

CREATE DATABASE exam78

USE exam78

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

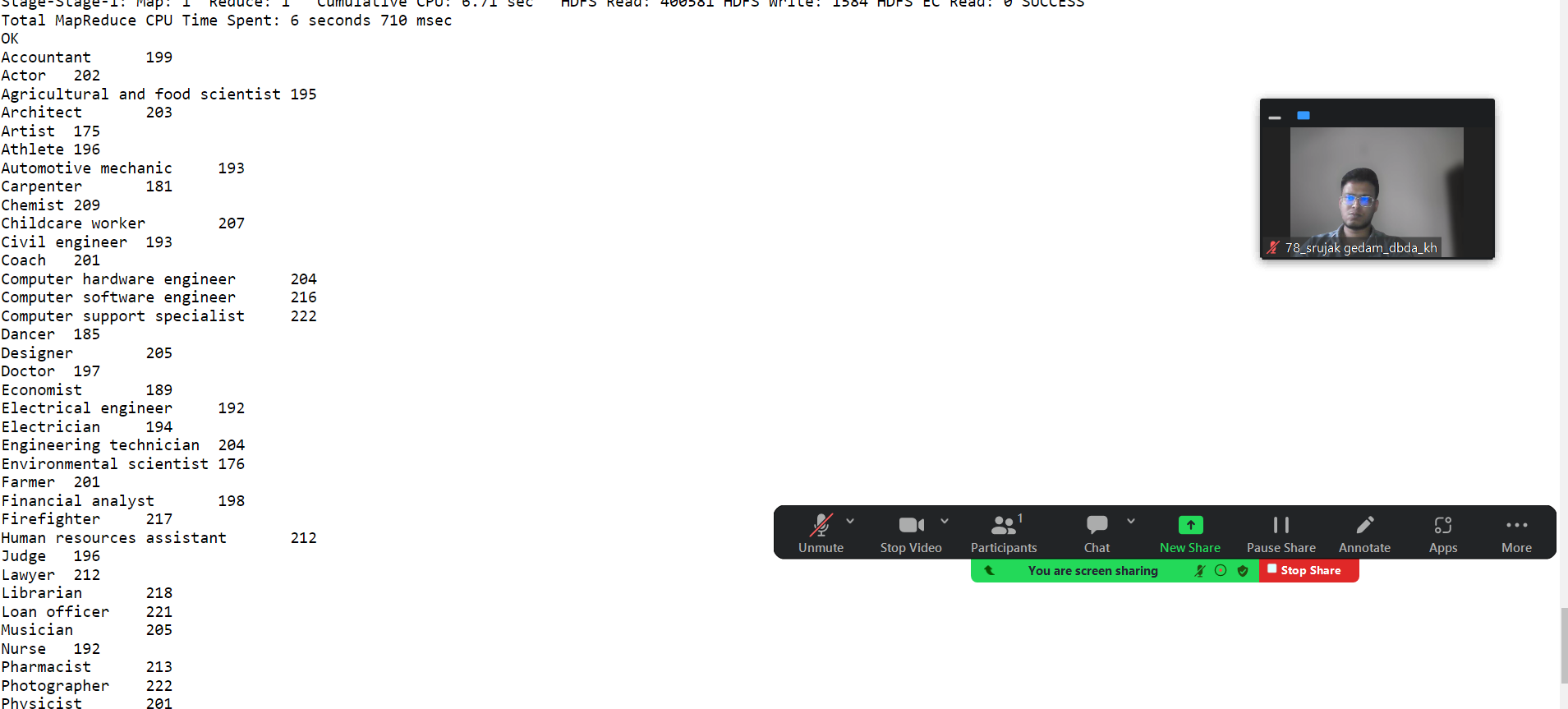
create table customer(custno INT, firstname STRING, lastname STRING, age INT, profession STRING)

row format delimited

fields terminated by ','

stored as textfile;

Select profession, count(custno) from customer group by profession;



