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

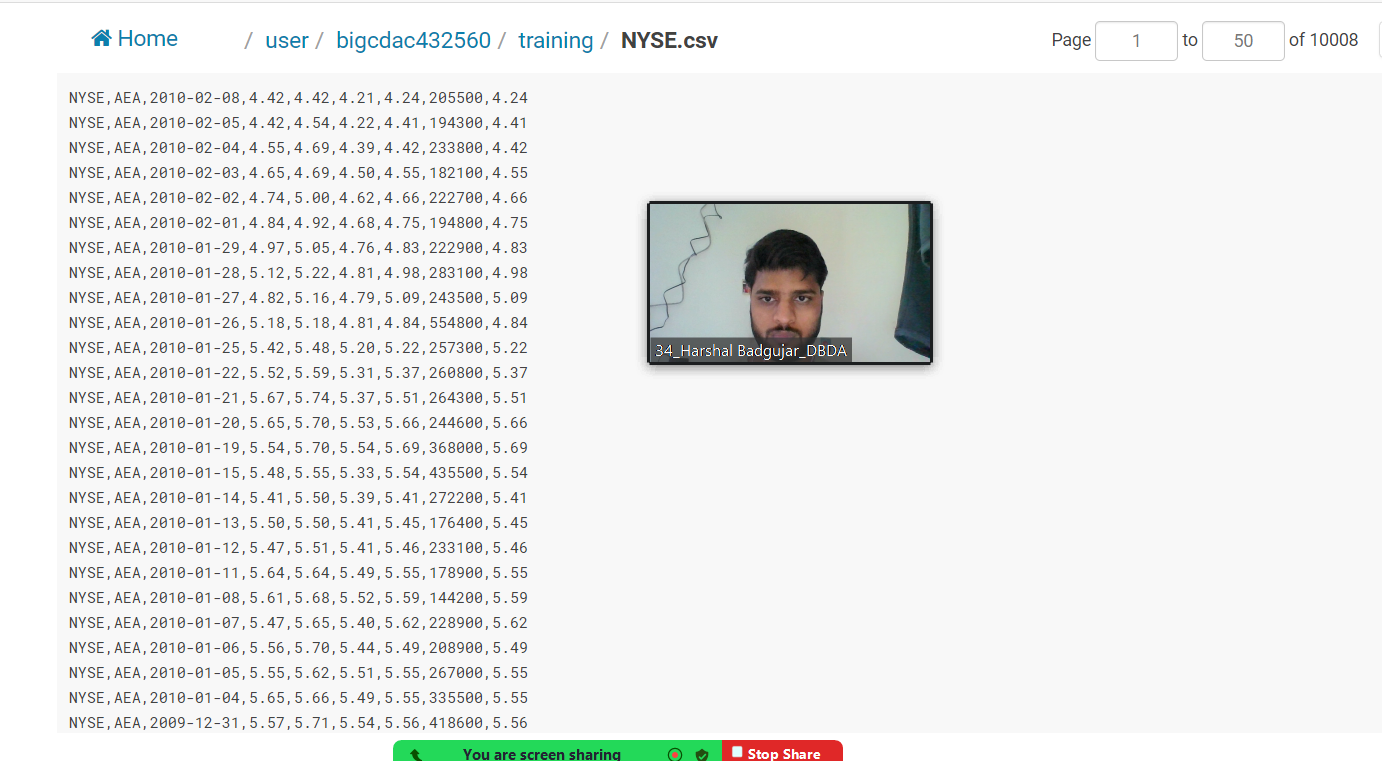
6. Intra day low price of the stock

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

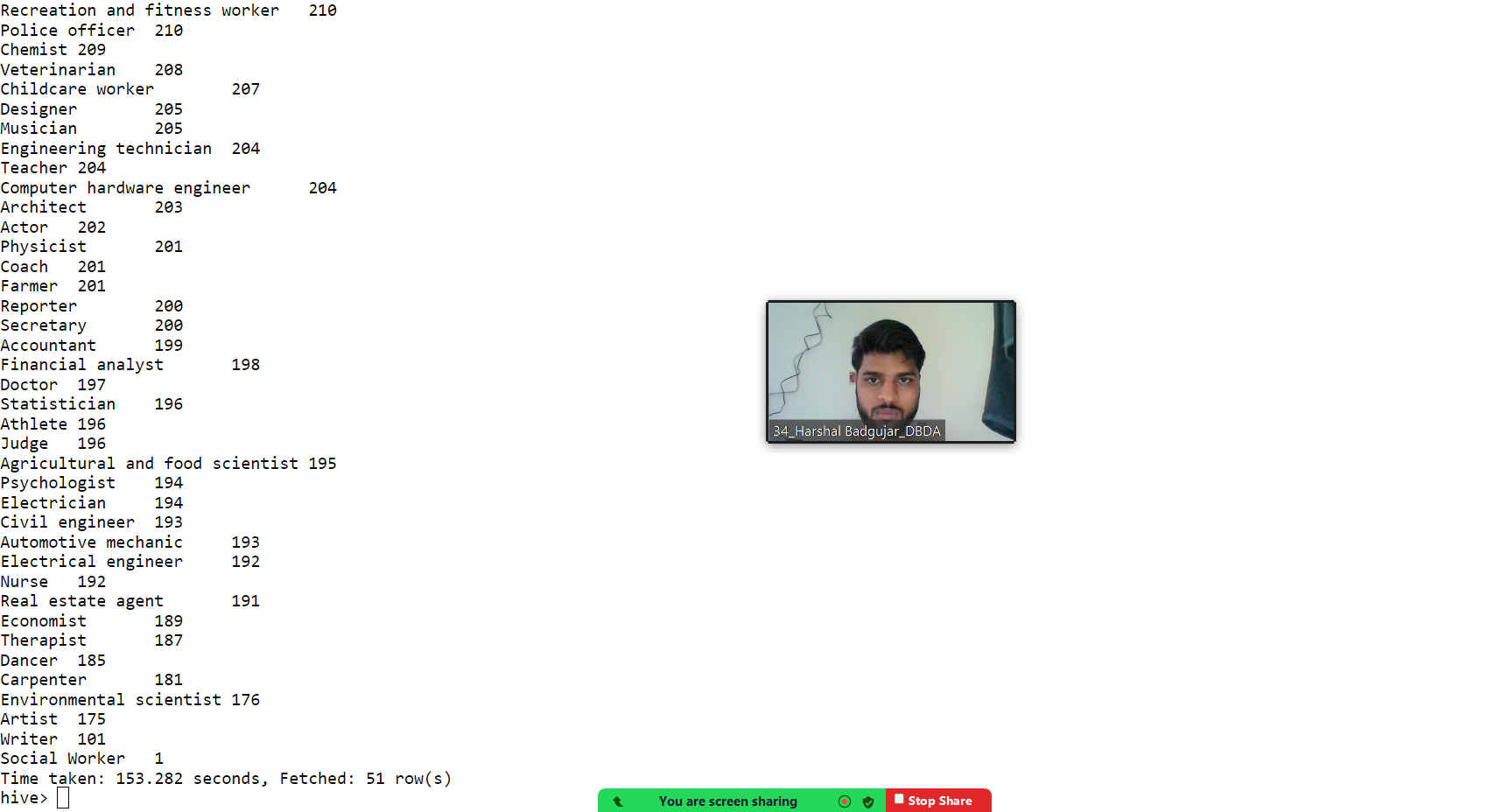
lastname

age

profession

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

