

1. Download vechile sales data -> https://github.com/shashank-mishra219/Hive-Class/blob/main/sales\_order\_data.csv

**-->sales\_order\_data.csv**

**In local path - /home/cloudera/sidd/Challenge/sales\_order\_data.csv**

2. Store raw data into hdfs location

**$ hadoop fs -put /home/cloudera/sidd/Challenge/sales\_order\_data.csv /sidd/challenge/**

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\_csv**

**(**

**ORDERNUMBER int,QUANTITYORDERED int,PRICEEACH int,ORDERLINENUMBER int,**

**SALES int,STATUS string,QTR\_ID int,MONTH\_ID int,YEAR\_ID int,PRODUCTLINE string,**

**MSRP int,PRODUCTCODE string,PHONE string,CITY string,STATE string,POSTALCODE int,**

**COUNTRY string,TERRITORY string,CONTACTLASTNAME string,CONTACTFIRSTNAME string,**

**DEALSIZE string**

**)**

**row format delimited**

**fields terminated by ','**

**TBLPROPERTIES(“skip.header.line.count” = “1”);**

4. Load data from hdfs path into "sales\_order\_csv"

**hive> load data inpath '/sidd/challenge/sales\_order\_data.csv' into table sales\_order\_csv;**

**hive> set hive.cli.print.header = true;**

**to show headers in result of select statement.**



5. Create an internal hive table which will store data in ORC format "sales\_order\_orc"