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

create table txnrec(txnid INT, txndate STRING, custid INT, amount DOUBLE,

category STRING, product STRING, city STRING, state STRING, spendby STRING)

row format delimited

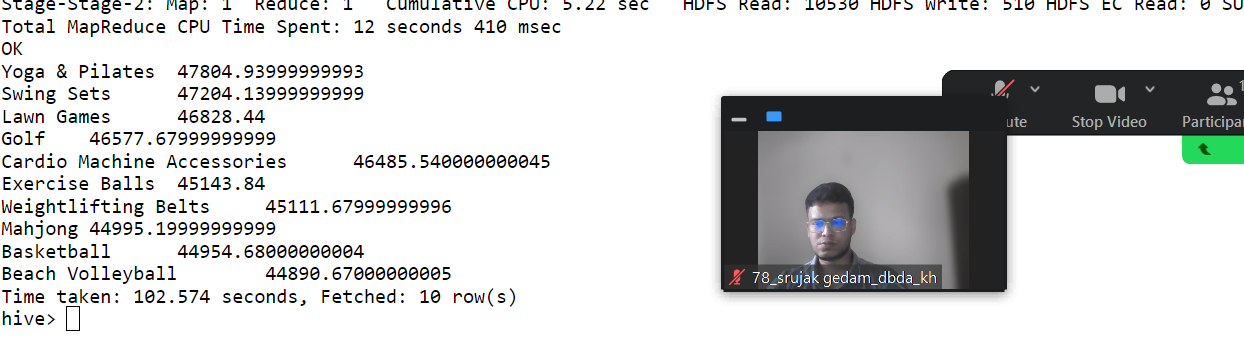
fields terminated by ','

stored as textfile

location '/user/bigdatamind4385/sales';

LOAD DATA LOCAL INPATH 'txns1.txt' OVERWRITE INTO TABLE txnrec;

Select product,sum(amount) as total from txnrec group by product order by total desc limit 10;



3) Write a program to create partiioned table on category

create table txnrecsByCat2(txnno INT, txndate STRING, custno INT, amount DOUBLE,

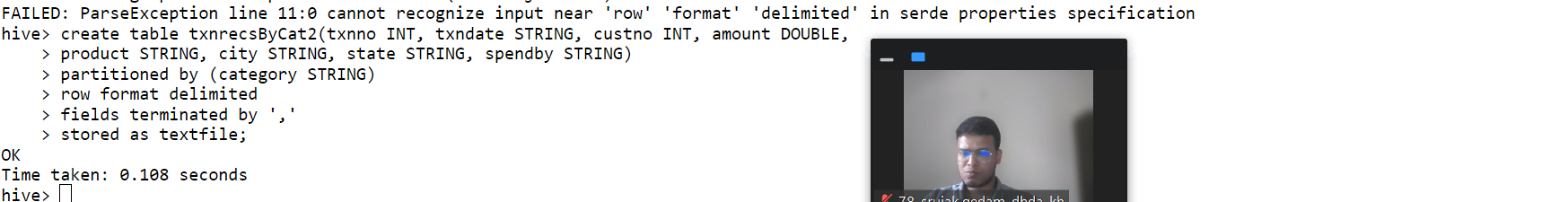
product STRING, city STRING, state STRING, spendby STRING)

