**Hive Assignment**

1. Download Vehicle sales data from ----> <https://github.com/shashank-mishra219/Hive-Class/blob/main/sales_order_data.csv>
2. Store raw data into hdfs location:

**hadoop fs -put sales\_order\_data.csv /tmp/hive/**

1. 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

**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**

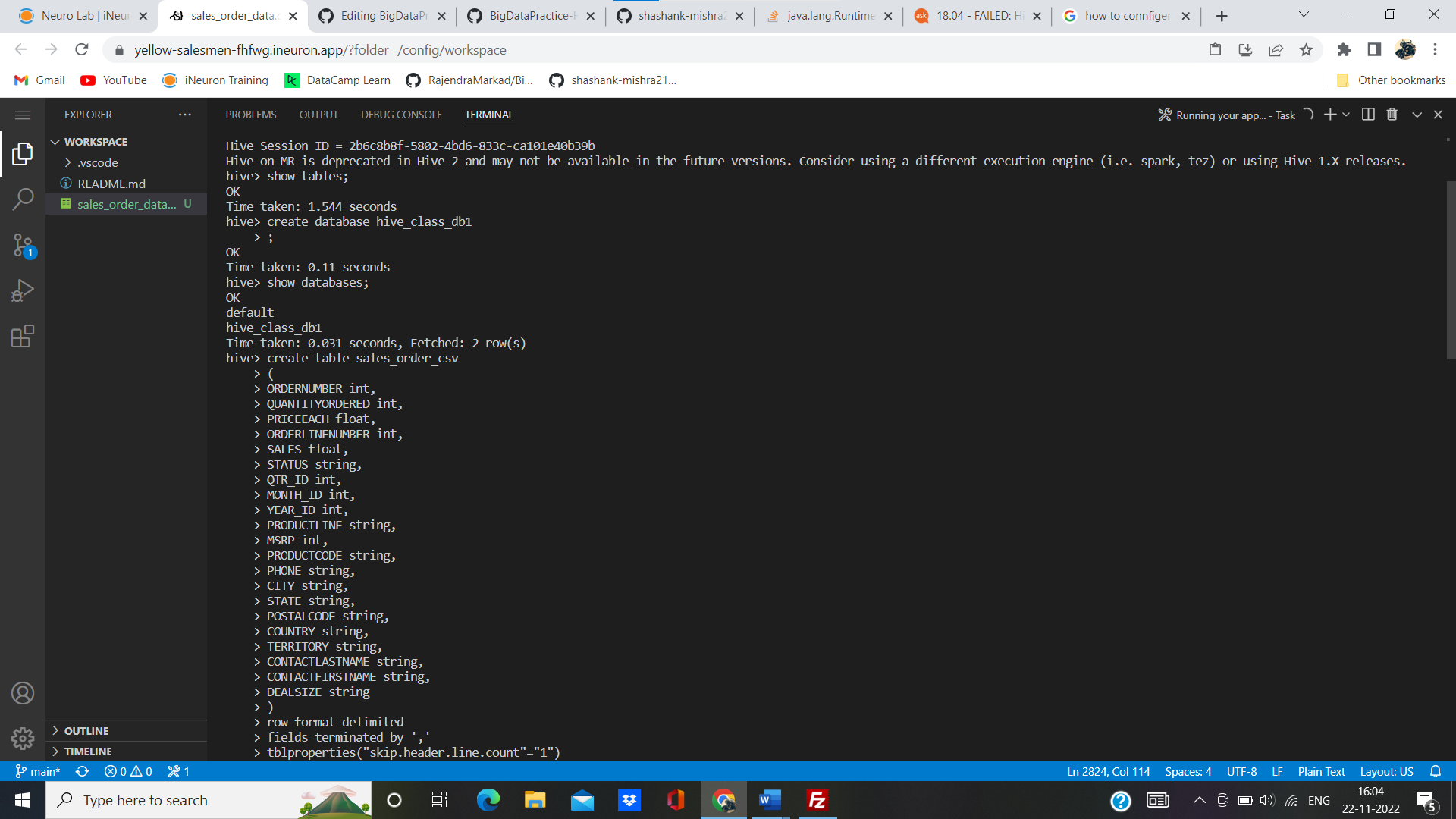
