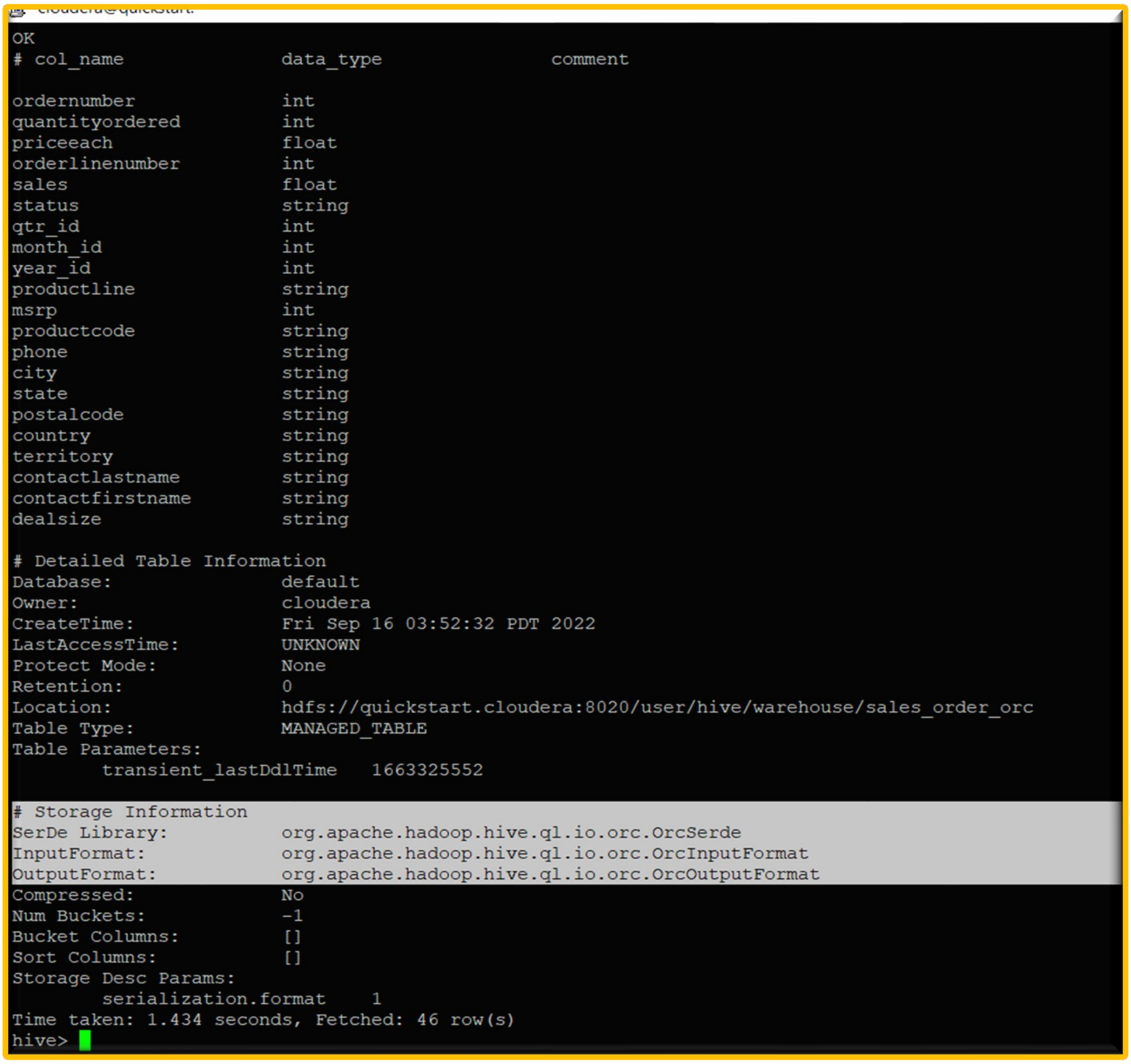
```
    > PHONE string,
    > CITY string,
    > STATE string,
    > POSTALCODE string,
    > COUNTRY string,
    > TERRITORY string,
    > CONTACTLASTNAME string,
    > CONTACTFIRSTNAME string,
    > DEALSIZE string
    > )
    > row format delimited
    > fields terminated by ','
    > tblproperties("skip.header.line.count"="1")
    > ;
OK
Time taken: 1.31 seconds
hive> load data inpath '/ineuron_hdfs/sales_order_data.csv' into table sales_order_csv;
Loading data to table default.sales_order_csv
Table default.sales_order_csv stats: [numFiles=1, totalSize=360233]
OK
Time taken: 2.017 seconds
hive>
```

