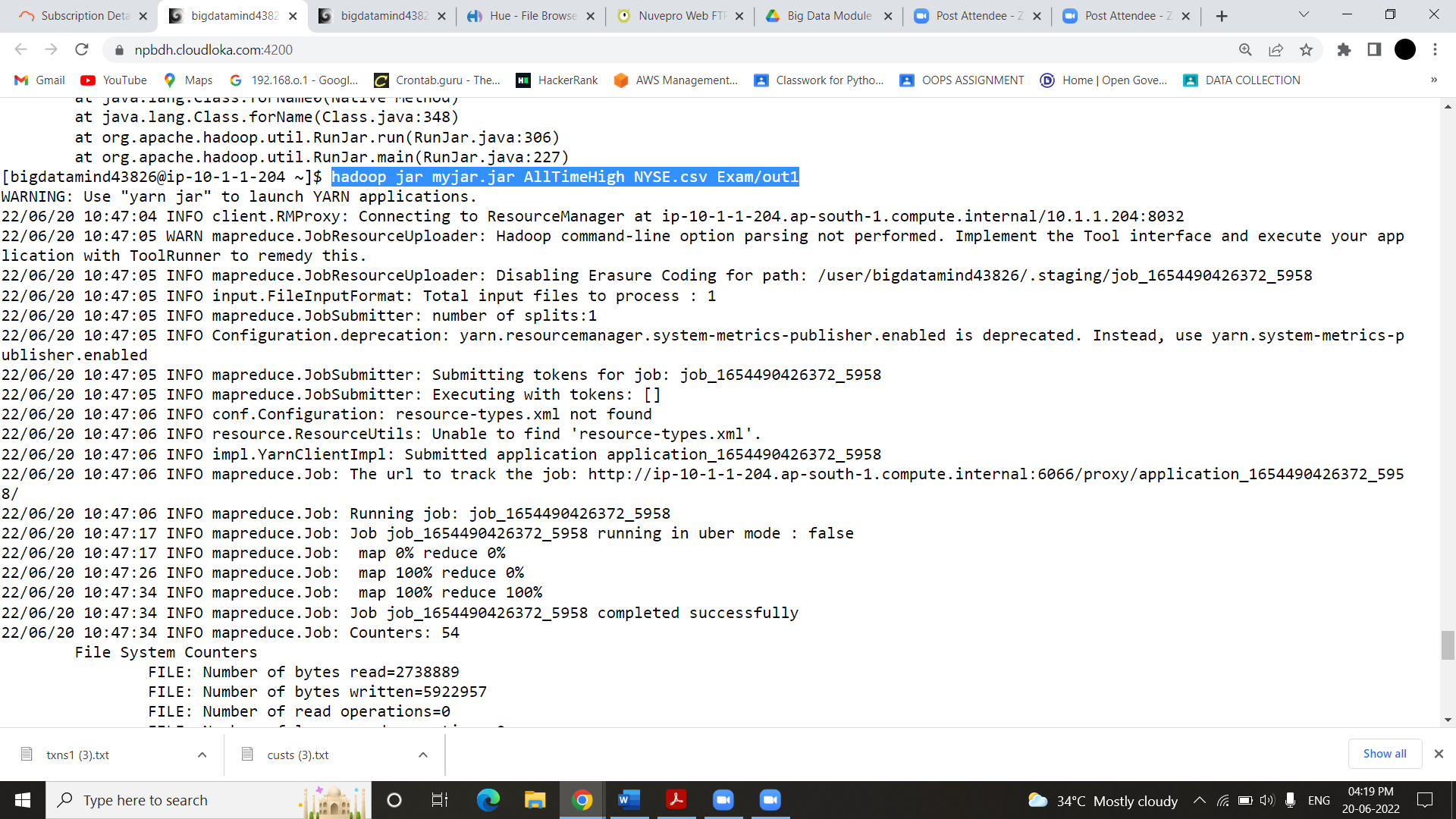
**Q1.**

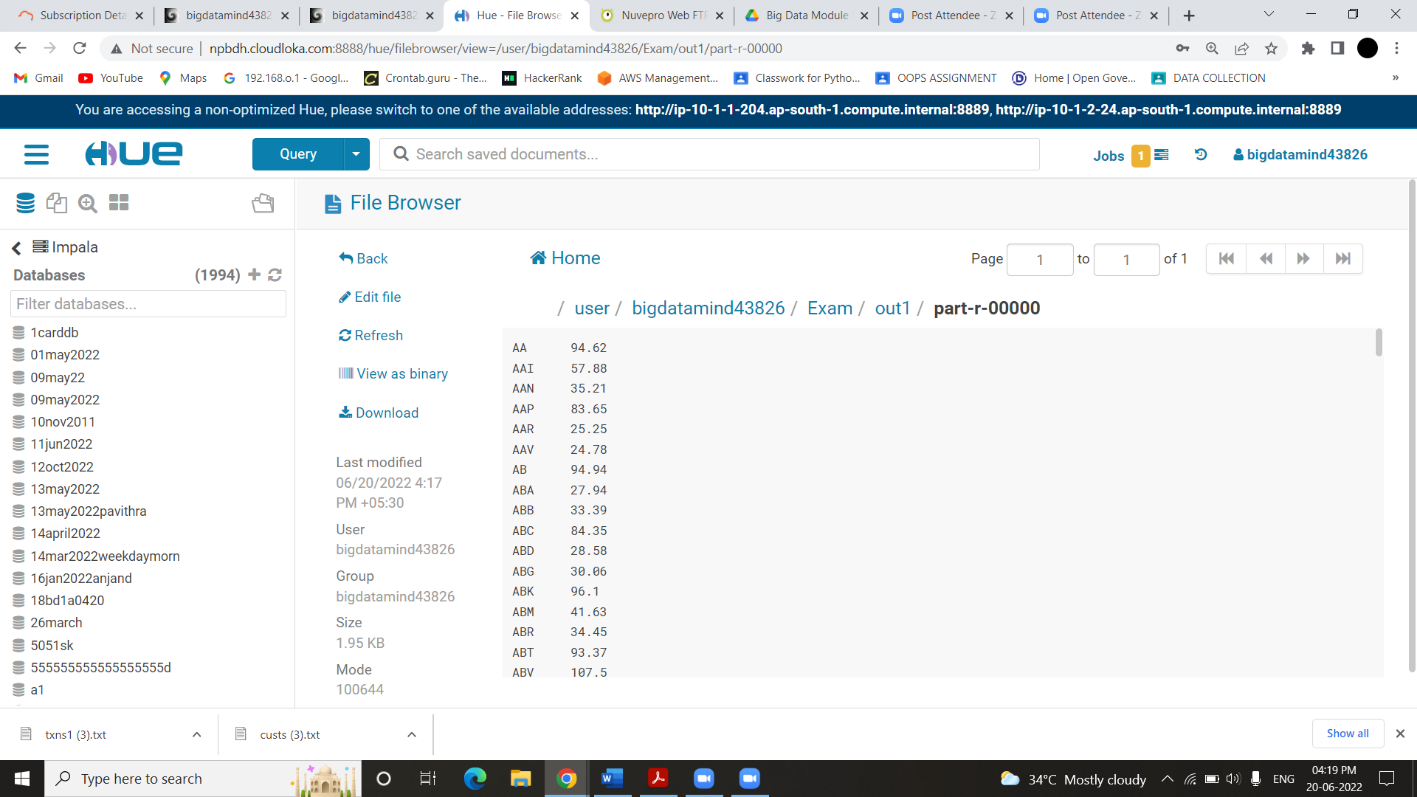
**MapReduce**

**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. Kindlyfind the solutions to the questions below.

hadoop jar myjar.jar AllTimeHigh NYSE.csv Exam/out1

****

****

**Ans--**

**Hive**

Please find the customer data set.