**)**

**row format delimited**

**fields terminated by ','**

**tblproperties("skip.header.line.count"="1")**

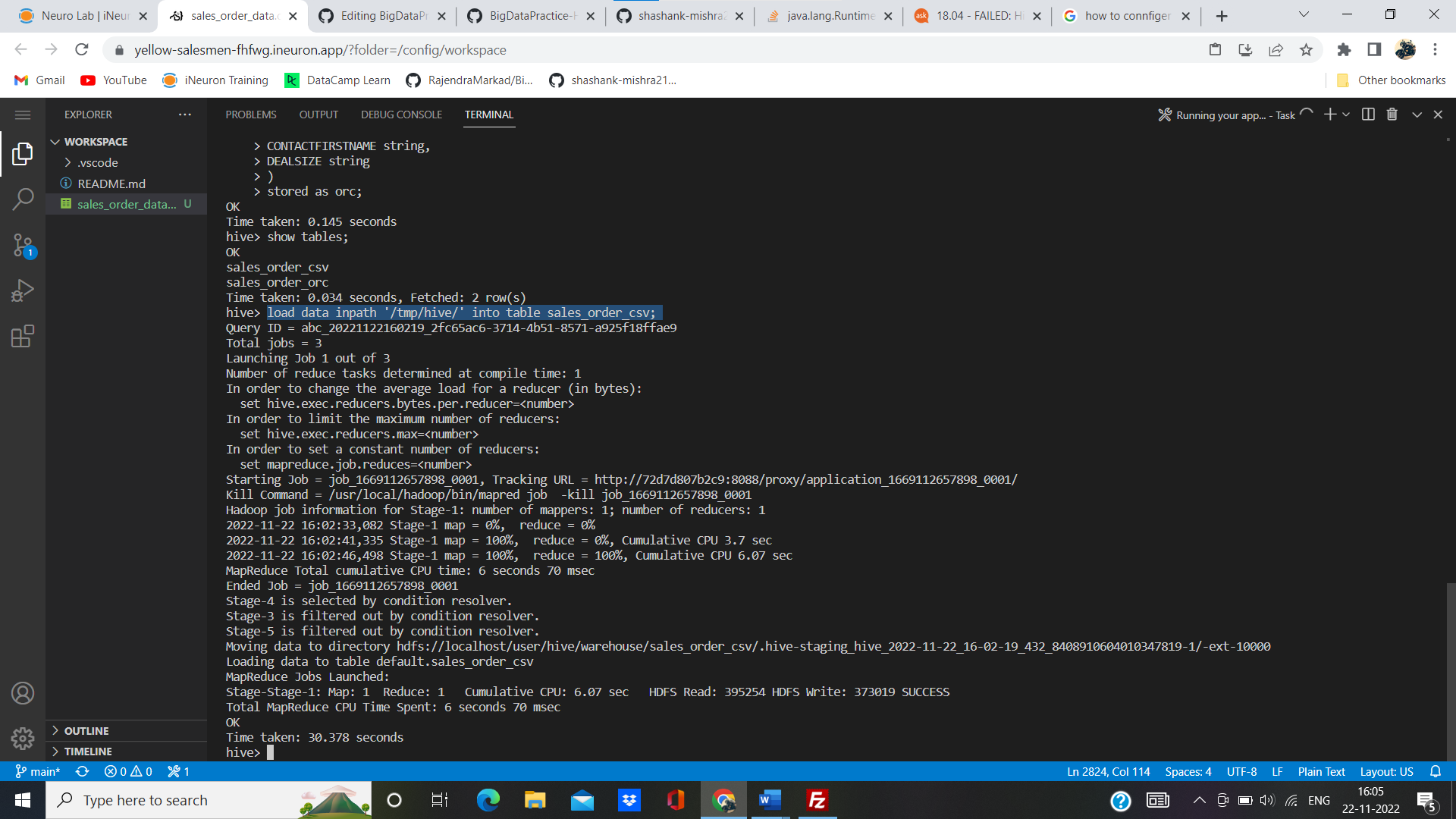
**;**

****

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

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

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

****

1. 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 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,**

