Question 2 : Find all time High price for each stock [15 marks] Hive

Please find the customer data set.

cust\_id int

firstname string

lastname string

age int

Profession string

**create table customer(**

**cust\_id int,**

**firstname string,**

**lastname string,**

**age int,**

**Profession string)**

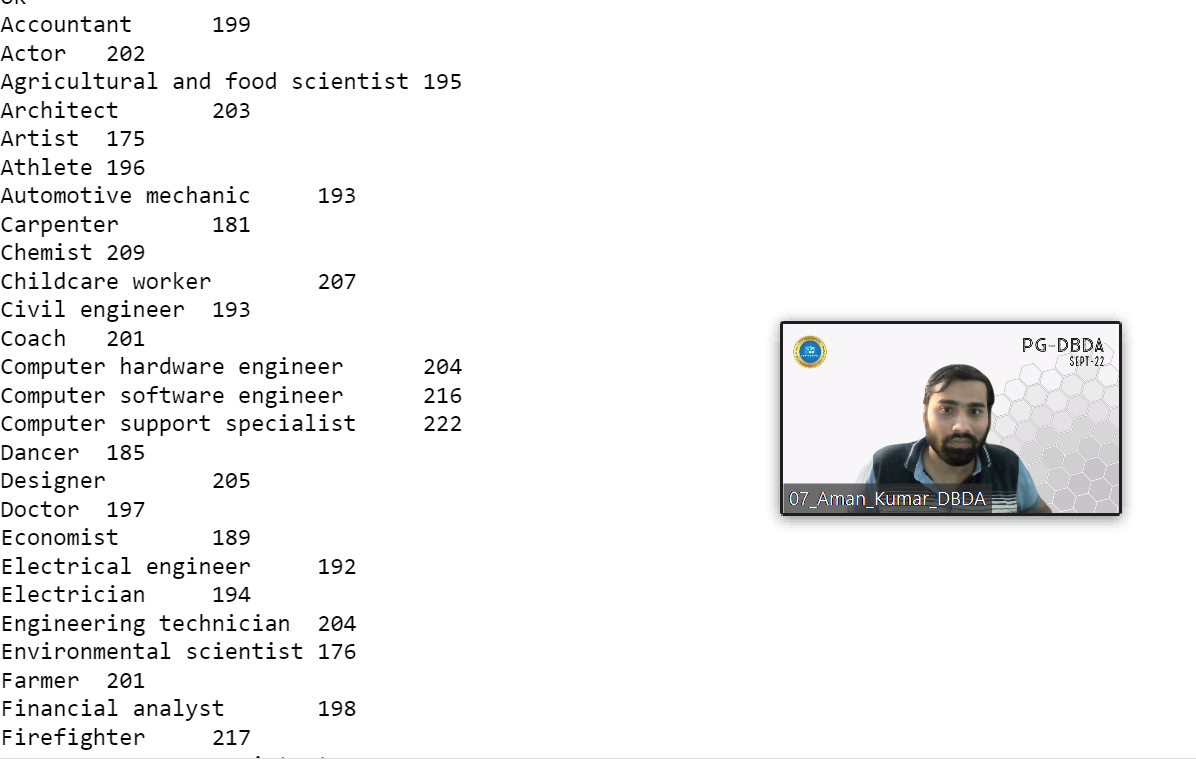
**row format delimited fields terminated by ‘,’**

**stored as textfile;**

**Load data local inpath ‘custs.txt’ overwrite into table customer;**

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

**select Profession , count(\*) from customer group by Profession;**

****

Please find the

sales data set.