4. Creating an internal hive table "sales_order_orc" which will store data in ORC format.



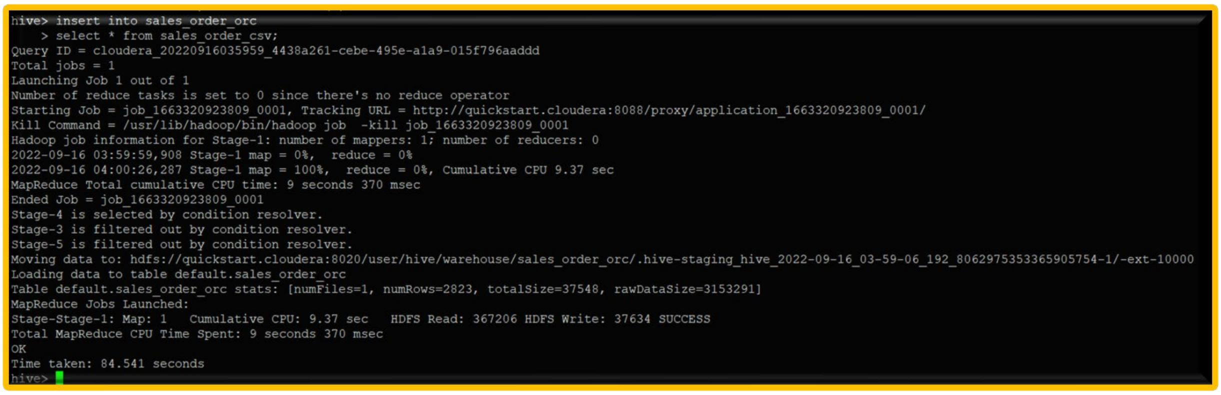
```
[cloudera@quickstart ~]$ hive

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 sales_order_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: 2.304 seconds
hive>
```

5. Checking the table format



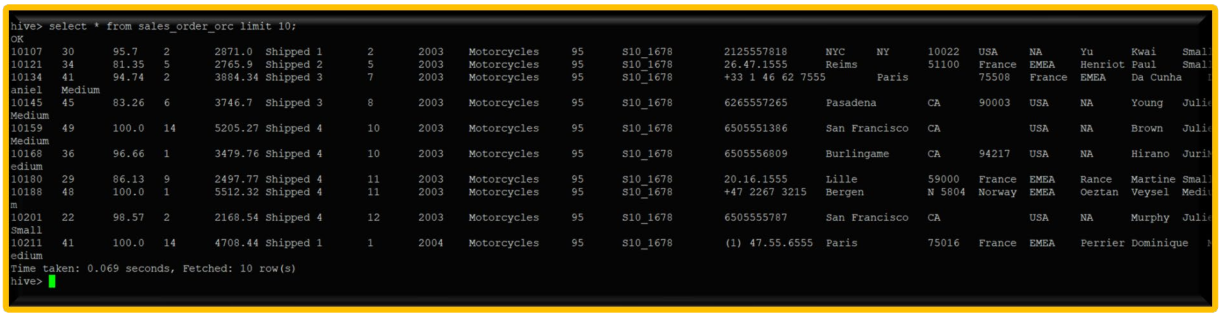
```
cloudera@quickstart:
OK
# col_name              data_type               comment

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

# Detailed Table Information
Database:               default
Owner:                  cloudera
CreateTime:             Fri Sep 16 03:52:32 PDT 2022
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://quickstart.cloudera:8020/user/hive/warehouse/sales_order_orc
Table Type:             MANAGED_TABLE
Table Parameters:
        transient_lastDdlTime   1663325552

# Storage Information
SerDe Library:          org.apache.hadoop.hive.ql.io.orc.OrcSerde
InputFormat:            org.apache.hadoop.hive.ql.io.orc.OrcInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 1.434 seconds, Fetched: 46 row(s)
hive>
```