cust id

firstname

lastname

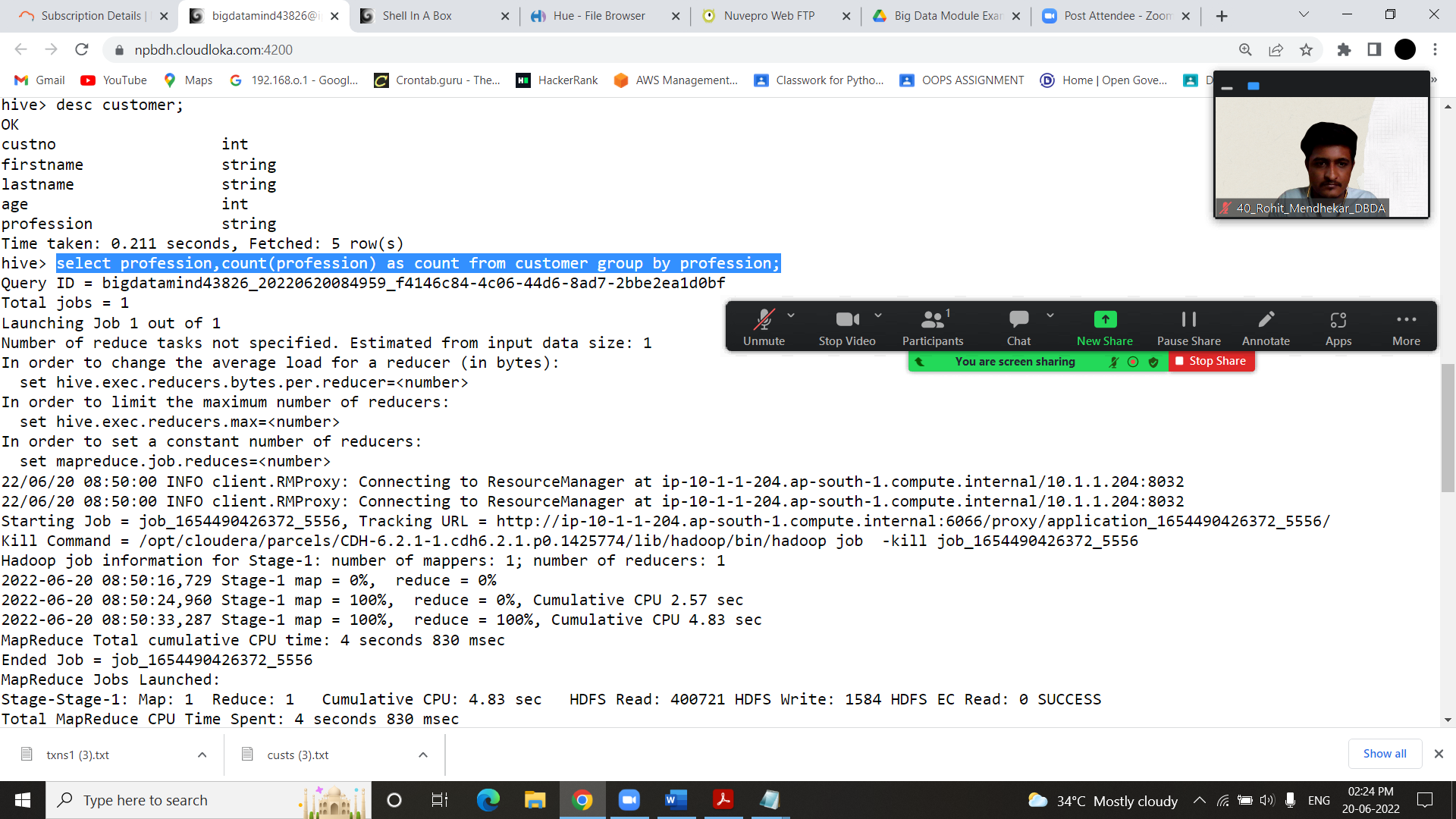
age

profession

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

Ans

select profession,count(profession) as count from customer group by profession;

