**CDAC MUMBAI**

**PG-DBDA SEP 2022 BATCH KHARGHAR**

**MODULE: BIG DATA ANALYTICS**

**DATE : 14TH DEC, 2022**

**MARKS : 40 MARKS**

--------------------------------------------------------------------------------------------------------------------------

- [10

marks]

**Q1.**

**MapReduce**

**Problem Statement**

Here, we have chosen the stock market dataset on which we have performed map-reduce operations. Following is the structure of the data. Kindly Find the solutions to the questions below.

Data Structure

1. Exchange Name

2 Stock symbol

3. Transaction date

4. Opening price of the stock

5. Intra day high price of the stock

6. Intra day low price of the stock

7. Closing price of the stock

8. Total Volume of the stock on the particular day

9. Adjustment Closing price of the stock

Field Separator – comma

**Question 2 : Find all time High price for each stock**

[15 marks]

**Hive**

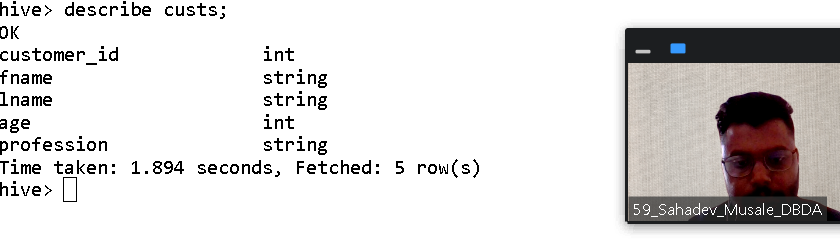
Please find the customer data set.

cust id

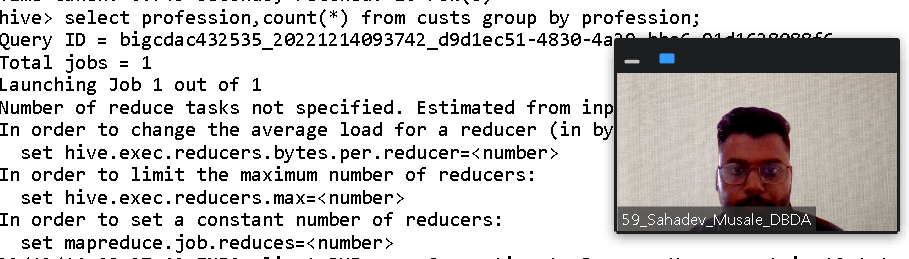
firstname

lastname

age

profession 

;



**1) Write a program to find the count of customers for each profession.**

**select profession,count(\*) from custs group by profession;**

**OK**

**Accountant 199**

**Actor 202**

**Agricultural and food scientist 195**

**Architect 203**

**Artist 175**

**Athlete 196**

**Automotive mechanic 193**

**Carpenter 181**

**Chemist 209**

**Childcare worker 207**

**Civil engineer 193**

**Coach 201**

**Computer hardware engineer 204**

**Computer software engineer 216**

**Computer support specialist 222**

**Dancer 185**

**Designer 205**

**Doctor 197**

**Economist 189**

**Electrical engineer 192**

**Electrician 194**

**Engineering technician 204**

**Environmental scientist 176**

**Farmer 201**

**Financial analyst 198**

**Firefighter 217**

**Human resources assistant 212**

**Judge 196**

**Lawyer 212**

**Librarian 218**

**Loan officer 221**

**Musician 205**

**Nurse 192**

**Pharmacist 213**

**Photographer 222**

**Physicist 201**

**Pilot 213**

**Police officer 210**

**Politician 228**

**Psychologist 194**

**Real estate agent 191**

**Recreation and fitness worker 210**

**Reporter 200**

**Secretary 200**

**Social Worker 1**

**Social worker 212**

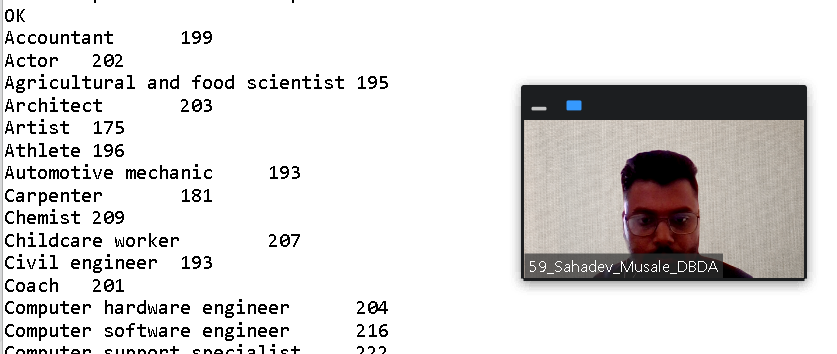
**Statistician 196**

**Teacher 204**

**Therapist 187**

**Veterinarian 208**

**Writer 101**

****

Please find the sales data set.