**hive>**

**create table sales\_order\_orc**

(**ORDERNUMBER int,QUANTITYORDERED int,PRICEEACH int,ORDERLINENUMBER int,**

**SALES int,STATUS string,QTR\_ID int,MONTH\_ID int,YEAR\_ID int,PRODUCTLINE string,**

**MSRP int,PRODUCTCODE string,PHONE string,CITY string,STATE string,POSTALCODE int,**

**COUNTRY string,TERRITORY string,CONTACTLASTNAME string,CONTACTFIRSTNAME string,**

**DEALSIZE string**

**)**

**stored as ORC;**

6. Load data from "sales\_order\_csv" into "sales\_order\_orc"

**hive> from sales\_order\_csv insert overwrite table sales\_order\_orc select \*;**

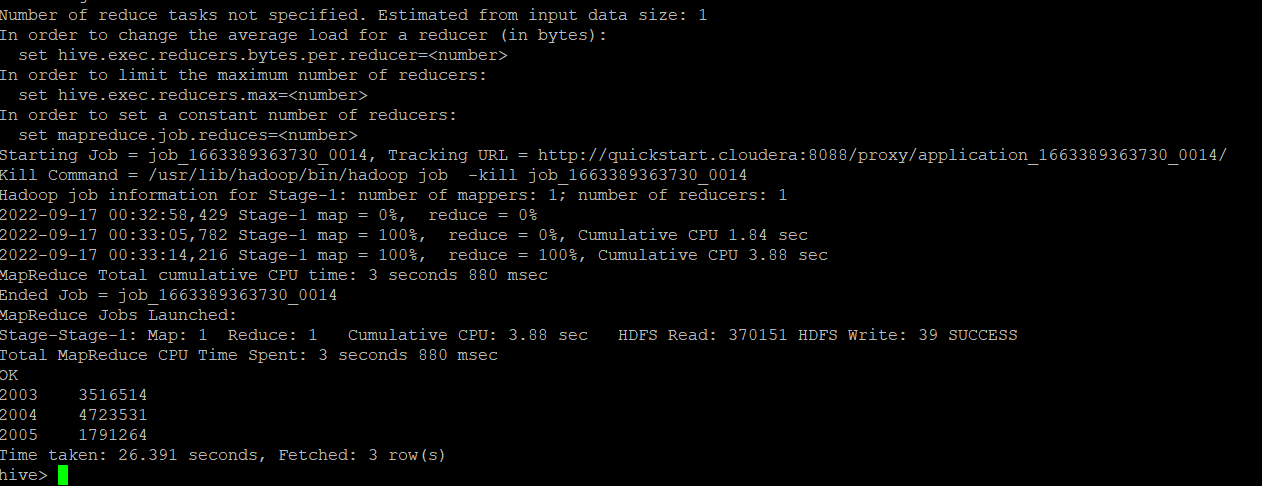
**or**

**hive> insert overwrite table sales\_order\_orc select \* from sales\_order\_csv;**

7. Perform below menioned queries on "sales\_order\_orc" table :

a. Calculatye total sales per year

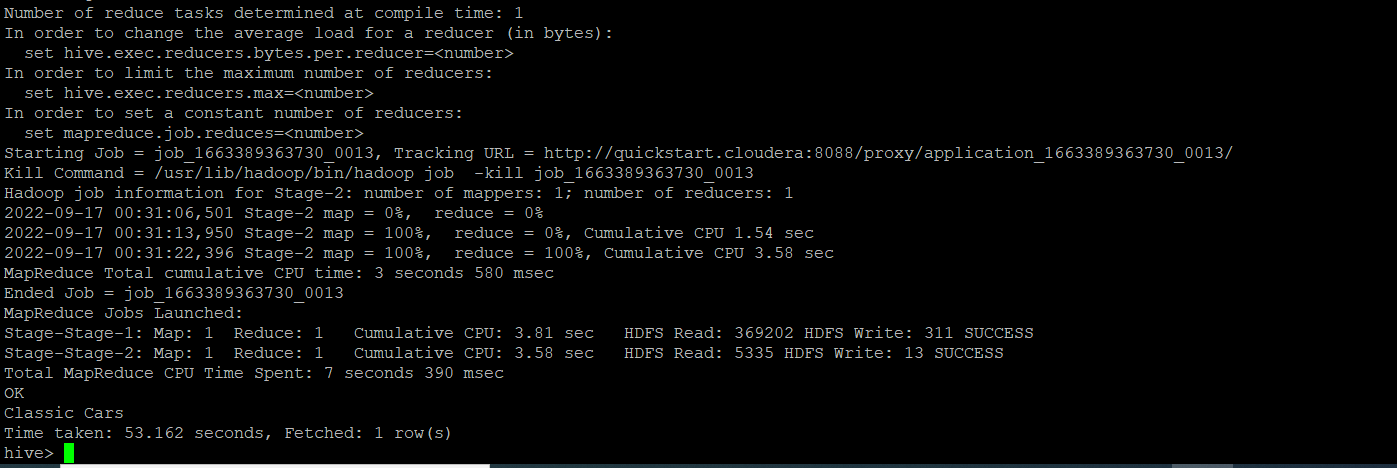
**hive> select year\_ID as year,sum(sales) as Total\_Sales from sales\_order\_csv group by year\_ID ;**





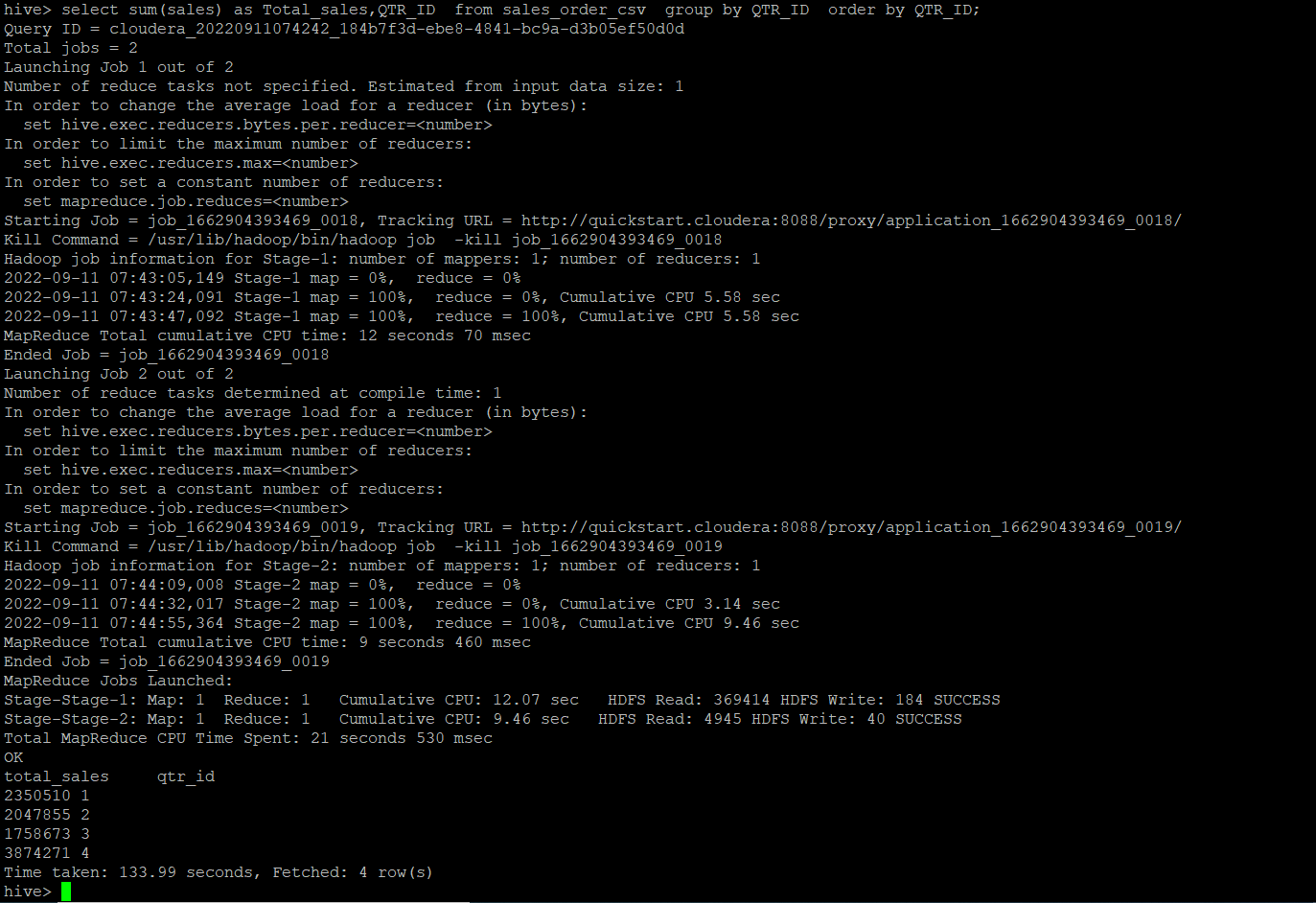
b. Find a product for which maximum orders were placed