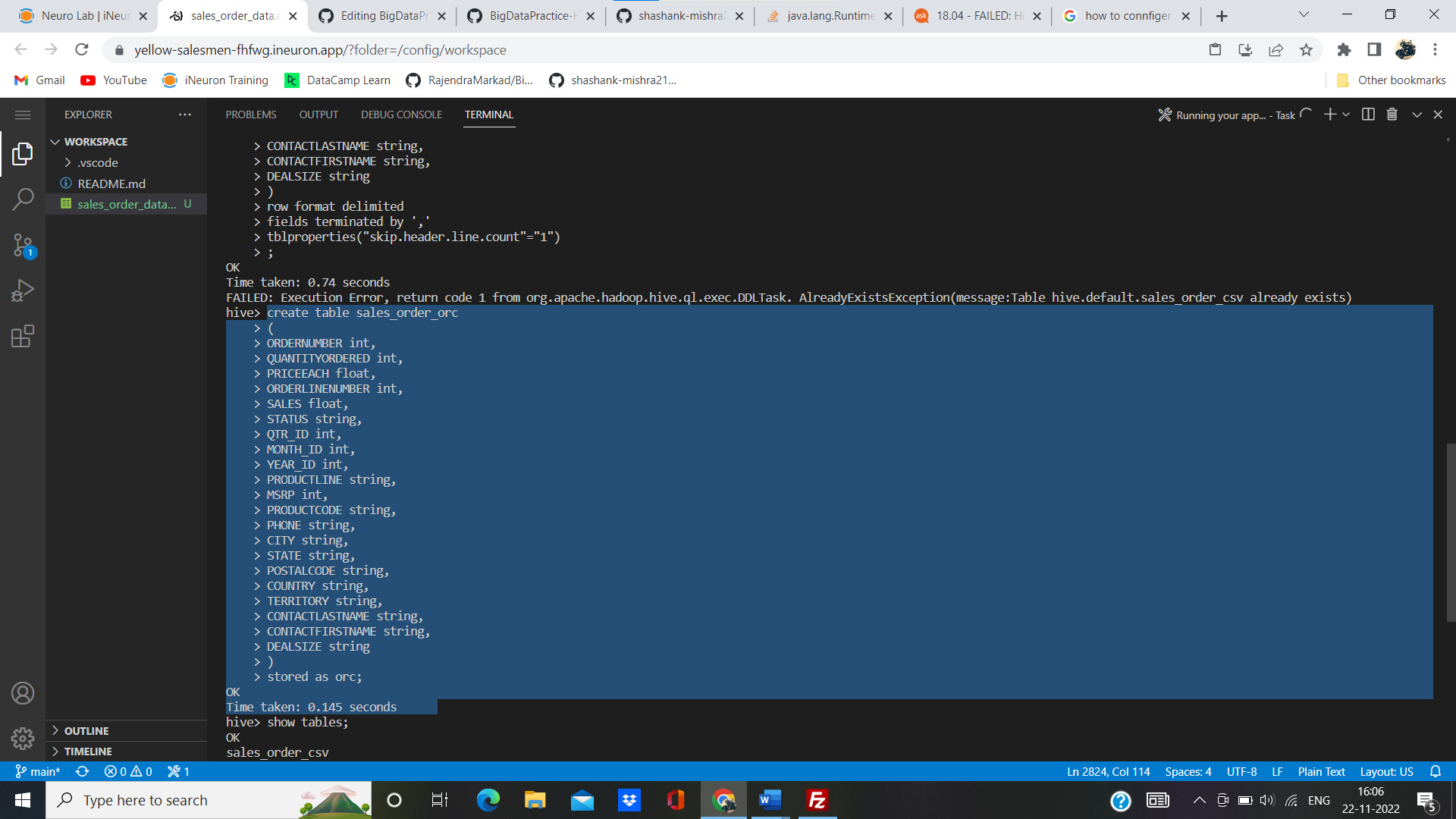
**> CONTACTLASTNAME string,**

**> CONTACTFIRSTNAME string,**

**> DEALSIZE string**

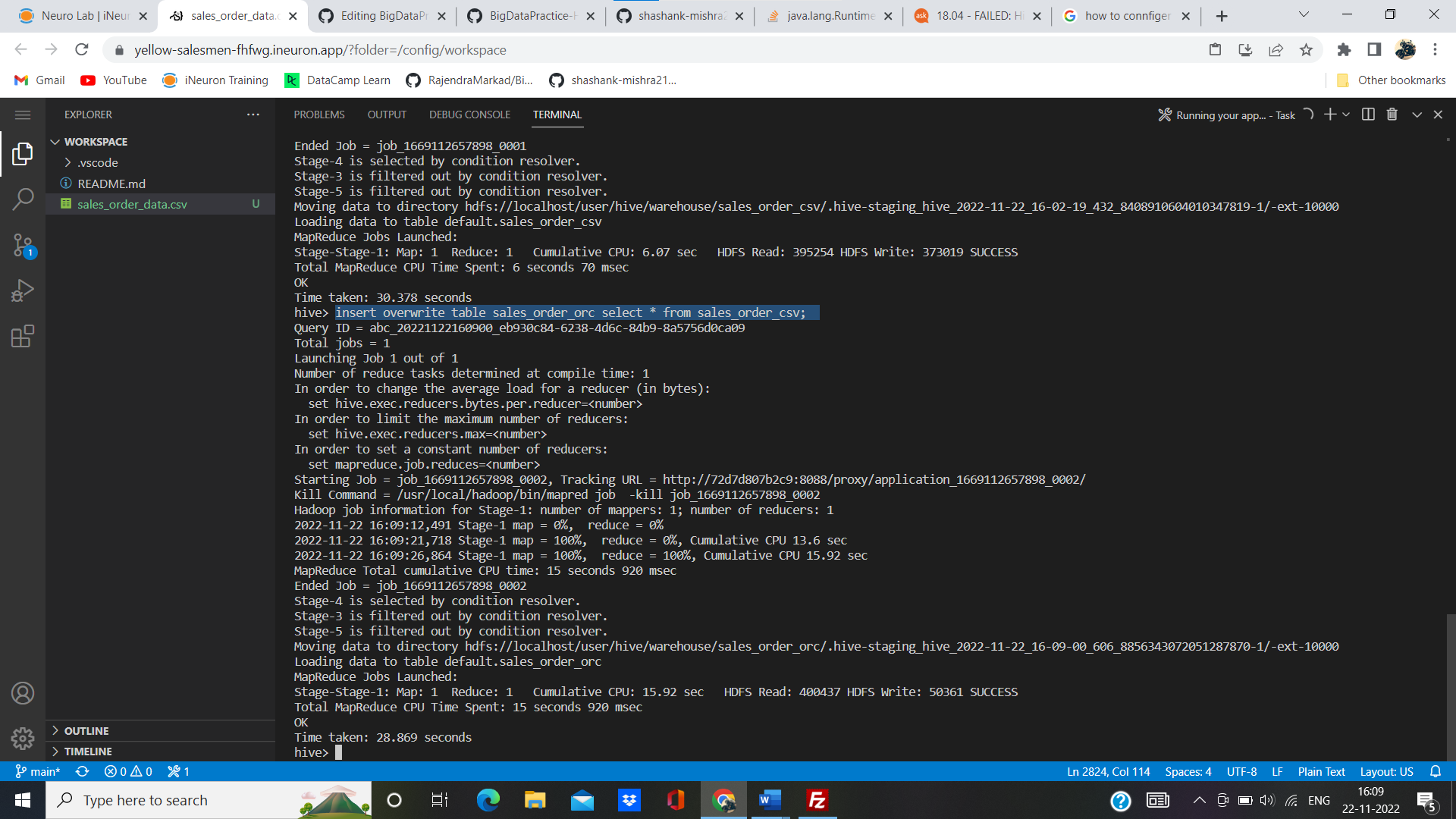