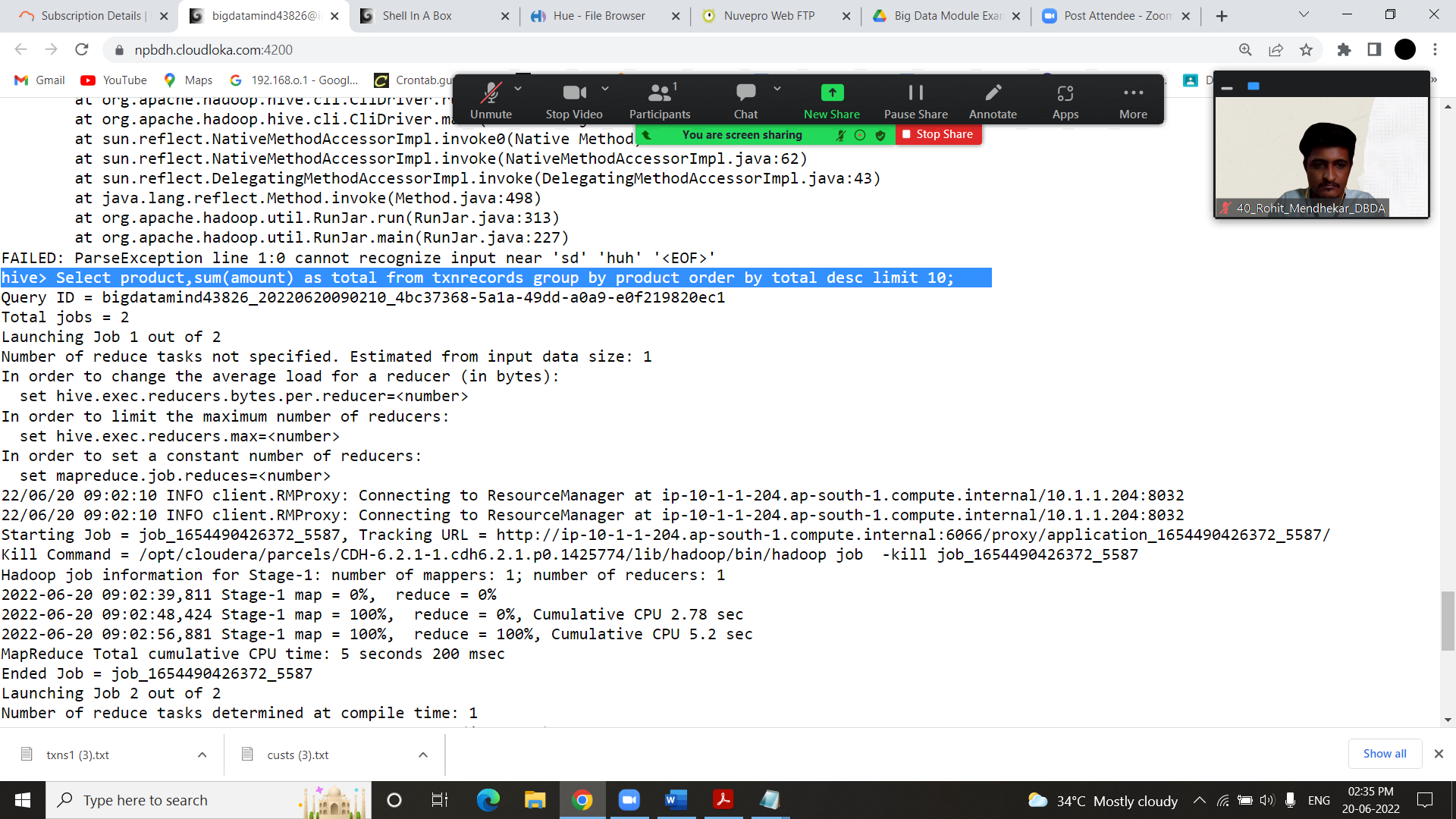


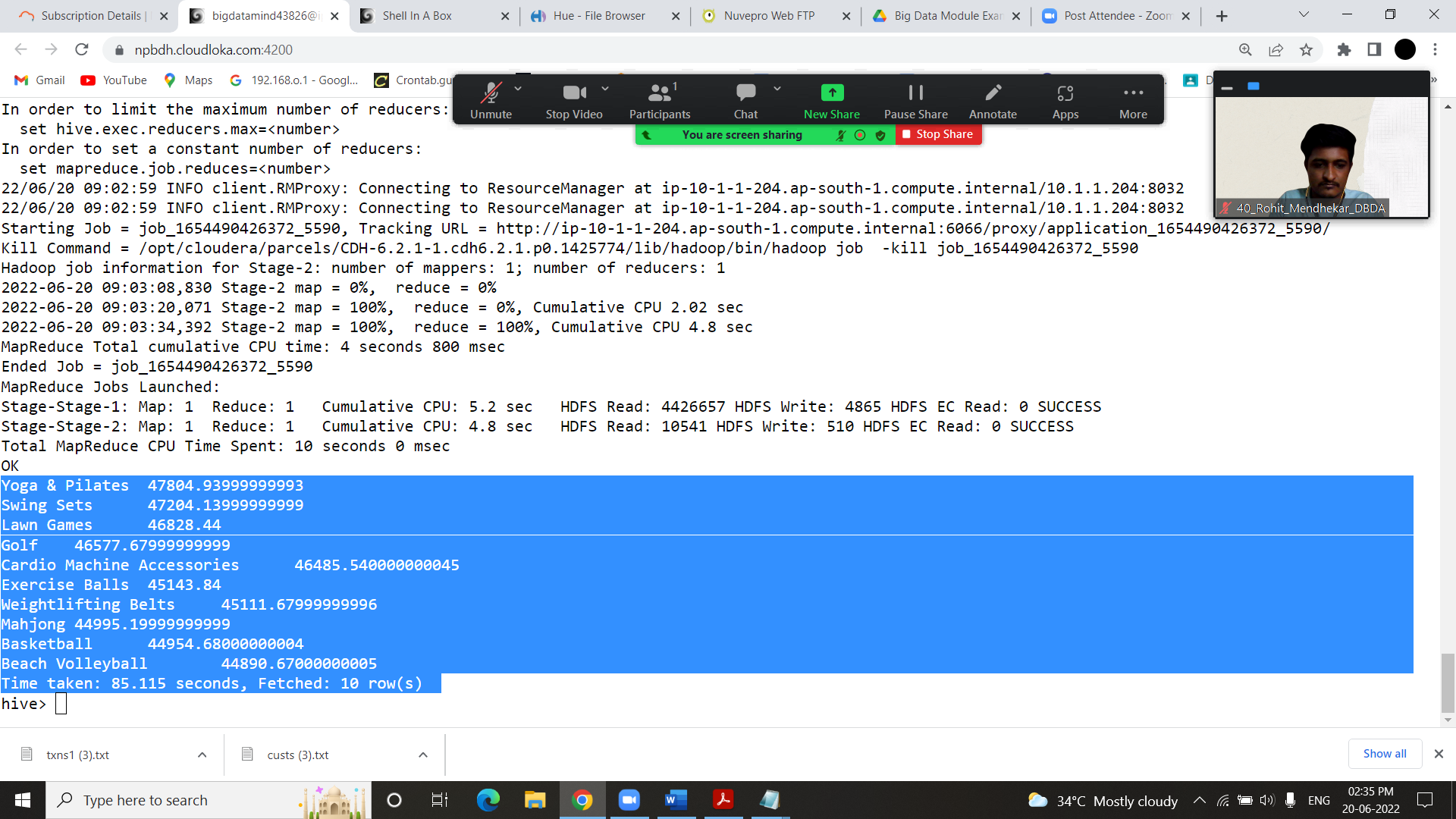


**2) Write a program to find the top 10 products sales wise**

Ans

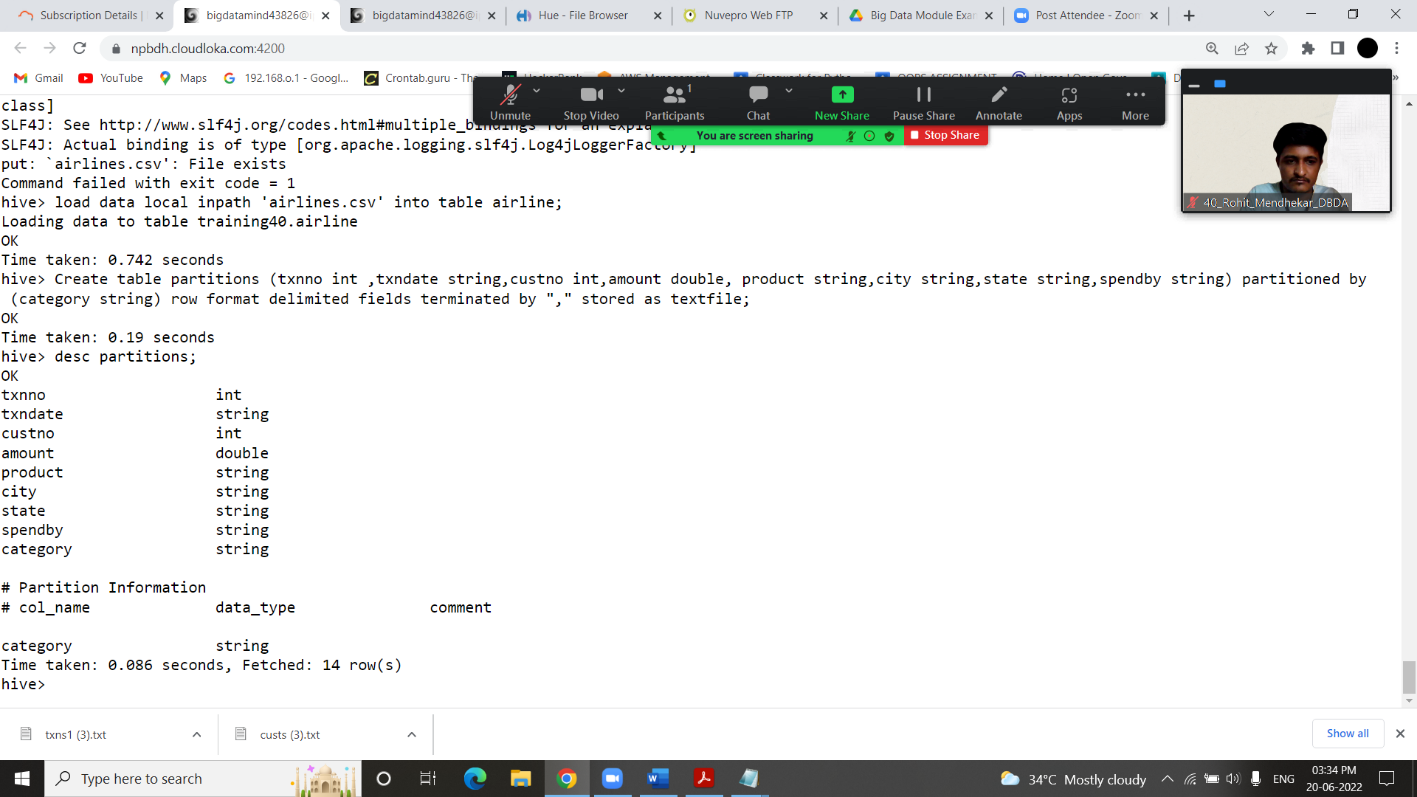
Select product,sum(amount) as total from txnrecords group by product order by total desc limit 10





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

Create table partitions (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;



**QUESTION 3** [15 marks]

**PySpark**

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

**Ans –**

>>> airlineRDD1 = airlineRDD.map(lamda a:a encode("ascii","ignore")

>>> airlineRDD1 = airlineRDD.map(lambda a:a.encode("ascii","ignore"))

>>> header = airlineRDD1.first()

>>> airlineRDD2 =airlineRDD1.filter(lambda a:a!=header)

>>> arrayRDD = airlineRDD2.map(lambda a:a.split(','))

>>> arrayRDD1=arrayRDD.map(lambda a:(a[0],int(a[3])))

>>> arrayRDD2 = arrayRDD1.reduceByKey(lambda a,b:a+b)

>>> sortbyval = arrayRDD2.sortBy(lambda a:-a[1])