**hive> select PRODUCTLINE from (select PRODUCTLINE ,sum(QUANTITYORDERED) max from sales\_order\_csv group by PRODUCTLINE order by max desc limit 1) a ;**



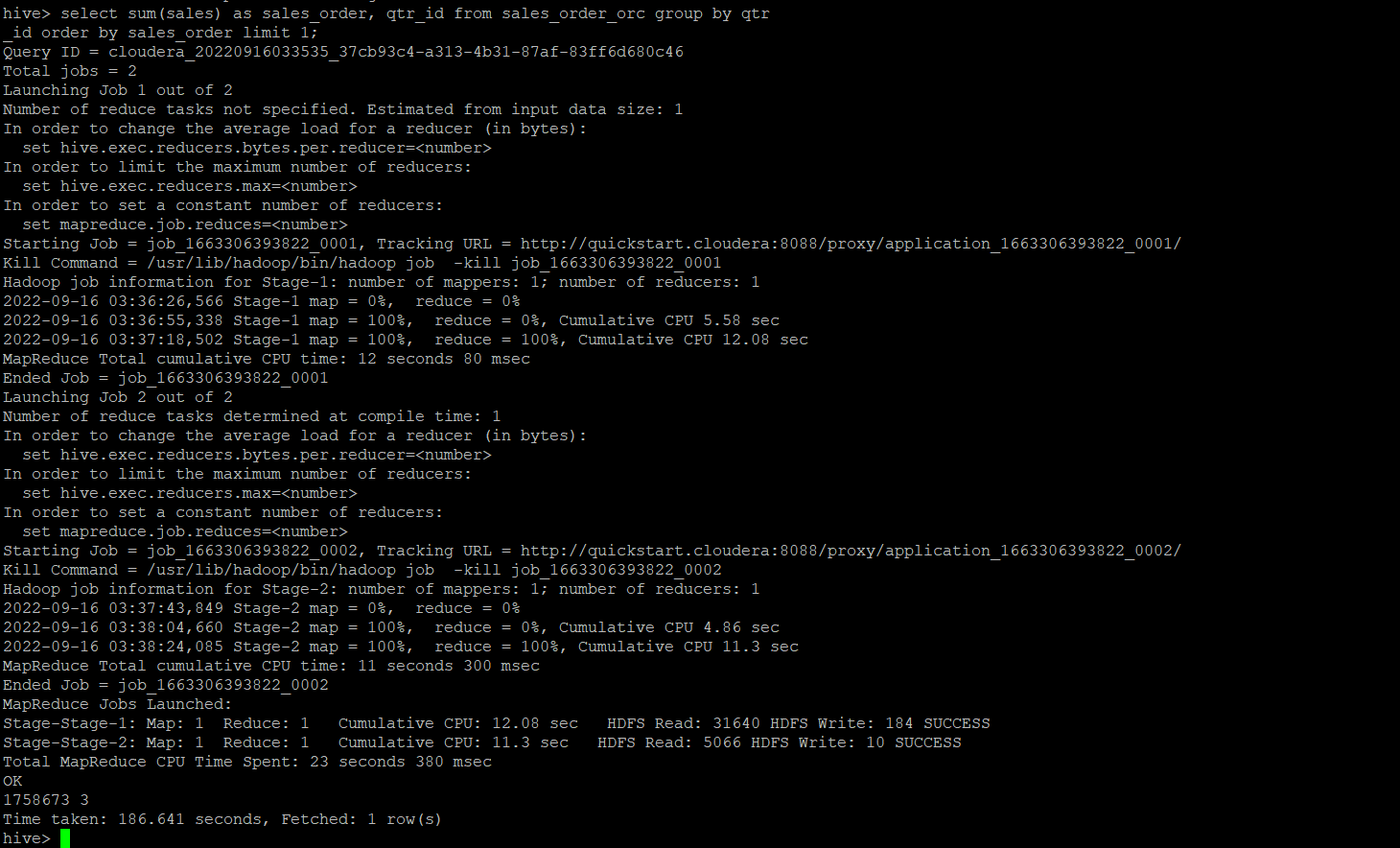
c. Calculate the total sales for each quarter