6. Loading data from "sales_order_csv" table into "sales_order_orc" table.



```
hive> insert into sales_order_orc
    > select * from sales_order_csv;
Query ID = cloudera_20220916035959_4438a261-cebe-495e-a1a9-015f796aaddd
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_1663320923809_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663320923809_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663320923809_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-09-16 03:59:59,908 Stage-1 map = 0%,  reduce = 0%
2022-09-16 04:00:26,287 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.37 sec
MapReduce Total cumulative CPU time: 9 seconds 370 msec
Ended Job = job_1663320923809_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/sales_order_orc/.hive-staging_hive_2022-09-16_03-59-06_192_8062975353365905754-1/-ext-10000
Loading data to table default.sales_order_orc
Table default.sales_order_orc stats: [numFiles=1, numRows=2823, totalSize=37548, rawDataSize=3153291]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 9.37 sec   HDFS Read: 367206 HDFS Write: 37634 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 370 msec
OK
Time taken: 84.541 seconds
hive>
```

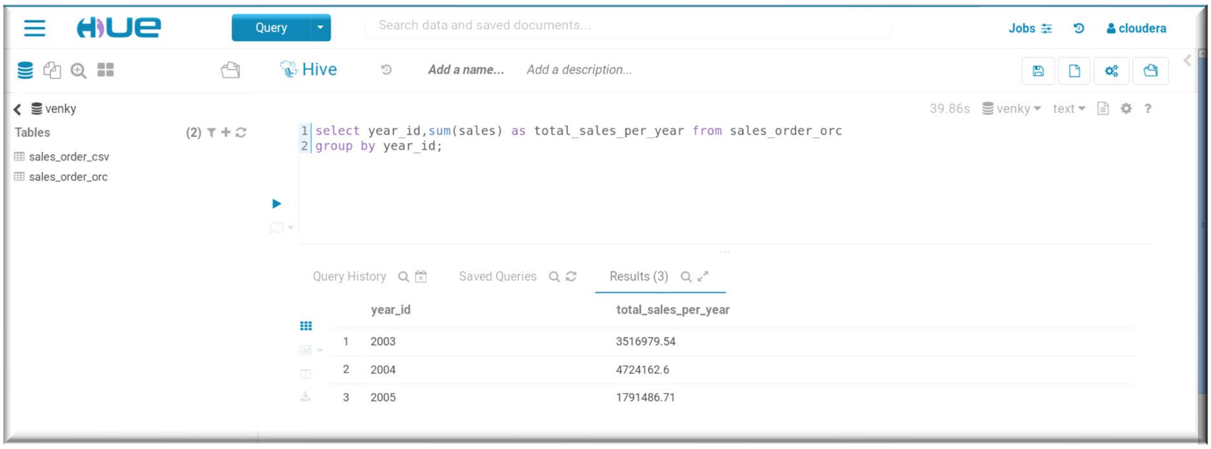
7. Checking the data in the table.



```
hive> select * from sales_order_orc limit 10;
OK
10107   30      95.7    2       2871.0  Shipped 1       2       2003    Motorcycles     95      S10_1678        2125557818      NYC     NY      10022   USA     NA      Yu      Kwai    Smal
10121   34      81.35   5       2765.9  Shipped 2       5       2003    Motorcycles     95      S10_1678        26.47.1555      Reims           51100   France  EMEA    Henriot Paul    Smal
10134   41      94.74   2       3884.34 Shipped 3       7       2003    Motorcycles     95      S10_1678        +33 1 46 62 7555        Paris           75508   France  EMEA    Da Cunha
aniel   Medium
10145   45      83.26   6       3746.7  Shipped 3       8       2003    Motorcycles     95      S10_1678        6265557265      Pasadena        CA      90003   USA     NA      Young   Julie
Medium
10159   49      100.0   14      5205.27 Shipped 4       10      2003    Motorcycles     95      S10_1678        6505551386      San Francisco   CA              USA     NA      Brown   Julie
Medium
10168   36      96.66   1       3479.76 Shipped 4       10      2003    Motorcycles     95      S10_1678        6505556809      Burlingame      CA      94217   USA     NA      Hirano  Juri
edium
10180   29      86.13   9       2497.77 Shipped 4       11      2003    Motorcycles     95      S10_1678        20.16.1555      Lille           59000   France  EMEA    Rance   Martine Smal
10188   48      100.0   1       5512.32 Shipped 4       11      2003    Motorcycles     95      S10_1678        +47 2267 3215   Bergen          N 5804  Norway  EMEA    Oeztan  Veysel  Medi
m
10201   22      98.57   2       2168.54 Shipped 4       12      2003    Motorcycles     95      S10_1678        6505555787      San Francisco   CA              USA     NA      Murphy  Julie
Small
10211   41      100.0   14      4708.44 Shipped 1       1       2004    Motorcycles     95      S10_1678        (1) 47.55.6555  Paris           75016   France  EMEA    Perrier Dominique
edium
Time taken: 0.069 seconds, Fetched: 10 row(s)
hive> 
```