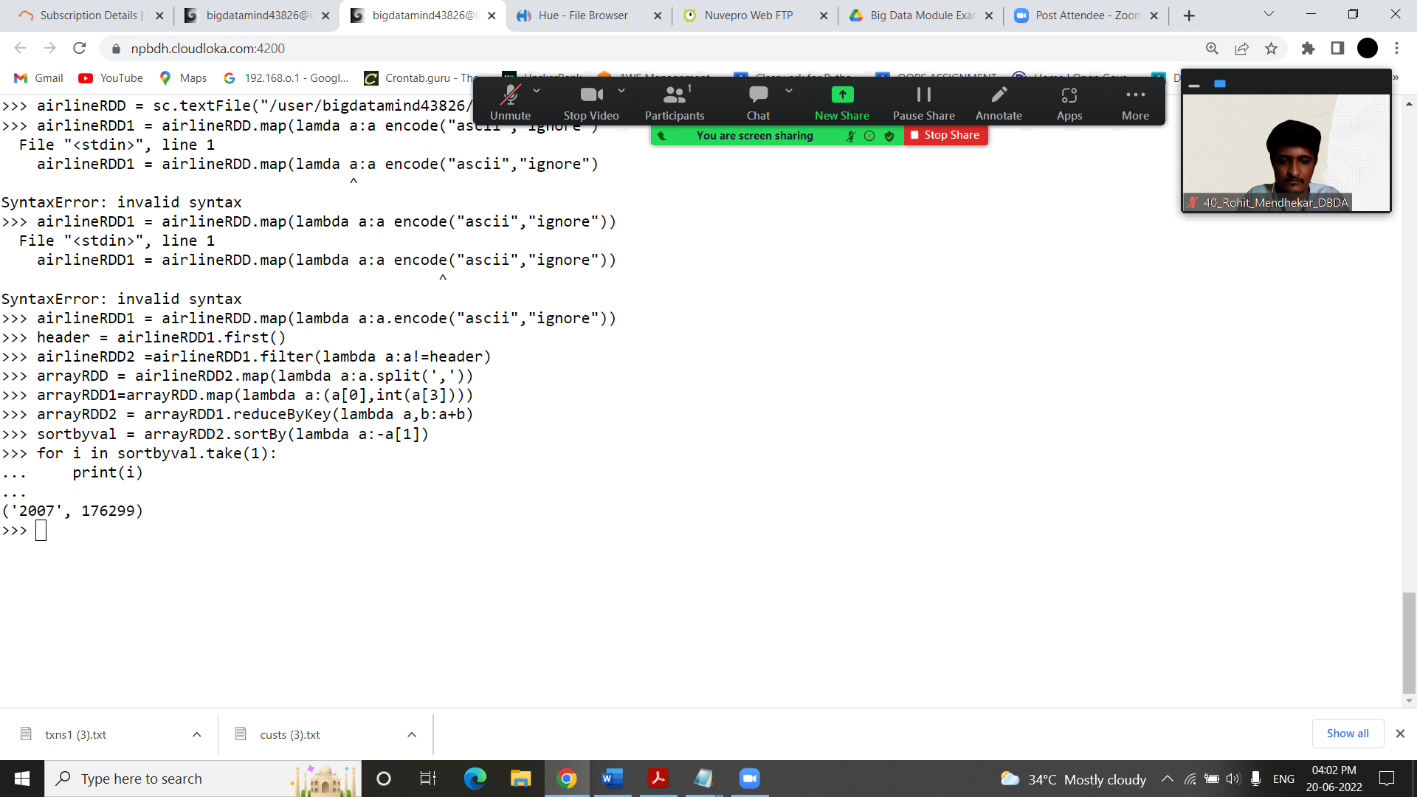
>>> for i in sortbyval.take(1):

... print(i)

...