txn\_id int

txn\_date string

cust\_id int

amount double

category string

product string

city string

state string

spendby string

**create table tax(**

**txn\_id int,**

**txn\_date string,**

**cust\_id int,**

**amount double,**

**category string,**

**product string,**

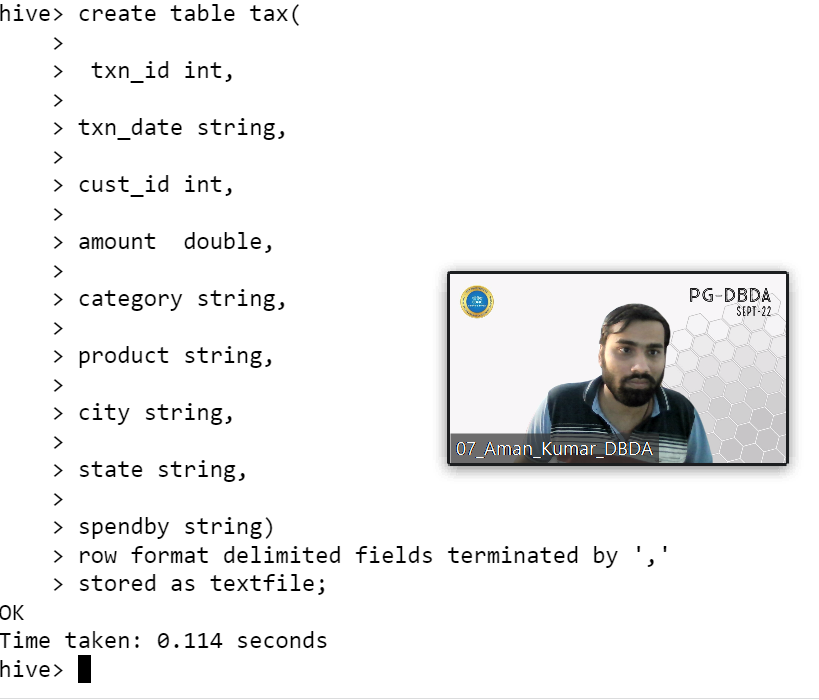
**city string,**

**state string,**

**spendby string)**

**row format delimited fields terminated by ‘,’**

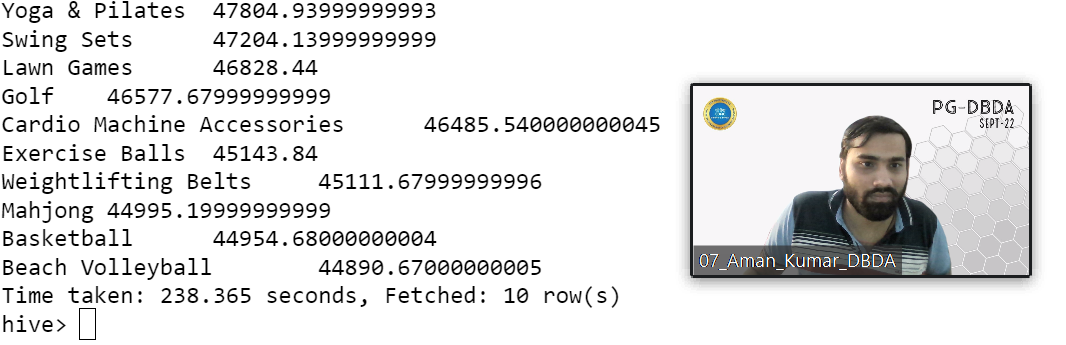
**stored as textfile;**

****

**Load data local inpath ‘txns1.txt’ overwrite into table tax;**

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

**select product , sum(amount) as total\_sales from tax group by product order by total\_sales desc limit 10;**

****

3) Write a program to create partiioned table on category

**create table tax\_partition(**

**txn\_id int,**

**txn\_date string,**

**cust\_id int,**

**amount double,**

**product string,**

**city string,**

**state string,**

**spendby string)**

**partitioned by (category string)**

**row format delimited fields terminated by ‘,’**

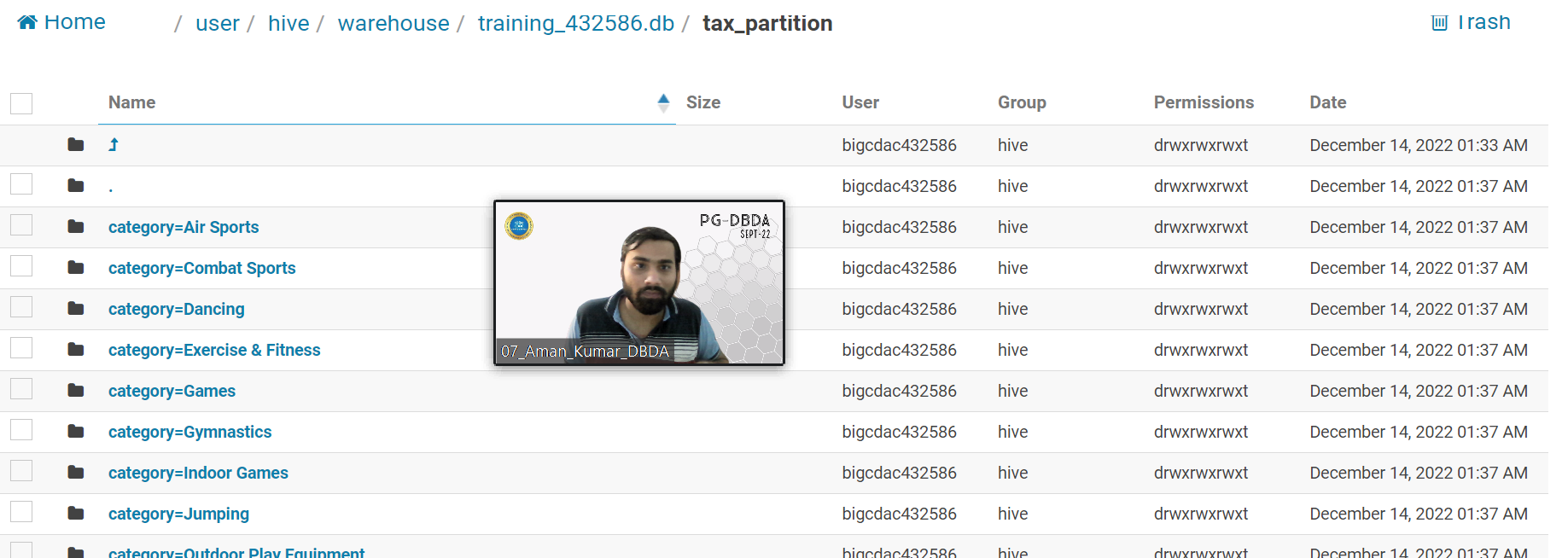