Performed below mentioned queries on "sales_order_orc" table:

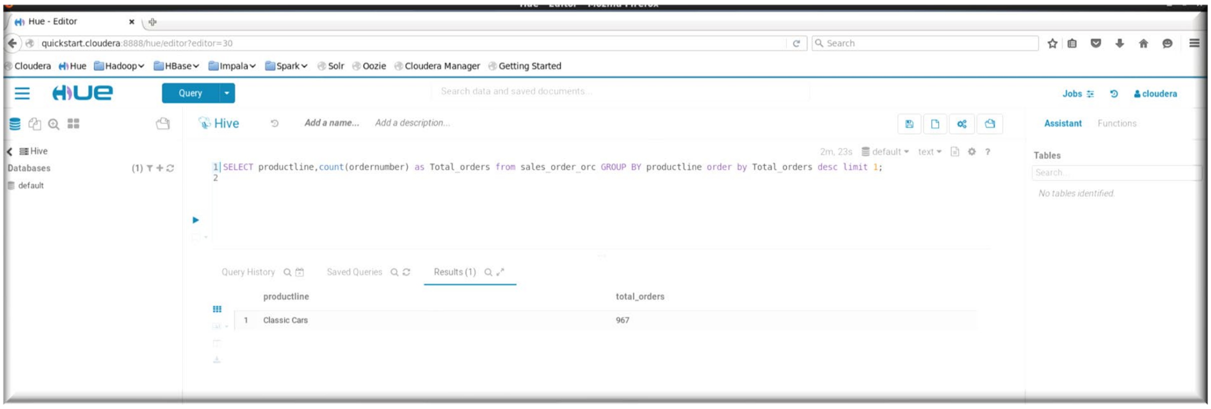
a. Calculate total sales per year



```
select year_id,sum(sales) as total_sales_per_year from sales_order_orc
group by year_id;
```

	year_id	total_sales_per_year
1	2003	3516979.54
2	2004	4724162.6
3	2005	1791486.71