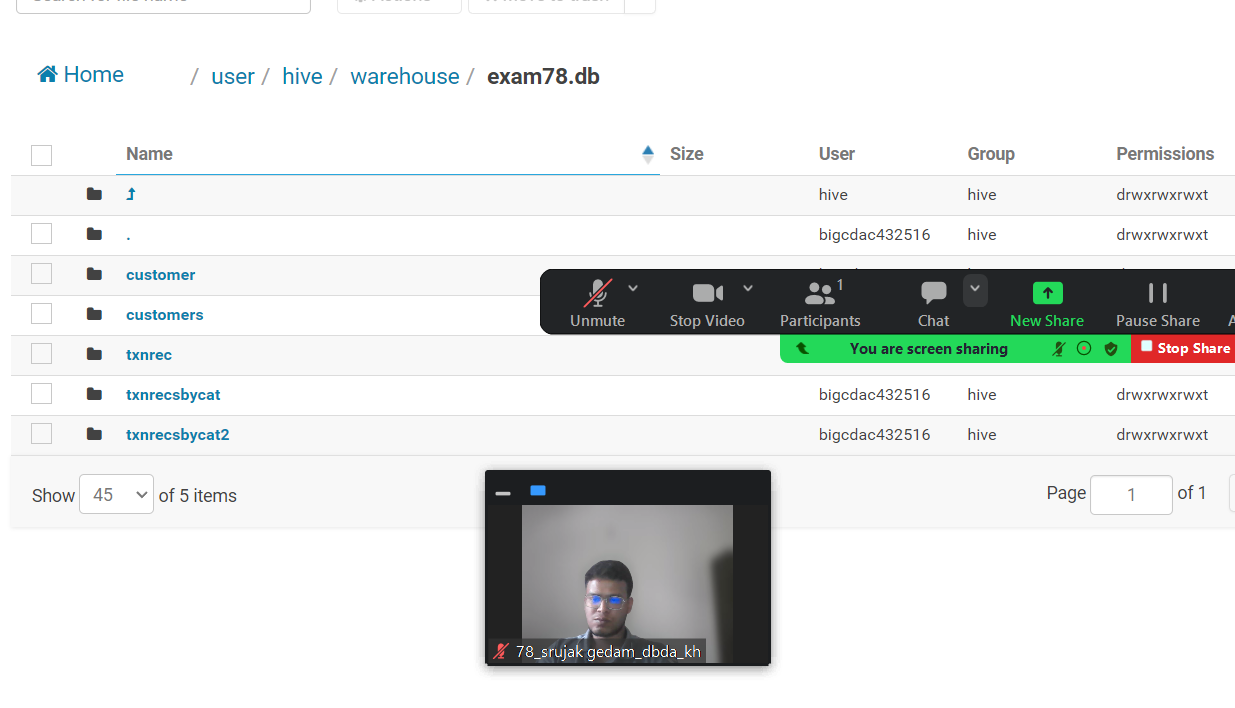
partitioned by (category STRING)

row format delimited

fields terminated by ','

stored as textfile;





3) 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

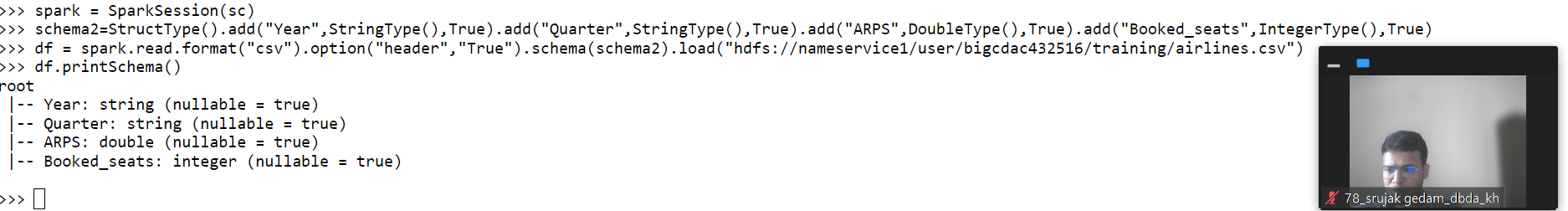
1) What was the highest number of people travelled in which year?