**stored as textfile;**

**Set hive.exec.dynamic.partition.mode = nonstrict;**

**Set hive.exec.dynamic.partition.mode = true;**

**From tax t insert overwrite table tax\_partition partition(category)**

**select t.txn\_id,t.txn\_date,t.cust\_id,t.amount,t.product,t.city,t.state,t.spendby,t.category**

**distribute by category;**

QUESTION 3 [15 marks]

PySpark

Please find the AIRLINES data set

Year

Quarter

Average

revenue\_per\_seat

Total\_number\_of\_booked\_seats

**from pyspark.context import SparkContext**

**from pyspark.sql.session import SparkSession**

**sc = SparkContext('local')**

**spark = SparkSession(sc)**

**spark = SparkSession.builder.config('spark.some.config.option','some-value').getOrCreate()**

**from pyspark.sql.types import StructType,IntegerType,StringType,FloatType**

**schema101 = StructType().add("Year",StringType(),True).add("Quarter",StringType(),True).add("revenue\_per\_seat",DoubleType(),True).add("Total\_number\_of\_booked\_seats",IntegerType(),True)**

**df.printSchema()**

**root**

**|-- Year: string (nullable = true)**

**|-- Quarter: string (nullable = true)**

**|-- Average revenue per seat: string (nullable = true)**

**|-- total no. of booked seats: string (nullable = true)**

**df = spark.read.format(“csv”).option(“header”,True).schema(schema101)load(“hdfs://nameservice1/user/bigcdac432586/training/airlines.csv”)**