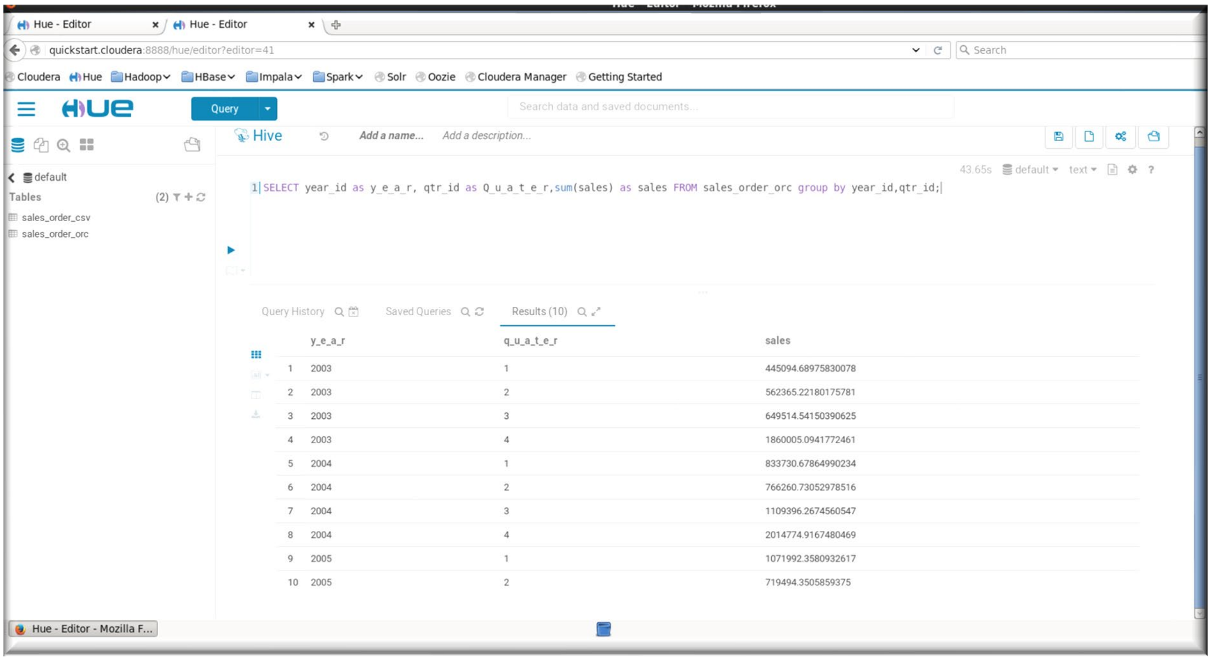
b. Find a product for which maximum orders were placed



```
SELECT productline,count(ordernumber) as Total_orders from sales_order_orc GROUP BY productline order by Total_orders desc limit 1;
```

	productline	total_orders
1	Classic Cars	967

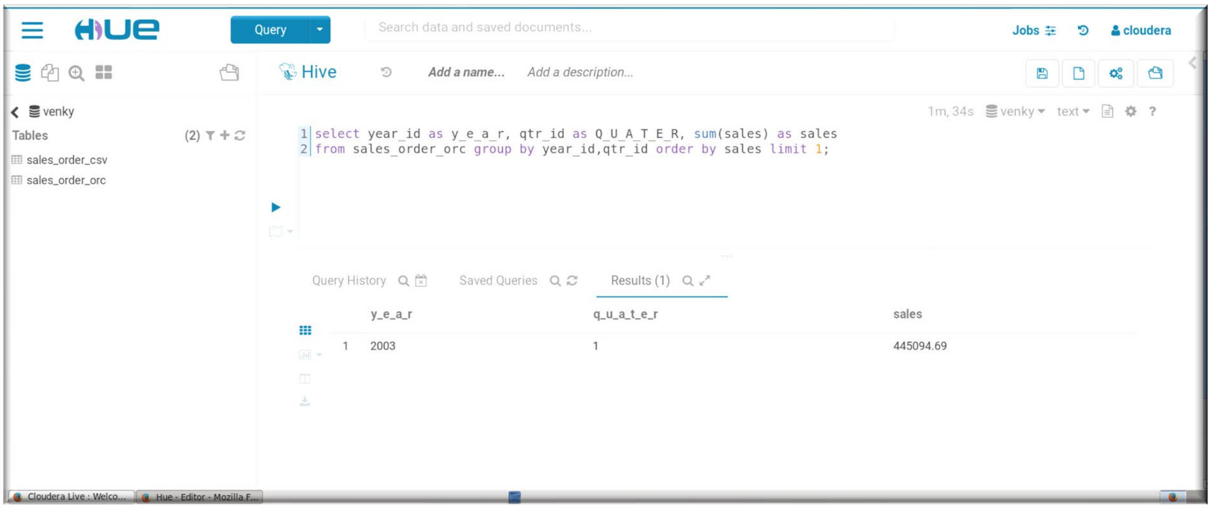
c. Calculate the total sales for each quarter



```
SELECT year_id as y_e_a_r, qtr_id as Q_u_a_t_e_r,sum(sales) as sales FROM sales_order_orc group by year_id,qtr_id;
```

	y_e_a_r	q_u_a_t_e_r	sales
1	2003	1	445094.68975830078
2	2003	2	562365.22180175781
3	2003	3	649514.54150390625
4	2003	4	1860005.0941772461
5	2004	1	833730.67864990234
6	2004	2	766260.73052978516
7	2004	3	1109396.2674560547
8	2004	4	2014774.9167480469
9	2005	1	1071992.3580932617
10	2005	2	719494.3505859375