**> )**

**> stored as orc;**

****

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

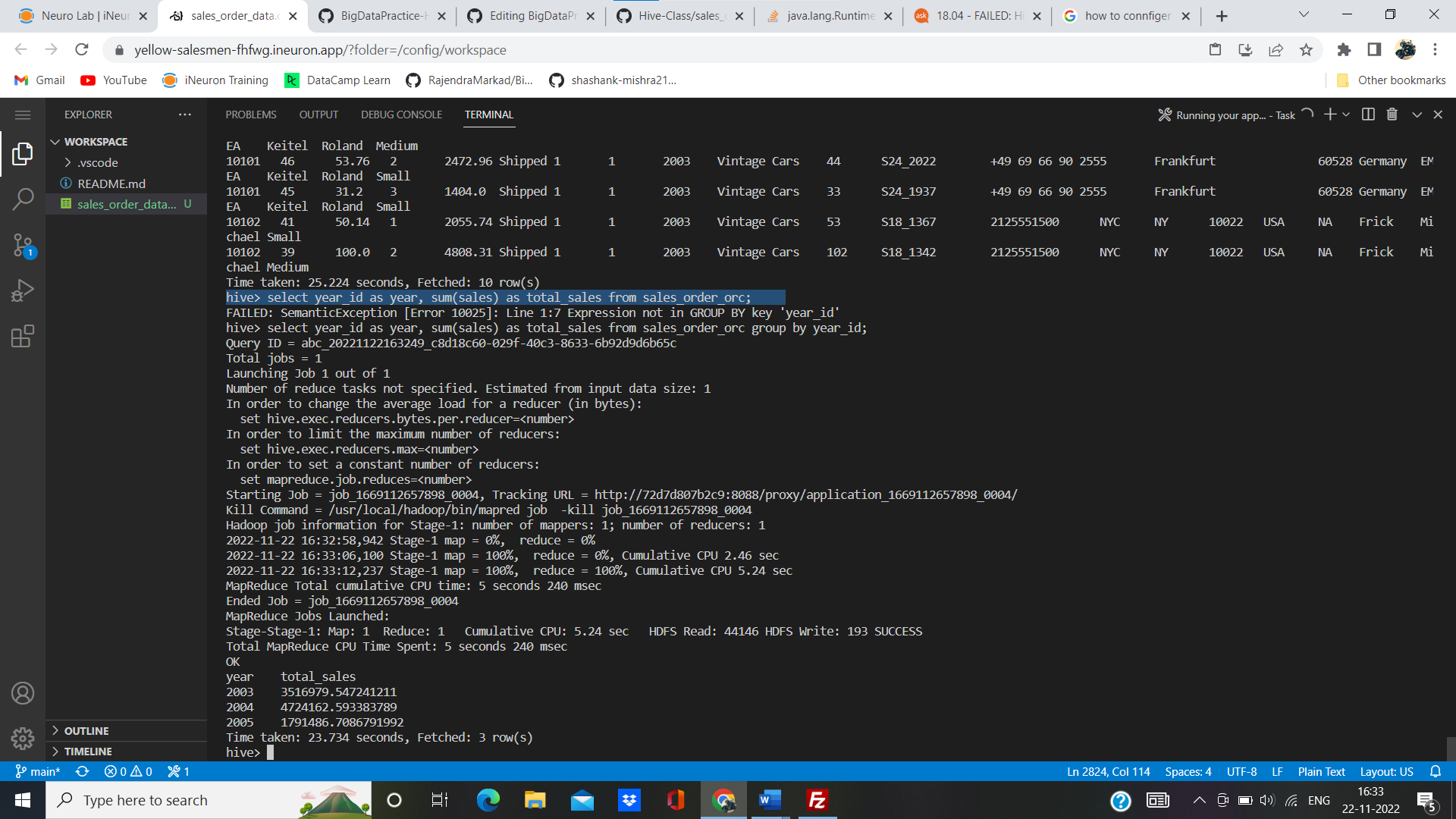
**Hive> insert overwrite table sales\_order\_orc select \* from sales\_order\_csv;**



