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.

Ans==

hadoop jar newmyjar1.jar cdac/AllTimeHigh /user/bigdatamind43814/NYSE.csv exam/out

package cdac;

import java.io.\*;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.DoubleWritable;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.conf.\*;

import org.apache.hadoop.fs.\*;

import org.apache.hadoop.mapreduce.lib.input.\*;

import org.apache.hadoop.mapreduce.lib.output.\*;

public class AllTimeHigh {

public static class MapClass extends Mapper<LongWritable,Text,Text,DoubleWritable>

{

private Text stock\_id = new Text();

private DoubleWritable High = new DoubleWritable();

public void map(LongWritable key, Text value, Context context)

{

try{

String[] str = value.toString().split(",");

double high = Double.parseDouble(str[4]);

stock\_id.set(str[1]);

High.set(high);

//context.write(new Text(str[1]),new LongWritable(vol));

context.write(stock\_id, High);

}

catch(Exception e)

{

System.out.println(e.getMessage());

}

}

}

public static class ReduceClass extends Reducer<Text,DoubleWritable,Text,DoubleWritable>

{

private DoubleWritable result = new DoubleWritable();

public void reduce(Text key, Iterable<DoubleWritable> values,Context context) throws IOException, InterruptedException {

double maxValue=0;

double temp\_val=0;

for (DoubleWritable value : values) {

temp\_val = value.get();

if (temp\_val > maxValue) {

maxValue = temp\_val;

}

}

result.set(maxValue);

context.write(key, result);

//context.write(key, new LongWritable(sum));

}

}

public static void main(String[] args) throws Exception {

Configuration conf = new Configuration();

//conf.set("name", "value")

//conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");

Job job = Job.getInstance(conf, "Highest Price for each stock");

job.setJarByClass(AllTimeHigh.class);

job.setMapperClass(MapClass.class);

//job.setCombinerClass(ReduceClass.class);

job.setReducerClass(ReduceClass.class);

job.setNumReduceTasks(1);

job.setOutputKeyClass(Text.class);

job.setOutputValueClass(DoubleWritable.class);