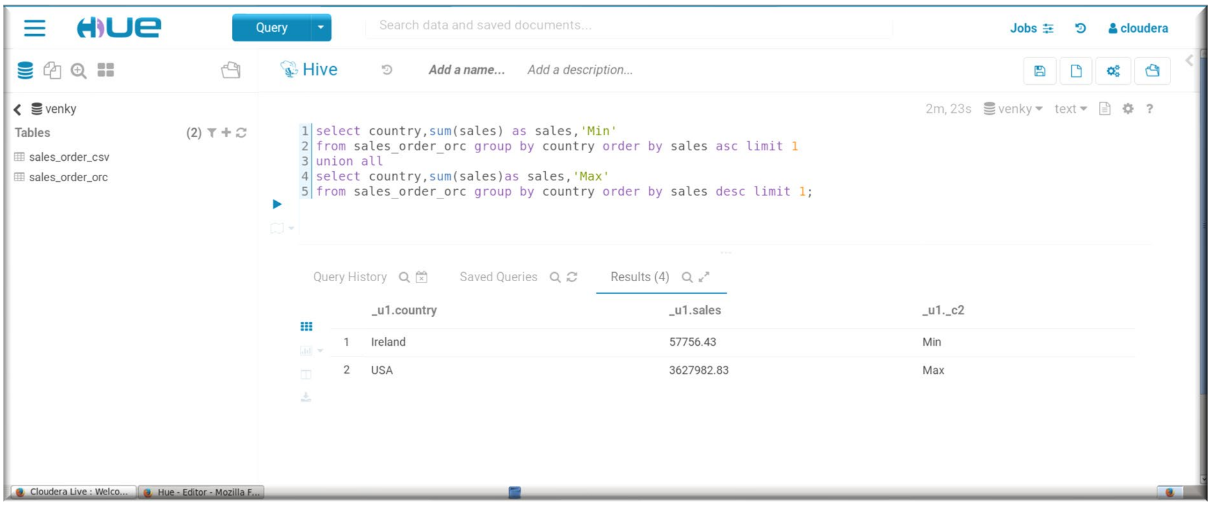
d. In which quarter sales was minimum



```
select year_id as y_e_a_r, qtr_id as Q_U_A_T_E_R, sum(sales) as sales
from sales_order_orc group by year_id,qtr_id order by sales limit 1;
```

	y_e_a_r	q_u_a_t_e_r	sales
1	2003	1	445094.69

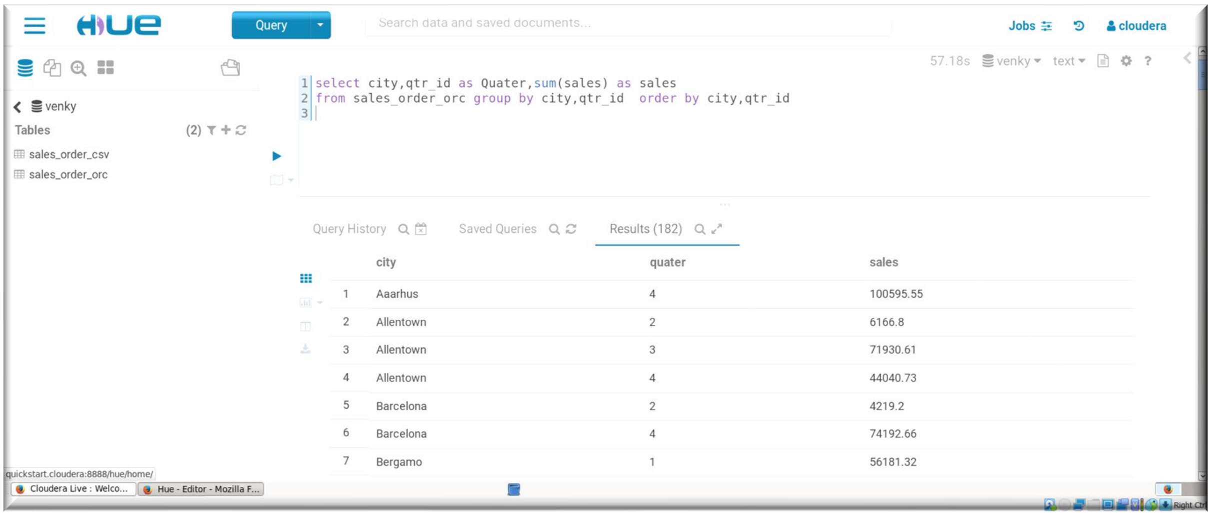
e. In which country sales was maximum and in which country sales was minimum



```
select country,sum(sales) as sales,'Min'
from sales_order_orc group by country order by sales asc limit 1
union all
select country,sum(sales)as sales,'Max'
from sales_order_orc group by country order by sales desc limit 1;
```