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

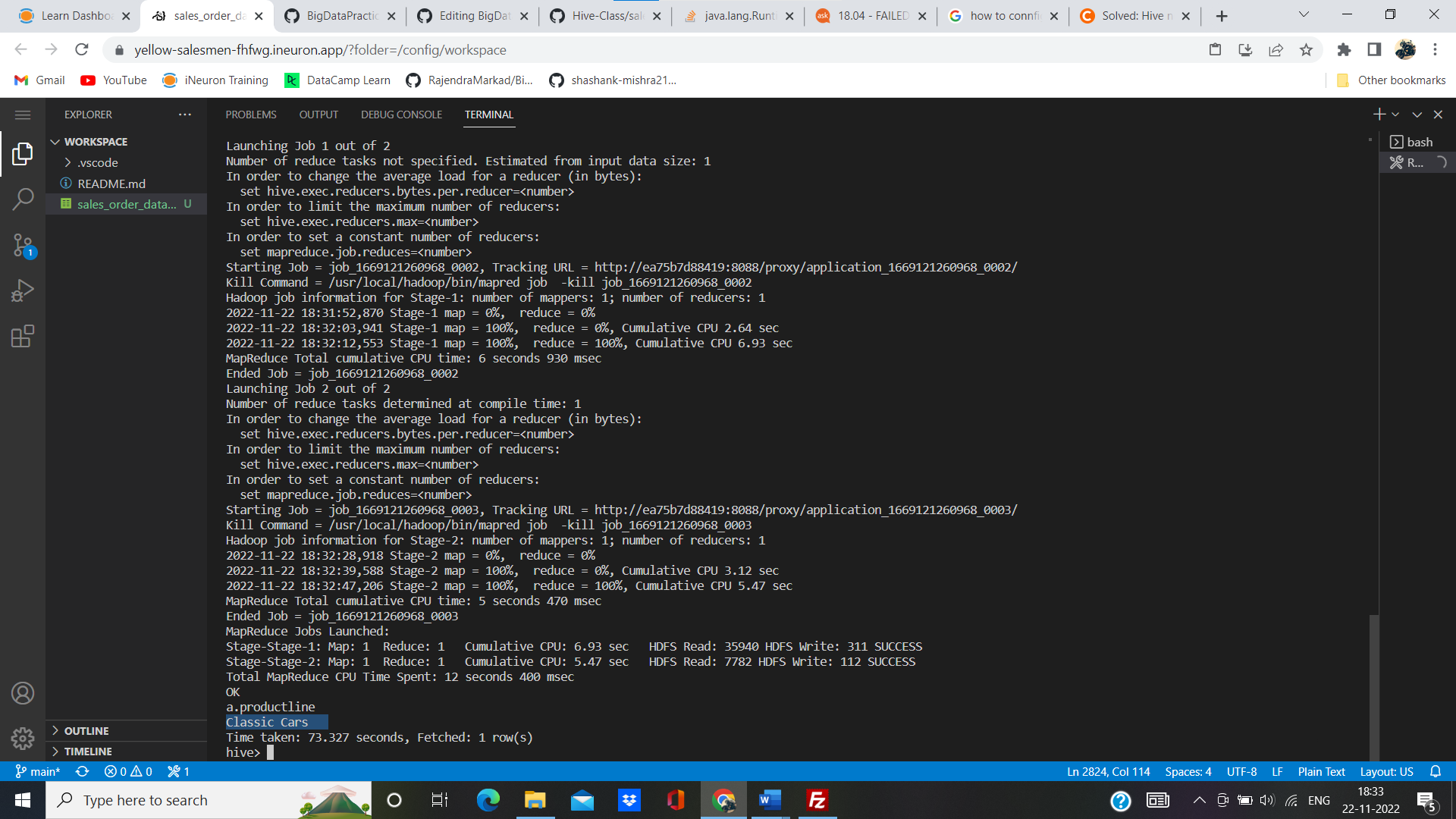
a. Calculate the total sales per year

**hive> select year\_id as year, sum(sales) as total\_sales from sales\_order\_orc group by year\_id ;**



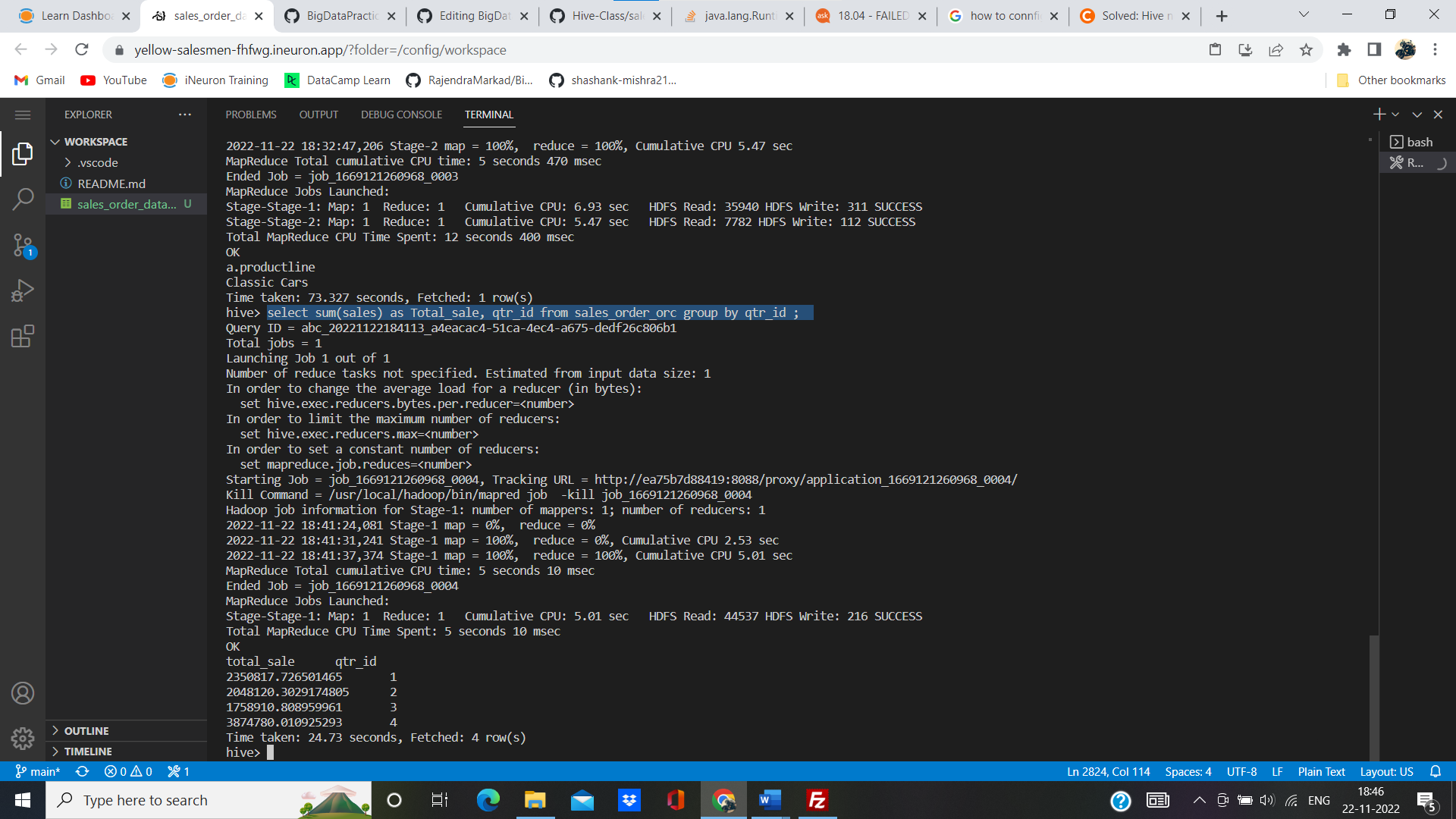
b. Find a product for which maximum orders were placed

hive>**Select A.productline from (Select sum(quantityordered) as maxquantity, productline from sales\_order\_orc group by productline order by maxquantity desc limit 1) A;**