select profession,count(profession)as cnt from customer1 group by profession order by cnt desc;



Please find the sales data set.

txn id

txn date

cust id

amount

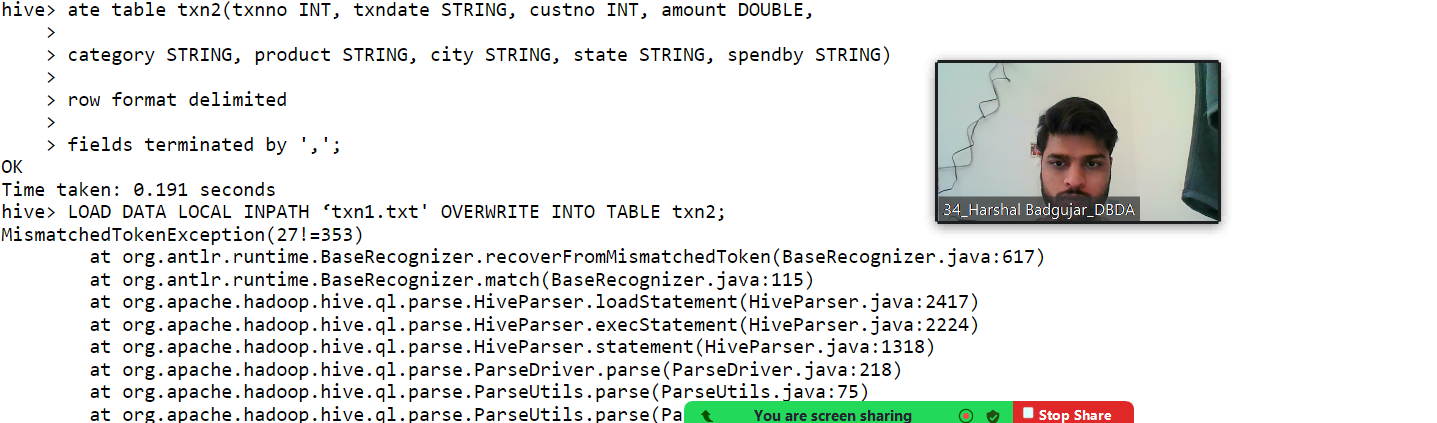
category

product

city

state

spendby



1. Write a program to find the top 10 products sales wise

select product,count(product) as top from txn group by product by order by top desc limit 10;