	_u1.country	_u1.sales	_u1._c2
1	Ireland	57756.43	Min
2	USA	3627982.83	Max

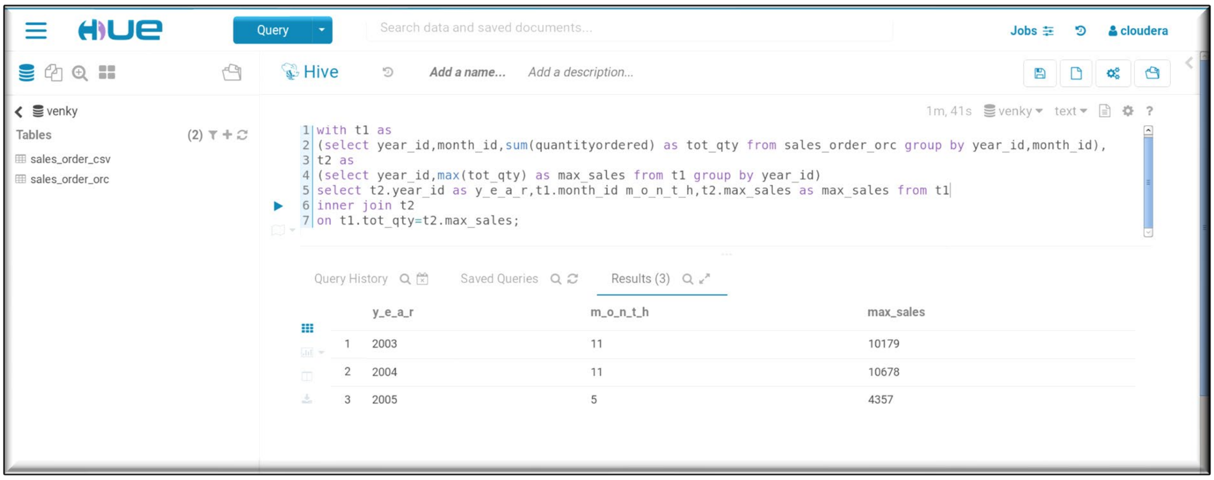
f. Calculate quarterly sales for each city



```
select city,qtr_id as Quater,sum(sales) as sales
from sales_order_orc group by city,qtr_id  order by city,qtr_id
```

	city	quater	sales
1	Aaarhus	4	100595.55
2	Allentown	2	6166.8
3	Allentown	3	71930.61
4	Allentown	4	44040.73
5	Barcelona	2	4219.2
6	Barcelona	4	74192.66
7	Bergamo	1	56181.32

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



```
with t1 as
(select year_id,month_id,sum(quantityordered) as tot_qty from sales_order_orc group by year_id,month_id),
t2 as
(select year_id,max(tot_qty) as max_sales from t1 group by year_id)
select t2.year_id as y_e_a_r,t1.month_id m_o_n_t_h,t2.max_sales as max_sales from t1
inner join t2
on t1.tot_qty=t2.max_sales;
```

	y_e_a_r	m_o_n_t_h	max_sales
1	2003	11	10179
2	2004	11	10678
3	2005	5	4357