**df.show()**

**+----+-------+------------------------+-------------------------+**

**|Year|Quarter|Average revenue per seat|total no. of booked seats|**

**+----+-------+------------------------+-------------------------+**

**|1995| 1| 296.9| 46561|**

**|1995| 2| 296.8| 37443|**

**|1995| 3| 287.51| 34128|**

**|1995| 4| 287.78| 30388|**

**|1996| 1| 283.97| 47808|**

**|1996| 2| 275.78| 43020|**

**|1996| 3| 269.49| 38952|**

**|1996| 4| 278.33| 37443|**

**|1997| 1| 283.4| 35067|**

**|1997| 2| 289.44| 46565|**

**|1997| 3| 282.27| 38886|**

**|1997| 4| 293.51| 37454|**

**|1998| 1| 304.74| 31315|**

**|1998| 2| 300.97| 30852|**

**|1998| 3| 315.25| 38118|**

**|1998| 4| 316.18| 35393|**

**|1999| 1| 331.74| 47453|**

**|1999| 2| 329.34| 38243|**

**|1999| 3| 317.22| 33048|**

**|1999| 4| 317.93| 31256|**

**+----+-------+------------------------+-------------------------+**

**only showing top 20 rows**

**df.registerTempTable(“airlines”)**

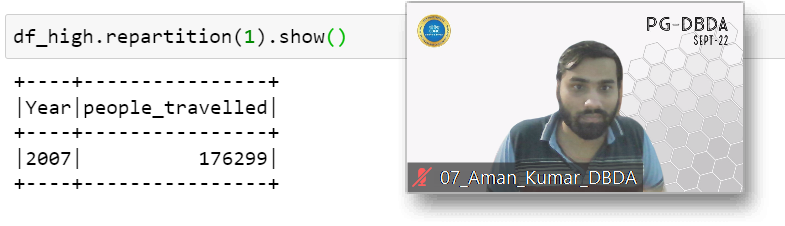
1) What was the highest number of people travelled in which year?

**df\_high= spark.sql(“select Year,sum(Total\_number\_of\_booked\_seats) as people\_travelled from airlines group by Year order by people\_travelled desc limit 1”)**

**df\_high.rdd.getNumPartitions()**

**1**

**df\_high.repartition(1).show()**

****

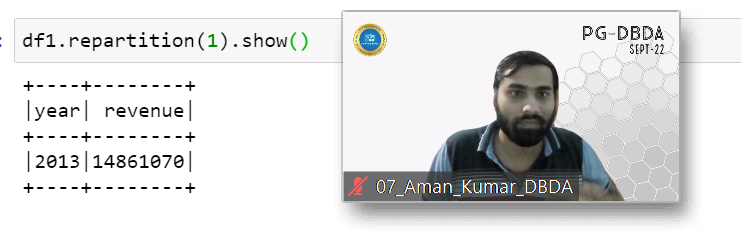
2) Identifying the highest revenue generation for which year

**df1=spark.sql(“select year,sum( revenue\_per\_seat\*Total\_number\_of\_booked\_seats) as revenue from airlines group by year order by revenue desc limit 1”)**

**df1.rdd.getNumPartitions()**

**1**

**df1.repartition(1).show()**

****

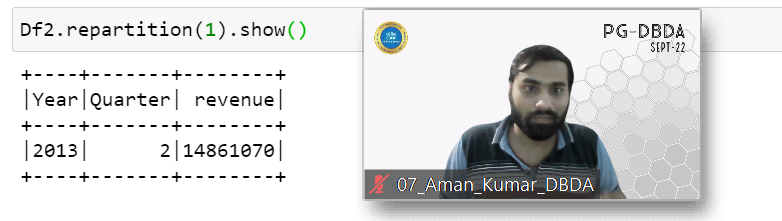
3) Identifying the highest revenue generation for which year and quarter (Common group)

**Df2 = spark.sql(“select Year,Quarter,sum( revenue\_per\_seat\*Total\_number\_of\_booked\_seats) as revenue from airlines group by Year,Quarter order by revenue desc limit 1”);**

**Df2.rdd.getNumPartitions()**

**1**

**Df2.repartition(1).show()**

****