1. Write a program to create partiioned table on category

QUESTION 3 [15 marks]

PySpark

Please find the AIRLINES data set

Year

Quarter

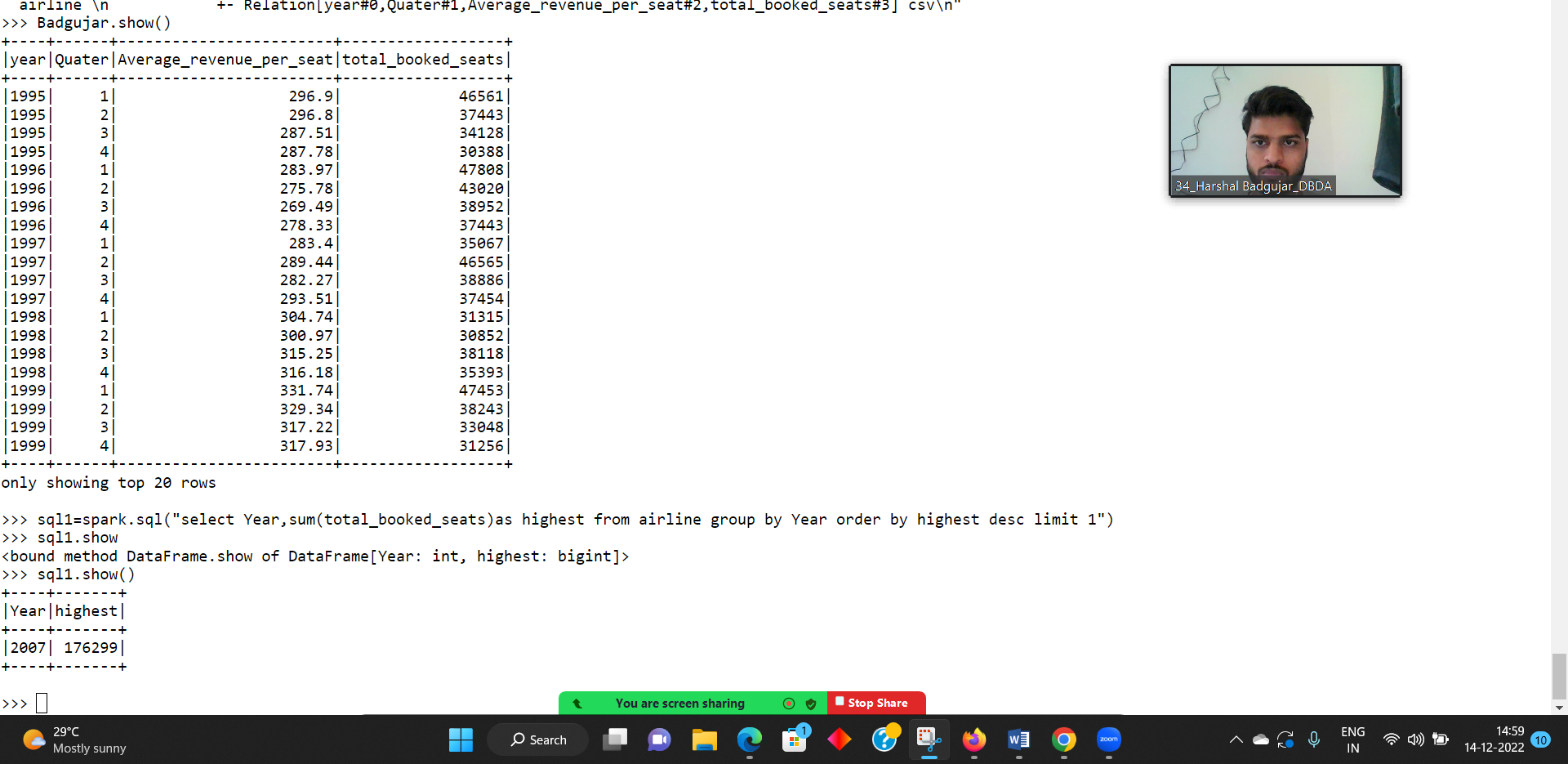
Average revenue per seat

Total number of booked seats

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

year?

sql1=spark.sql("select Year,sum(total\_booked\_seats)as highest from airline group by Year order by highest desc limit 1")



2) Identifying the highest revenue generation for which year

sql2=spark.sql("select Year,sum(Average\_revenue\_per\_seat\*total\_booked\_seats)/1000000 as Avgs\_in\_million from airline group by Year or

der by Avgs\_in\_million desc limit 1")



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

group)

sql3=spark.sql("select Year,Quater,sum((Average\_revenue\_per\_seat\*total\_booked\_seats))/1000000 as Avgs\_in\_million from airline group b

y Year,Quater having Year=2014 order by Avgs\_in\_million desc")

