Name:Keshav Nivrutti Yawale

Prn.number:220940325036

github:

3) Write a program to create partiioned table on category

QUESTION 3 [15 marks]

PySpark

Please find the AIRLINES data set

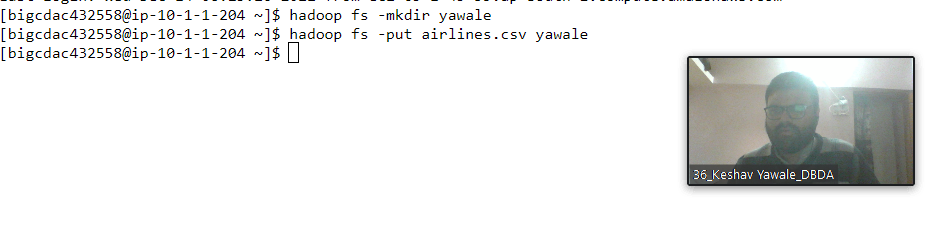
Year

Quarter

Average revenue per seat

Total number of booked seats

Answer:



from pyspark.sql.types import StructType,StringType,IntegerType,DoubleType,LongType

sch= StructType().add("Year",IntegerType(),True).add("Quarter",IntegerType(),True).add("Average\_revenue\_per",DoubleType(),True).add("total\_booked\_seats",IntegerType(),True)

air=spark.read.format("csv").option("header","True").schema(sch).load("hdfs://nameservice1//user/bigcdac432558/yawale/airlines.csv")

air.registerTempTable("airline")

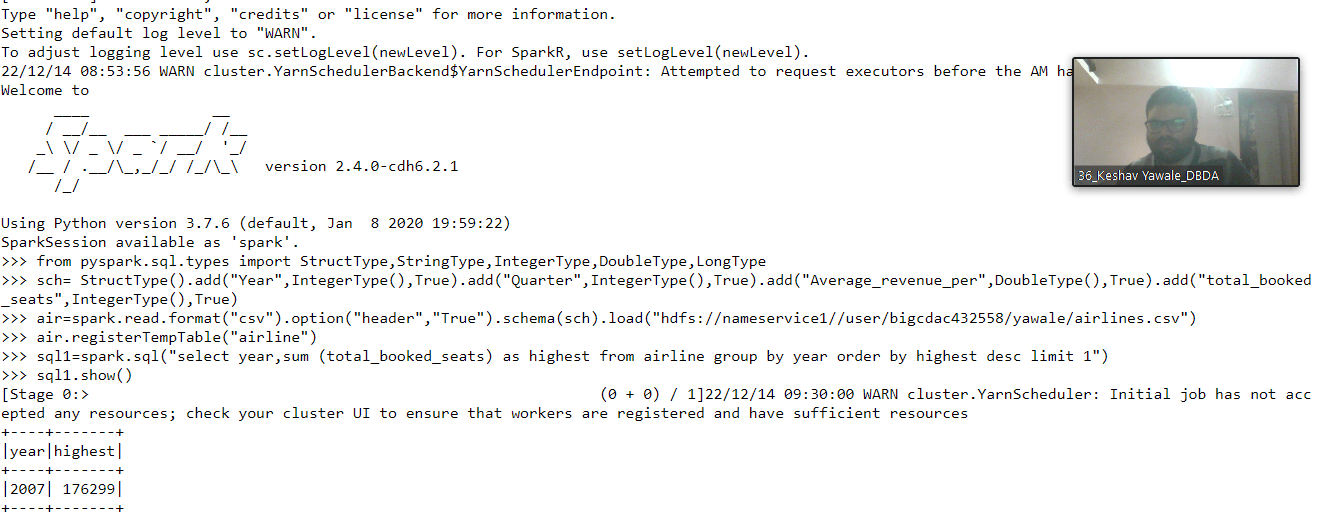
1) What was the highest number of people travelled in which

Year?

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

sql1.show()

**output:**

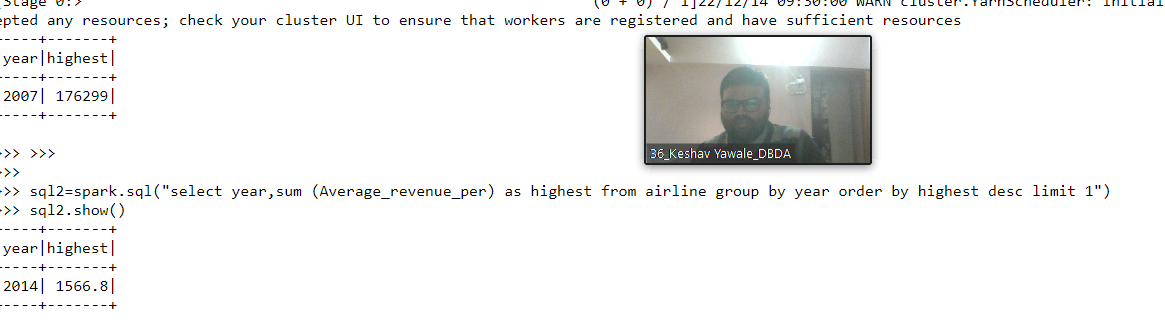


2) Identifying the highest revenue generation for which year

sql2=spark.sql("select year,sum (Average\_revenue\_per) as highest from airline group by year order by highest desc limit 1")

sql2.show()