****

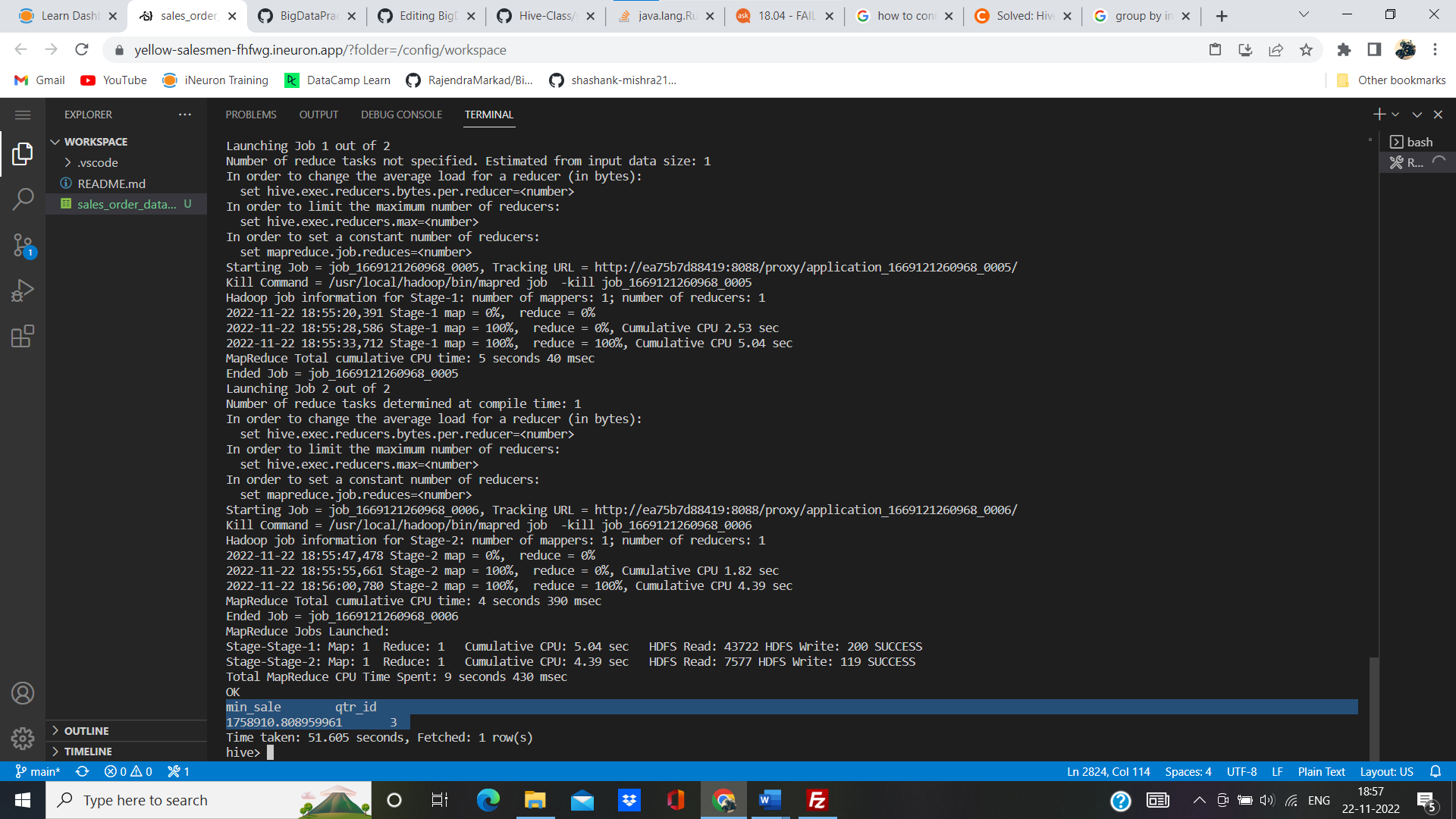
c. Calculate the total sales for each quarter

**Select sum(sales), qtr\_id from sales\_order\_orc group by qtr\_id;**



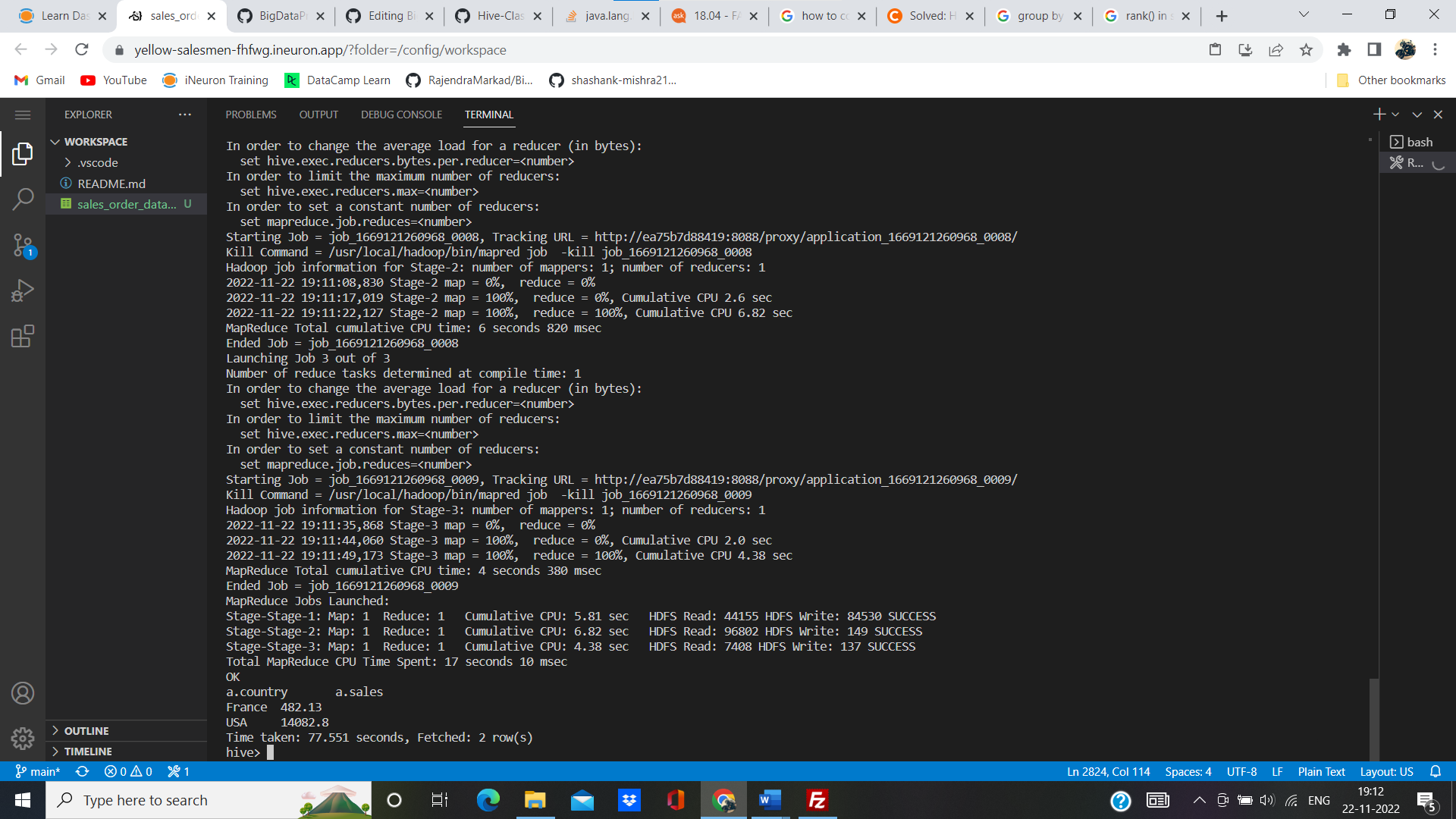
d. In which quarter sales was minimum?