('2007', 176299)

>>>



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

**Ans--**

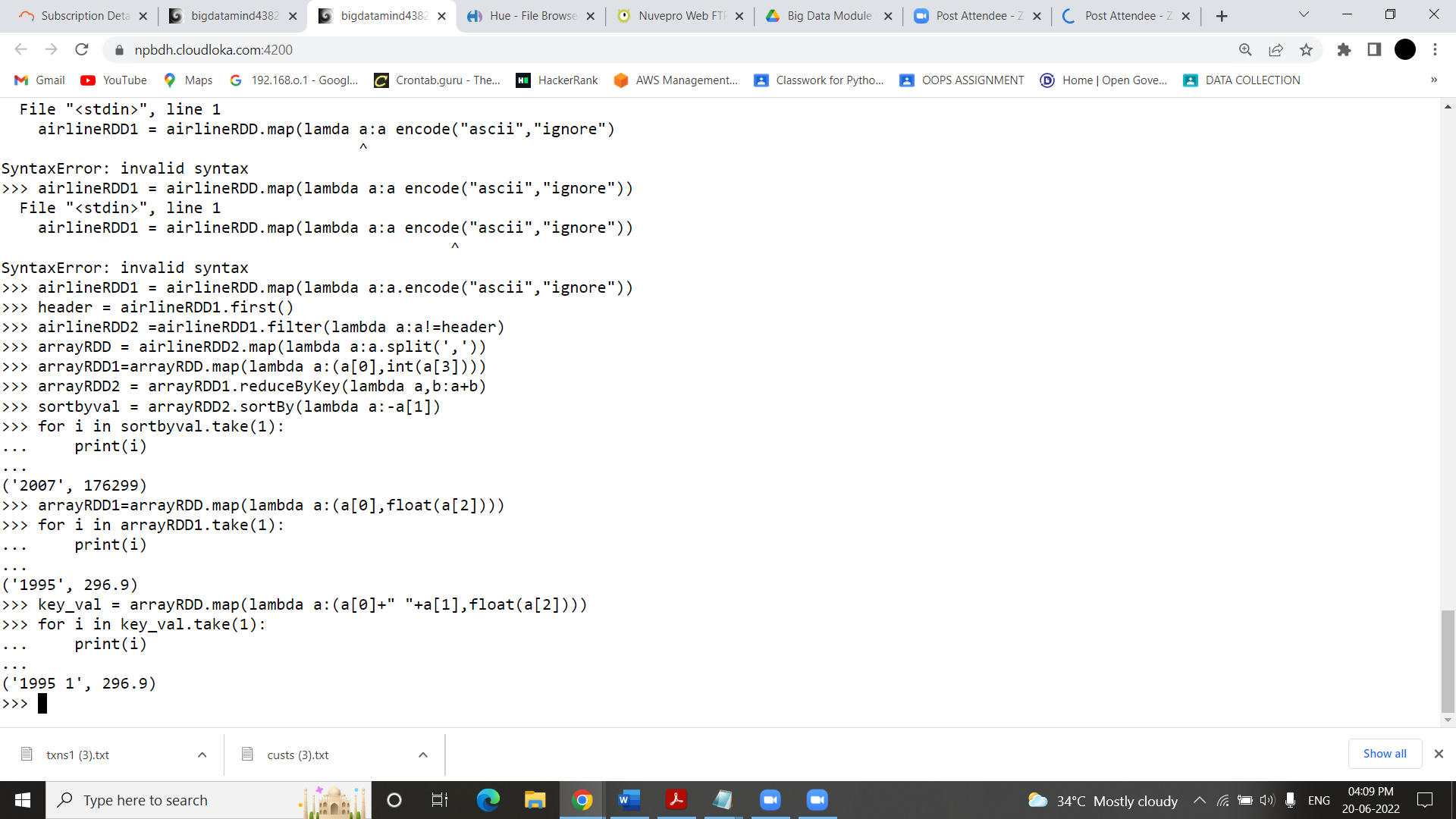
>>> arrayRDD1=arrayRDD.map(lambda a:(a[0],float(a[2])))

>>> for i in arrayRDD1.take(1):

... print(i)

...

('1995', 296.9)



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

**Ans**

>>> arrayRDD2 = key\_val.reduceByKey(lambda a,b:a+b)

>>> sortbyval = arrayRDD2.sortBy(lambda a:-a[1])

>>> for i in sortbyval.take(1):

... print(i)

...

('2014 3', 396.37)