**output:**

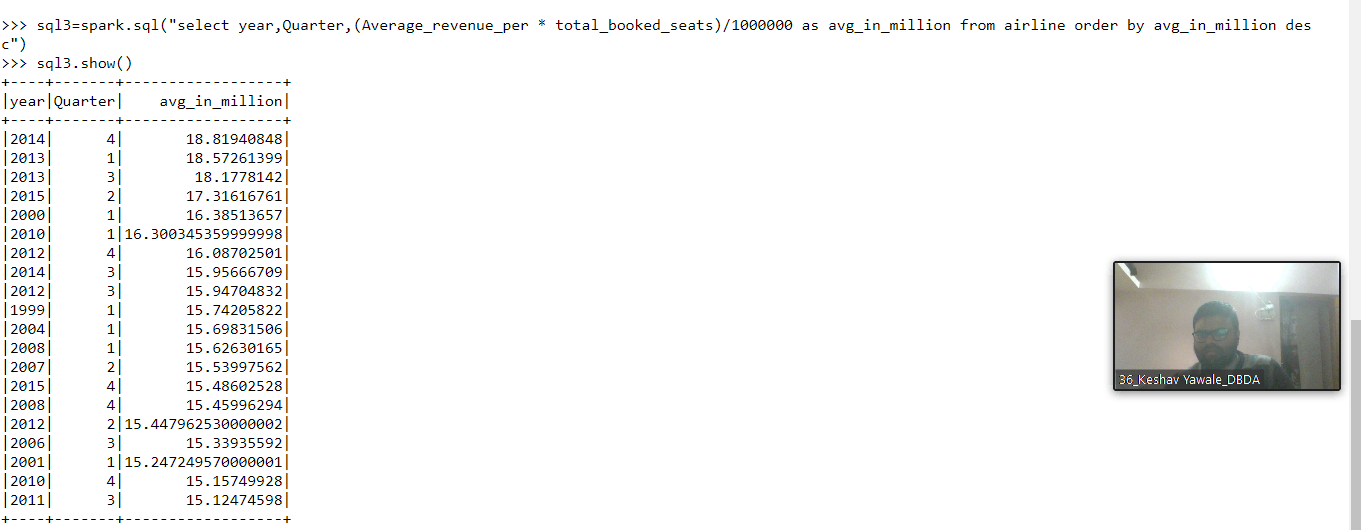


3) Identifying the highest revenue generation for which year and quarter (Common

group)

sql3=spark.sql("select year,Quarter,(Average\_revenue\_per \* total\_booked\_seats)/1000000 as avg\_in\_million from airline order by avg\_in\_million desc")

sql3.show()



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

Answer:

Hive

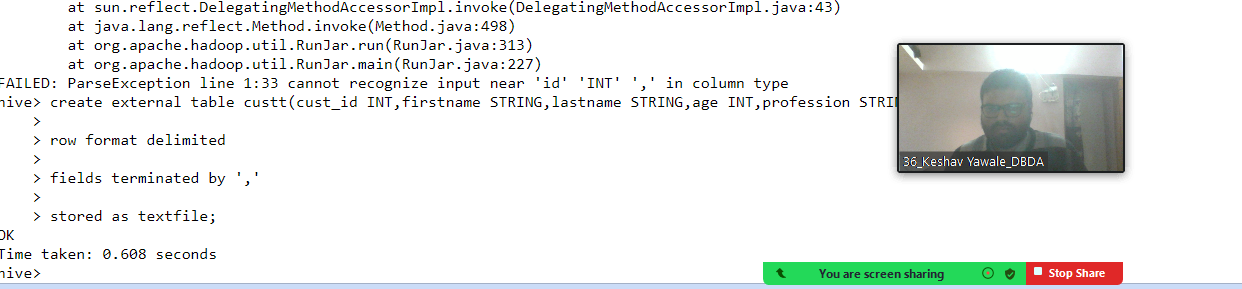
Use keshav;

create external table custt(cust\_id INT,firstname STRING,lastname STRING,age INT,profession STRING)