2) Identifying the highest revenue generation for which year

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

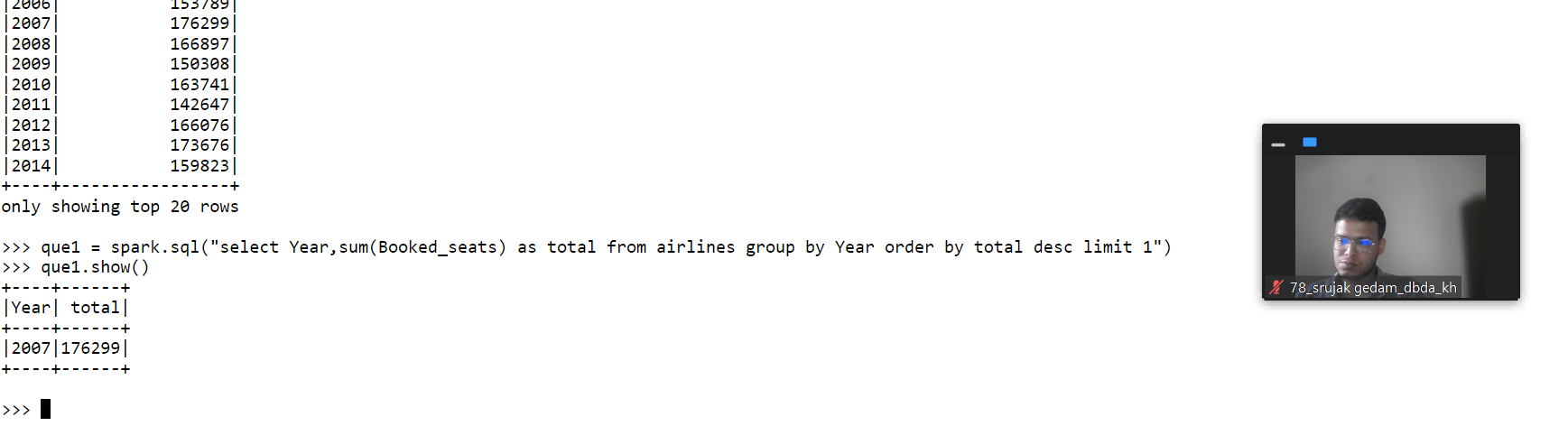
schema2=StructType().add("Year",StringType(),True).add("Quarter",StringType(),True).add("ARPS",DoubleType(),True).add("Booked\_seats",IntegerType(),True)

df = spark.read.format("csv").option("header","True").schema(schema2).load("hdfs://nameservice1/user/bigcdac432516/training/airlines.csv")



df.registerTempTable("airlines")

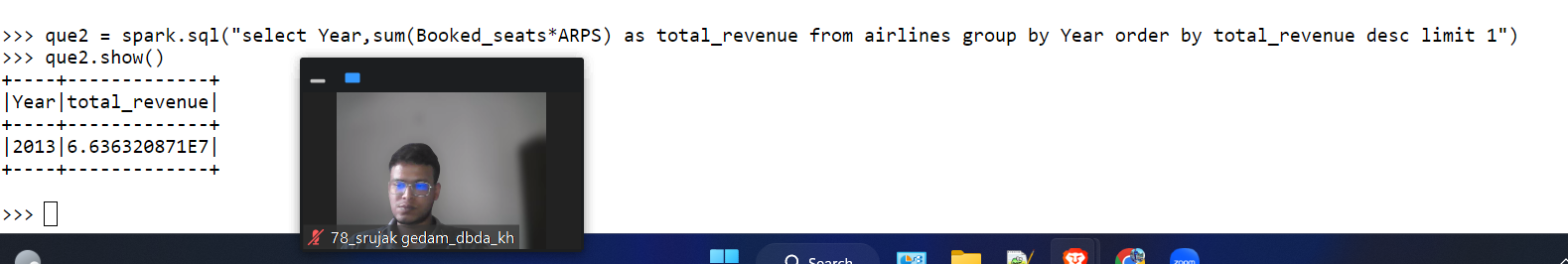
1.What was the highest number of people travelled in which year?



>>> que1 = spark.sql("select Year,sum(Booked\_seats) as total from airlines group by Year order by total desc limit 1")

>>> que1.show()

2) Identifying the highest revenue generation for which year



>>> que2 = spark.sql("select Year,sum(Booked\_seats\*ARPS) as total\_revenue from airlines group by Year order by total\_revenue desc limit 1")

>>> que2.show()

+----+-------------+

|Year|total\_revenue|

+----+-------------+

|2013|6.636320871E7|

+----+-------------+

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

df3=spark.sql("Select year,quarter,round(sum(ARPS\*BookedSeats)/1000000,2) as totalRevenue from airlines group by year,quarter order by totalRevenue desc limit 1");