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

****

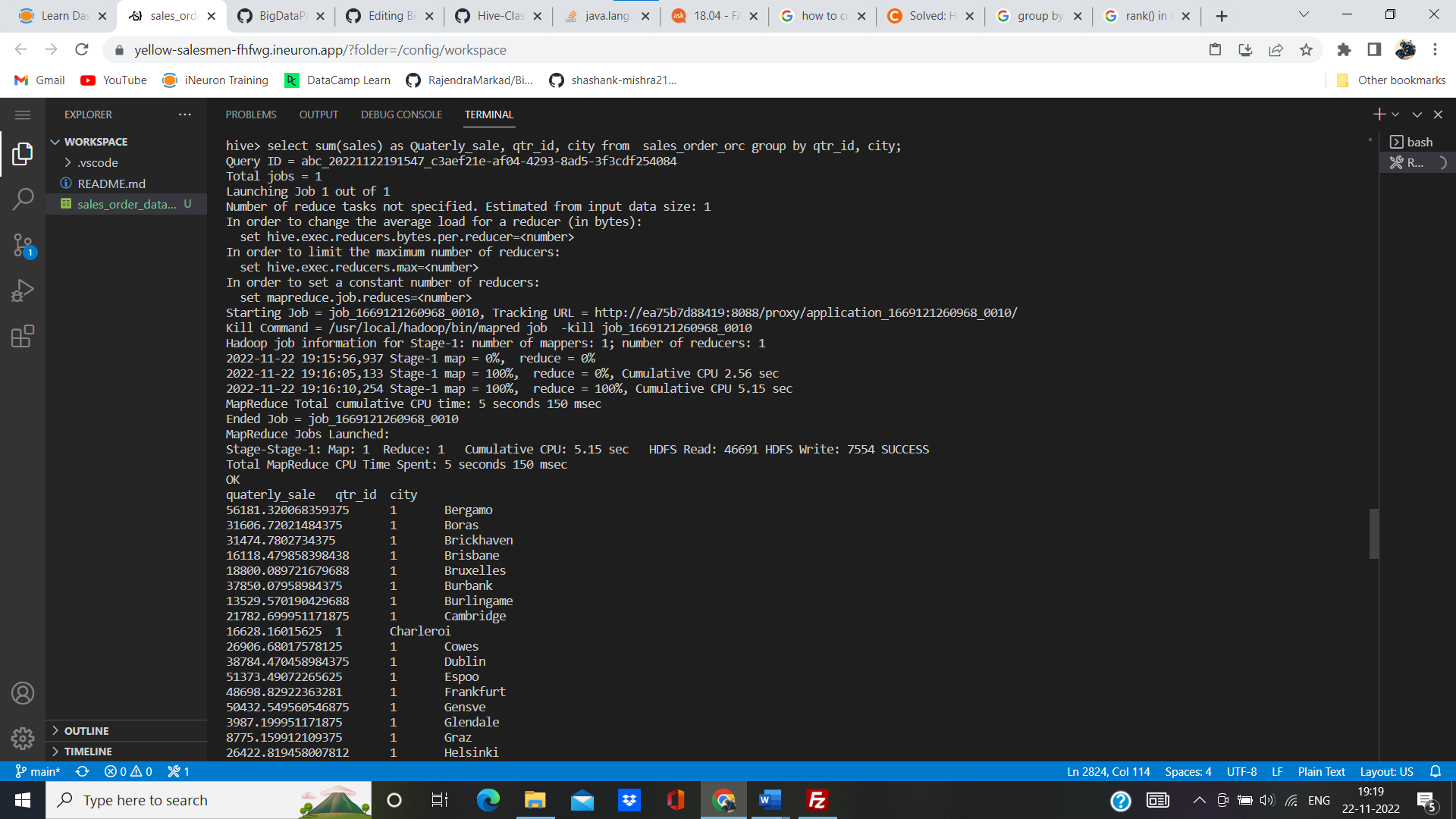
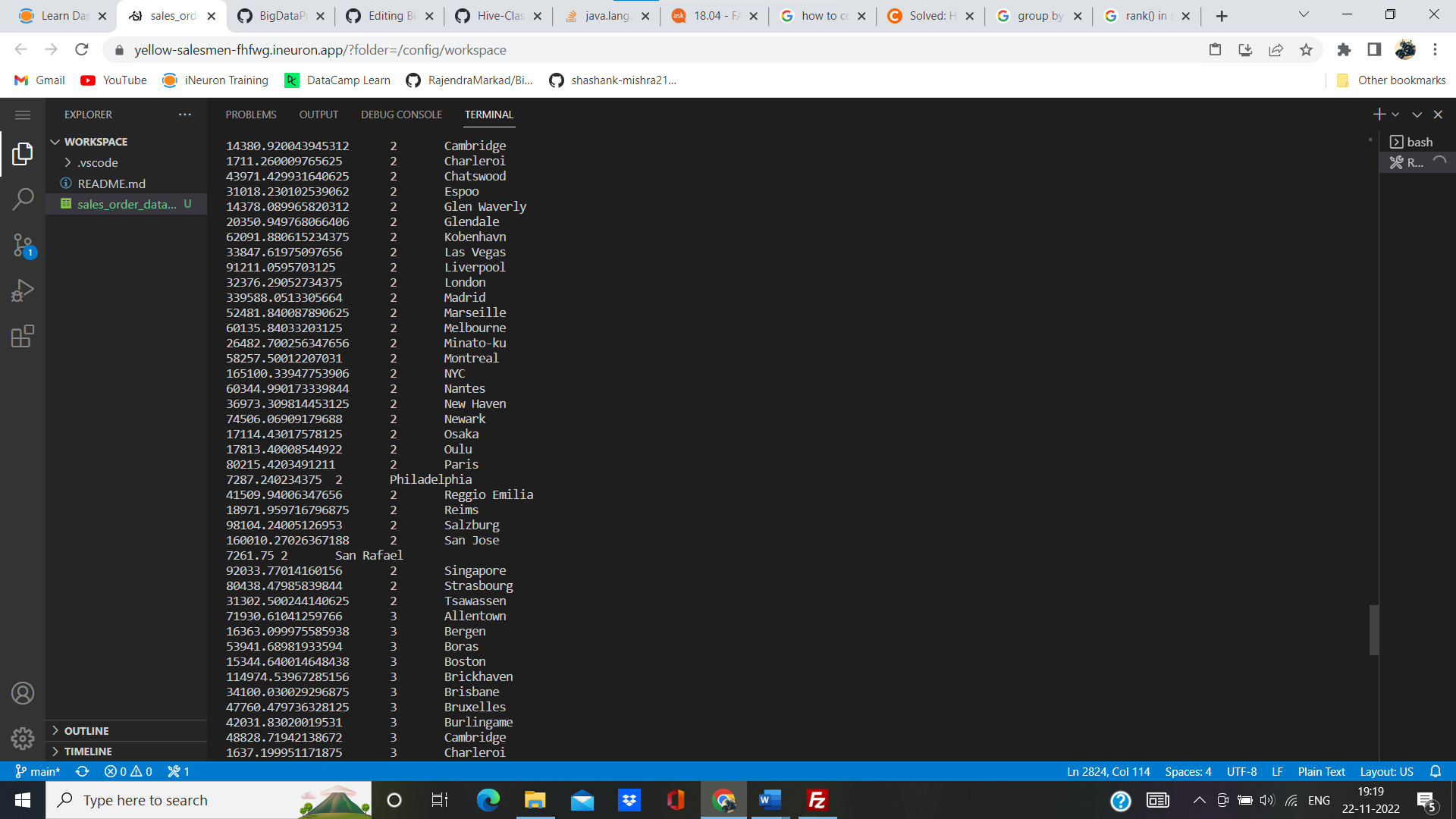
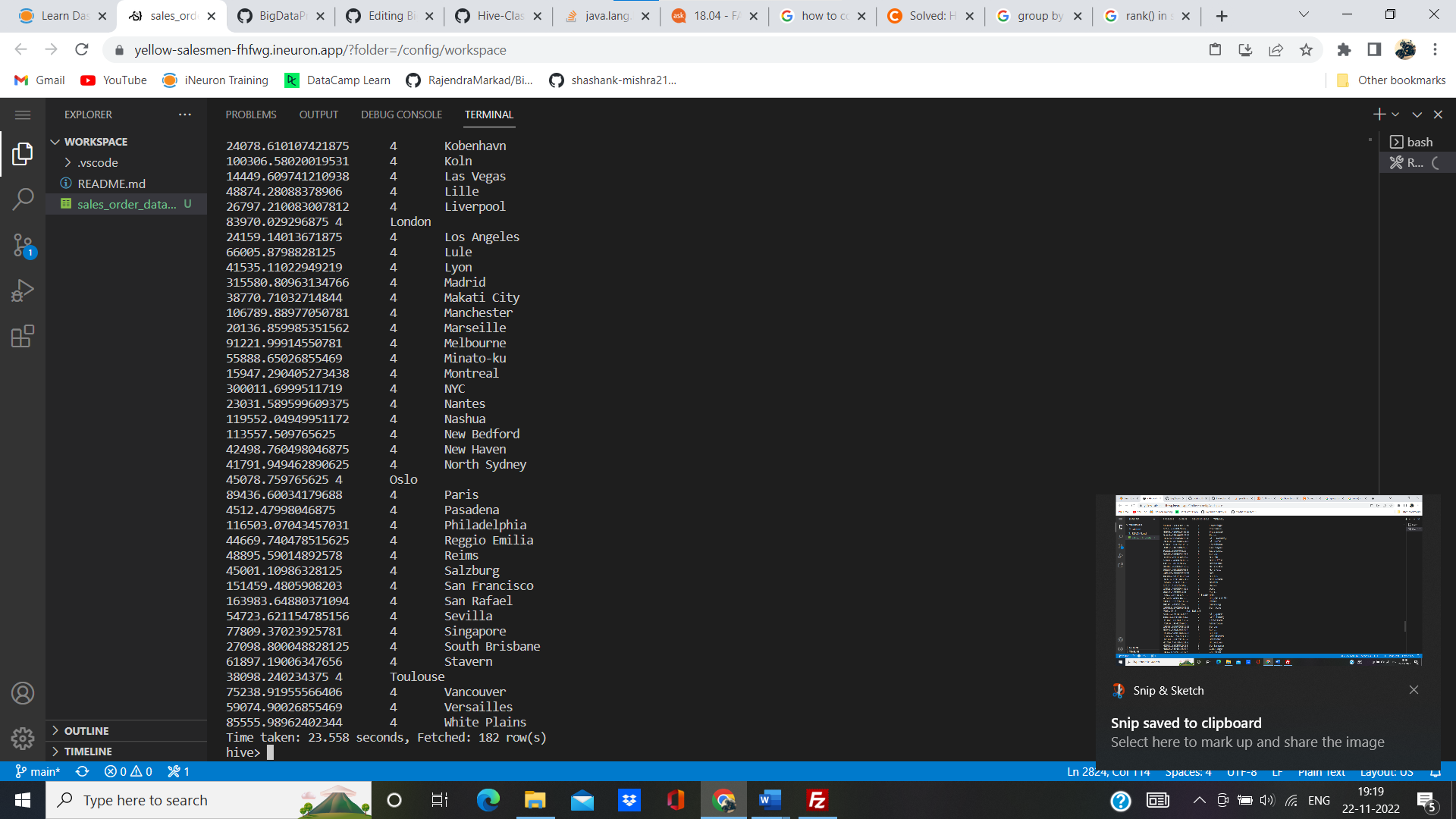
e. In which country sales was maximum and in which country sales was minimum

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

****

f. Calculate quarterly sales for each city.

**Hive>select sum(sales) as Quaterly\_sale, qtr\_id, city from sales\_order\_orc group by qtr\_id, city;**

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

**select month\_id, year\_id, quantityordered from (select 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;**