**hive> select sum(sales) as Total\_sales,QTR\_ID from sales\_order\_csv group by QTR\_ID order by QTR\_ID;**



d. In which quarter sales was minimum

**hive> select sum(sales) as sales\_order, qtr\_id from sales\_order\_orc group by qtr\_id order by sales\_order limit 1;**

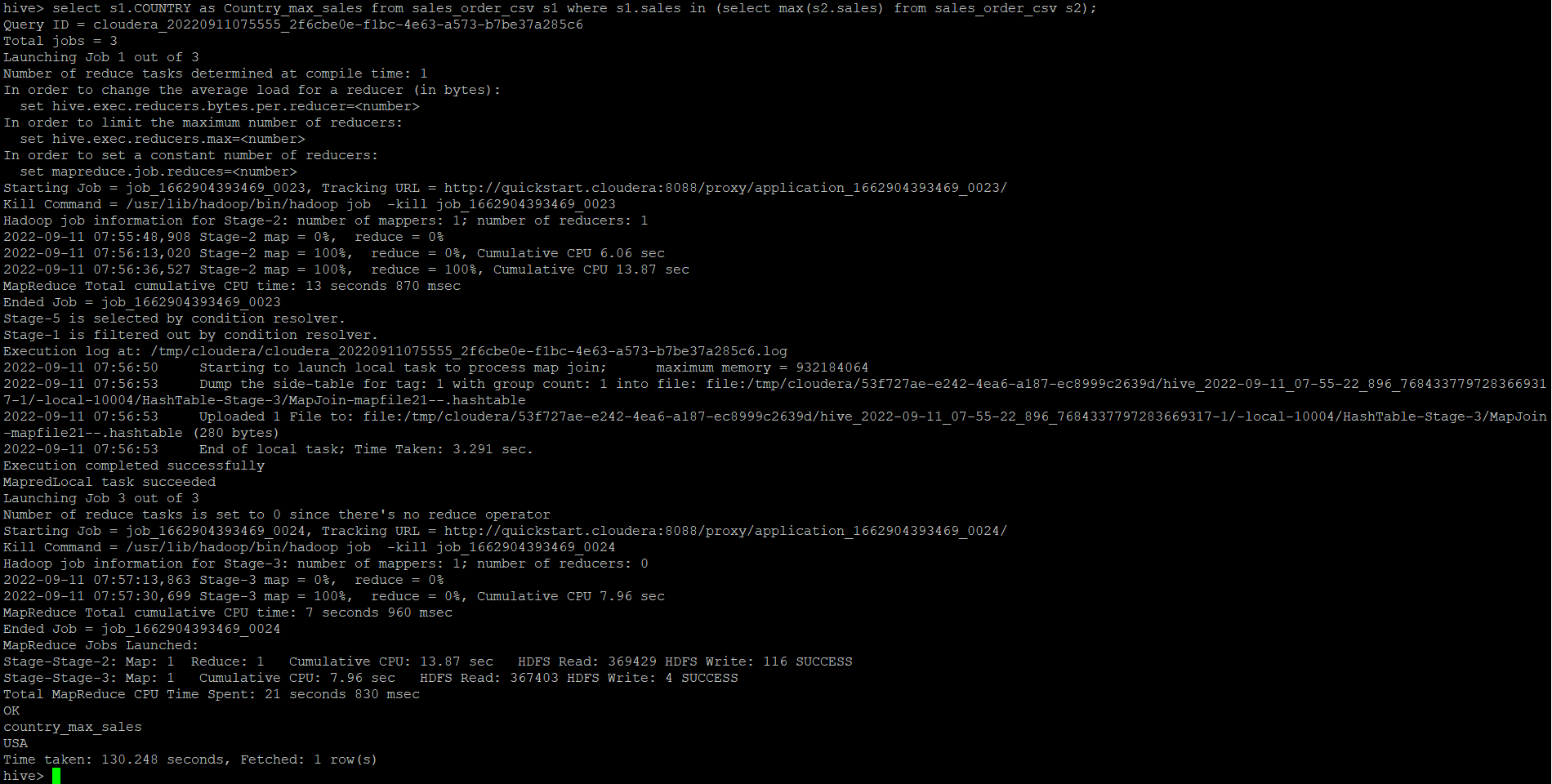


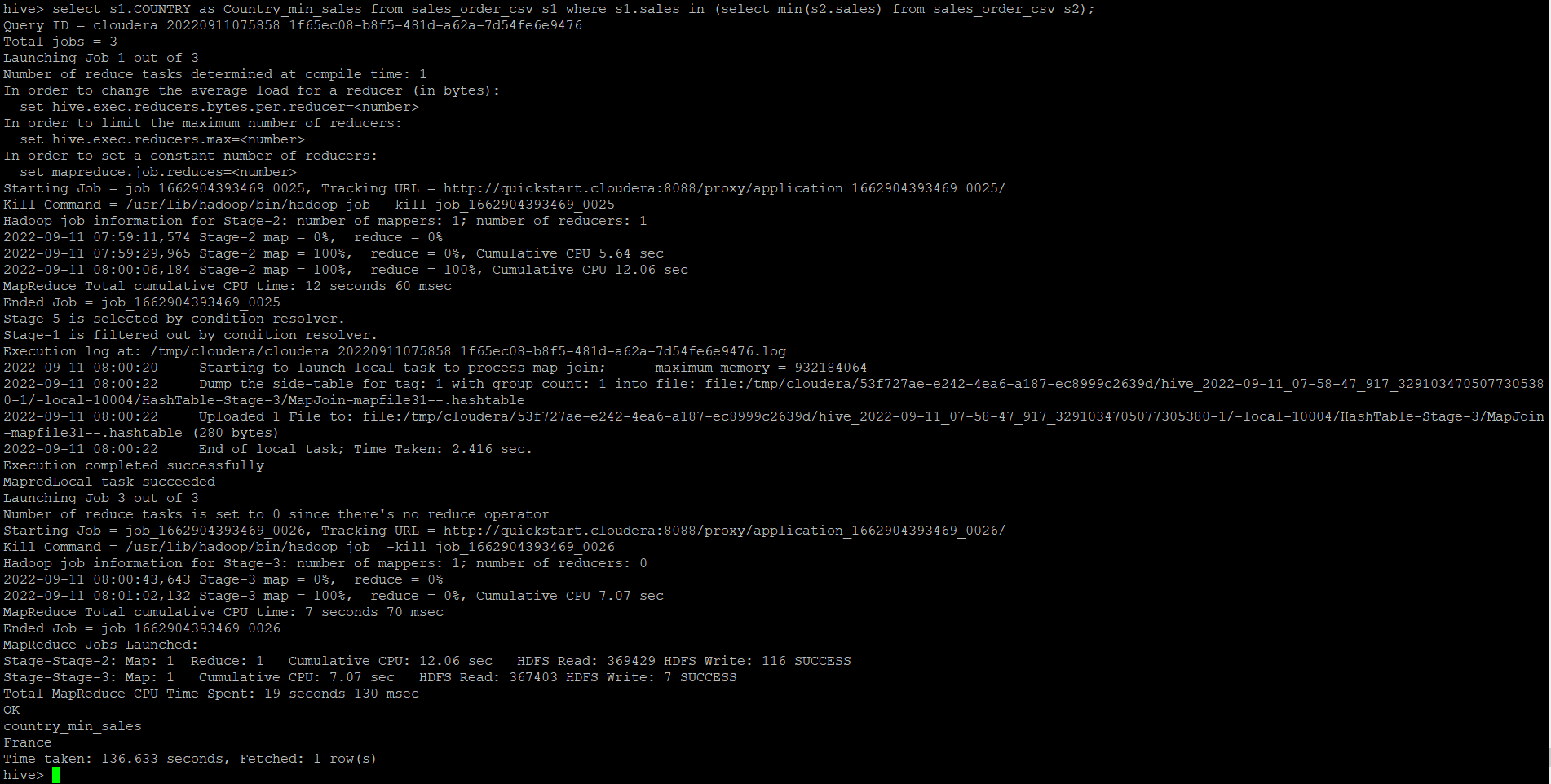
e. In which country sales was maximum and in which country sales was minimum