txn id

txn date

cust id

amount

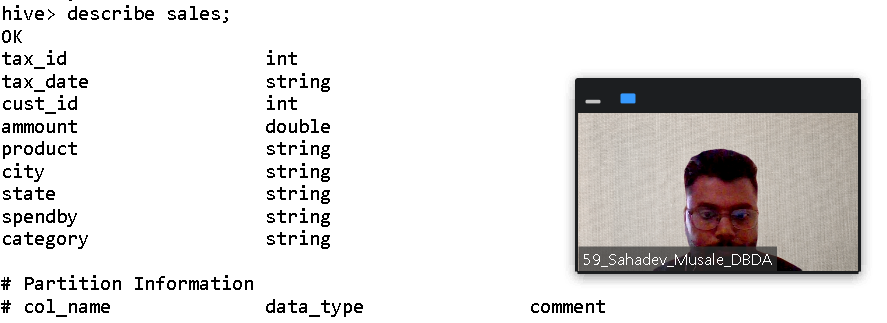
category

product

city

state

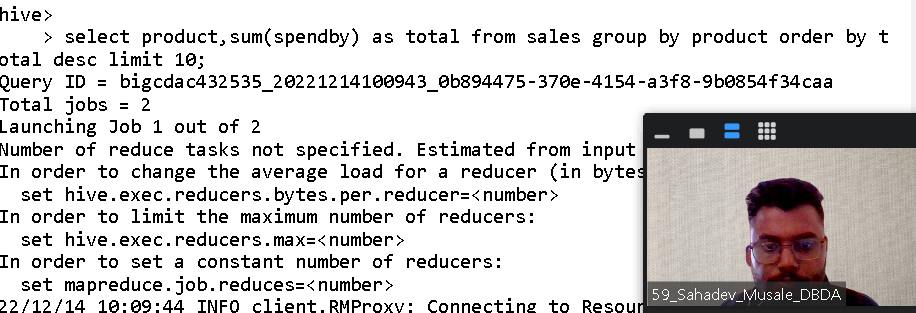
spendby

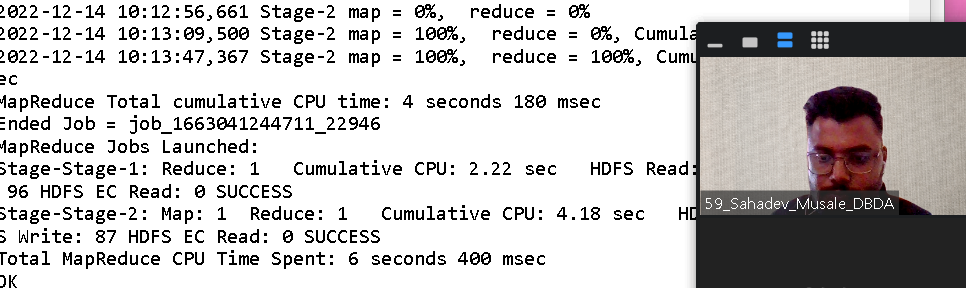


**2) Write a program to find the top 10 products sales wise**

**select product,sum(spendby) as total from sales group by product order by t**

**otal desc limit 10;**

****

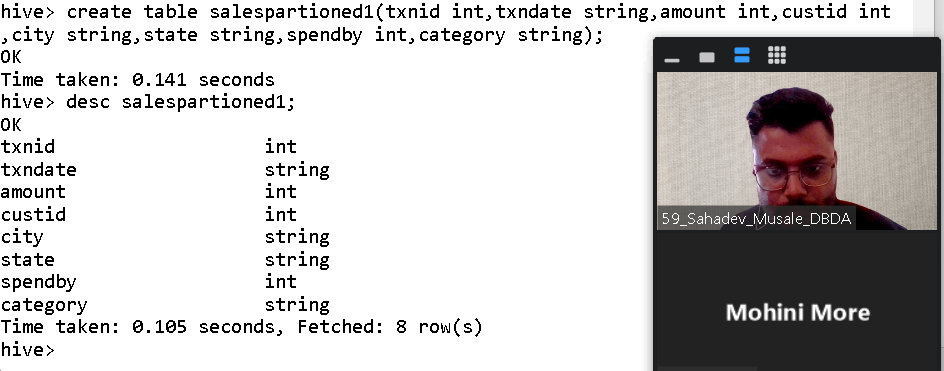
****

**3) Write a program to create partiioned table on category**

**QUESTION 3** [15 marks]

create table salespartioned1(txnid int,txndate string,amount int,custid int

,city string,state string,spendby int,category string);



hive> partitioned by (category string) row format delimited field terminated by '

,' stoared as textfile;

**PySpark**

Please find the AIRLINES data set

Year

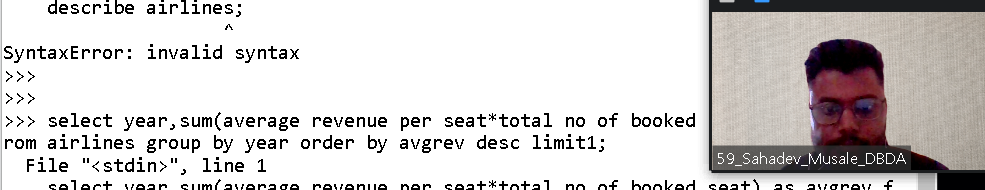
Quarter

Average revenue per seat

Total number of booked seats

**1) What was the highest number of people travelled in which**

**year? 2) Identifying the highest revenue generation for which year**

****

**3) Identifying the highest revenue generation for which year and quarter (Common group)**

*select year,quarter,sum(Average revenue per seat \* Total number of booked seats) as Rev from airlines group by year,quarter order by Rev desc limit 1;*