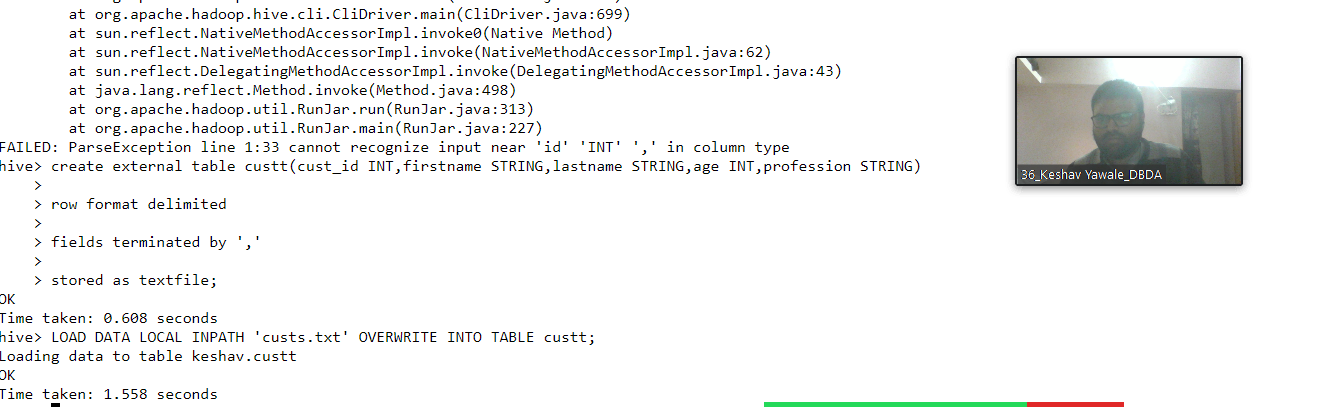
row format delimited

fields terminated by ','

stored as textfile



LOAD DATA LOCAL INPATH 'custs.txt' OVERWRITE INTO TABLE custt;



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

Please find the sales data set.

txn id

txn date

cust id

amount

category

product

city

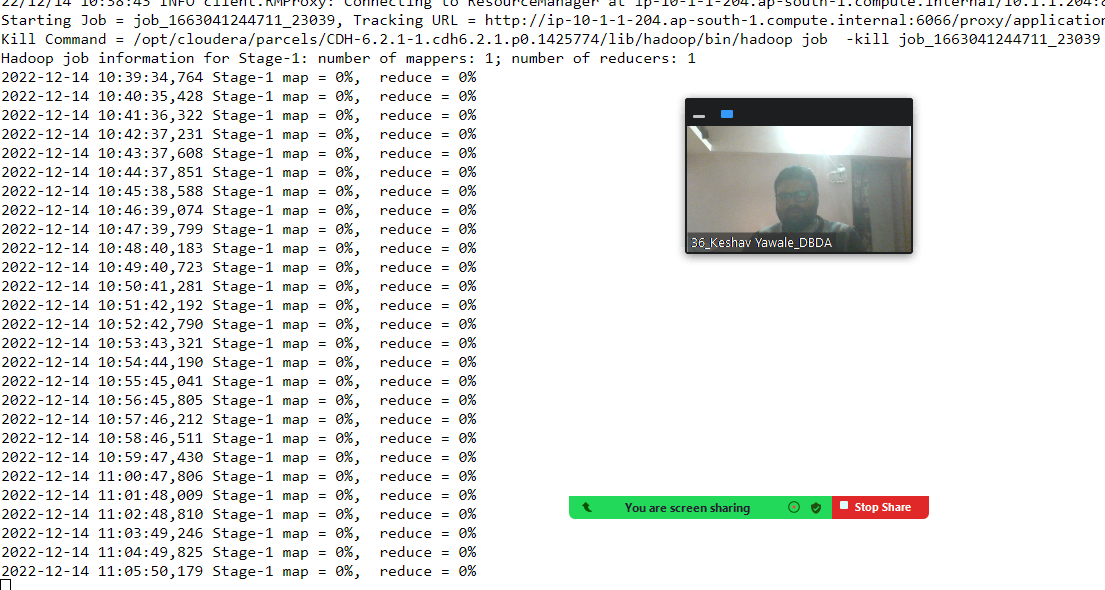
state

Spendby

answer:

Select profession, count(\*) as headcount from custt group by profession order by headcount;

Output: (prosessing is taking too much time)



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

Answer:

create table txnrecords (txnno INT, txndate STRING,custno INT,amount DOUBLE,category STRING,product STRING,city STRING,state STRING,spendy STRING )

row format delimited

fields terminated by ','

stored as textfile;

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

Select product,round(sum(amount),2) from txnrecords group by product order by amount desc limit 10;

Output: (prosessing is taking too much time)