**hive> select s1.COUNTRY as Country\_max\_sales from sales\_order\_csv s1 where s1.sales in (select max(s2.sales) from sales\_order\_csv s2);**

**hive> select s1.COUNTRY as Country\_min\_sales from sales\_order\_csv s1 where s1.sales in (select min(s2.sales) from sales\_order\_csv s2);**

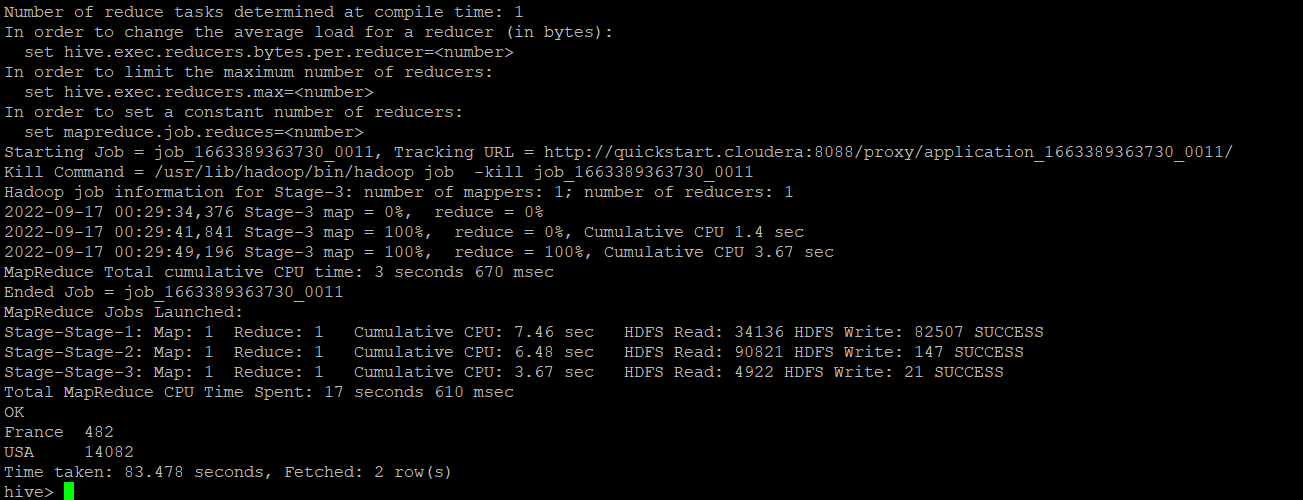






**Hive> select A.country,A.sales from (select country,sales,rank() over (order by sales) rnk\_min,rank() over(order by sales desc) rnk\_max from sales\_order\_orc)A where rnk\_min = 1 or rnk\_max =1 order by sales;**





f. Calculate quartelry sales for each city

**hive>select quartelry\_sales,city from (select sum(s.sales) as quartelry\_sales,s.QTR\_ID,s.city from sales\_order\_csv s group by s.QTR\_ID,s.city)a;**

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

**Hive> select month\_id,year\_id,QUANTITYORDERED from (select distinct month\_id,year\_id,QUANTITYORDERED ,dense\_rank() over(partition by year\_id order by QUANTITYORDERED desc) as rnk from sales\_order\_csv)a where a.rnk = 1;**