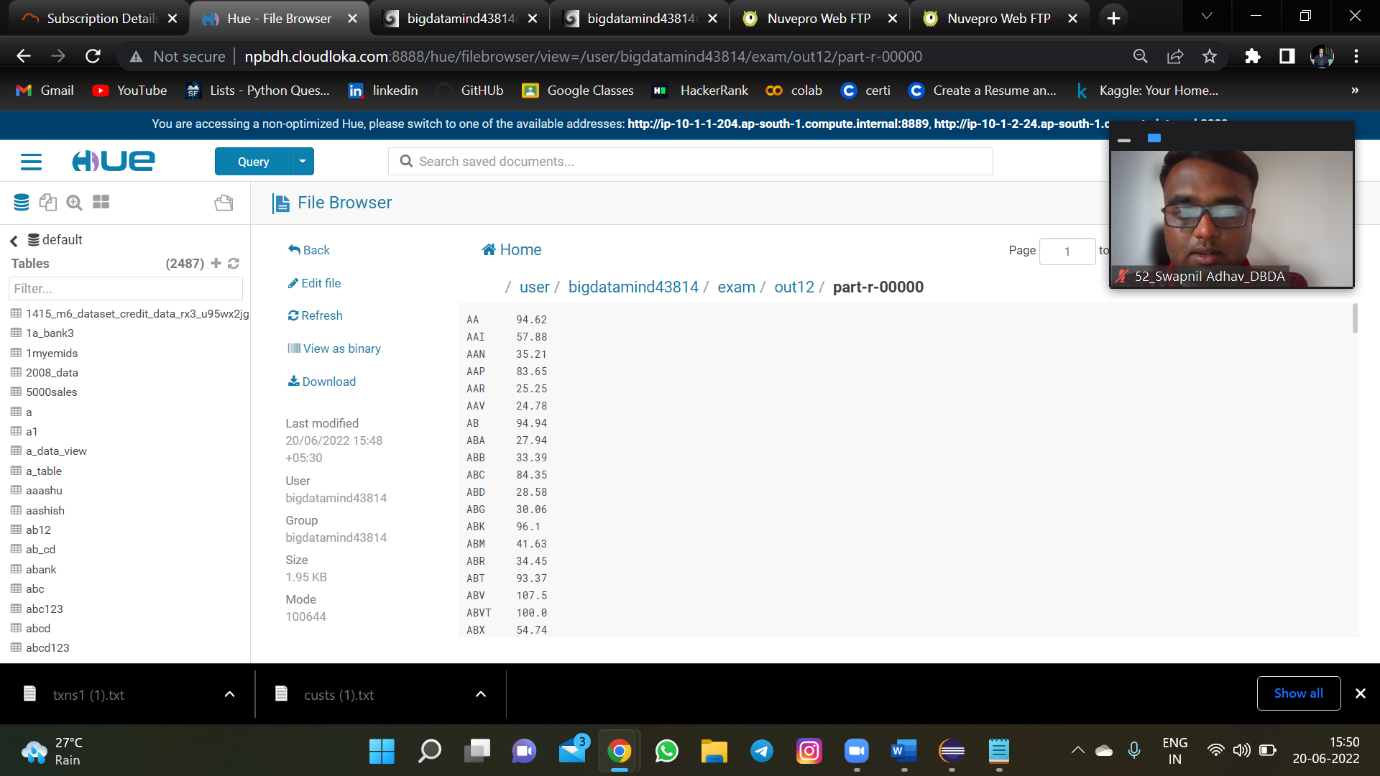
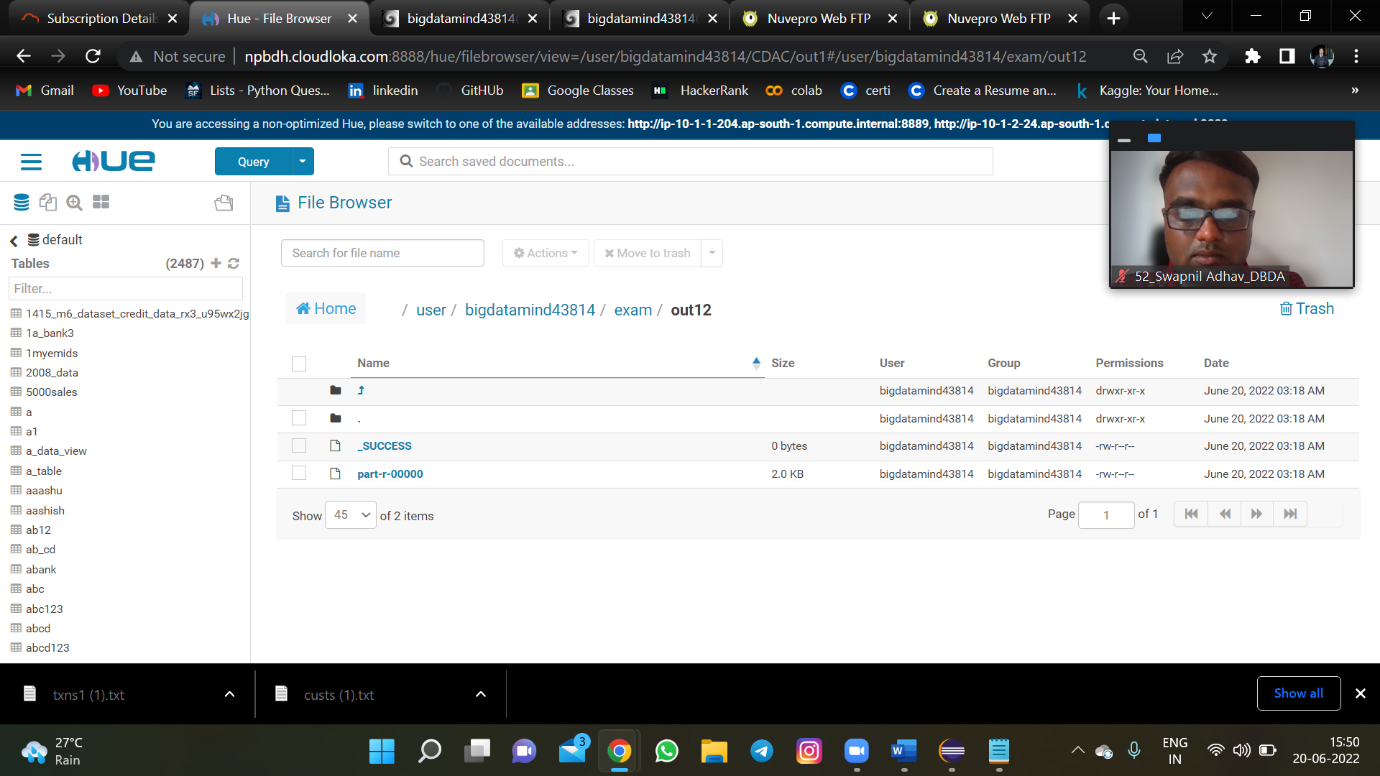
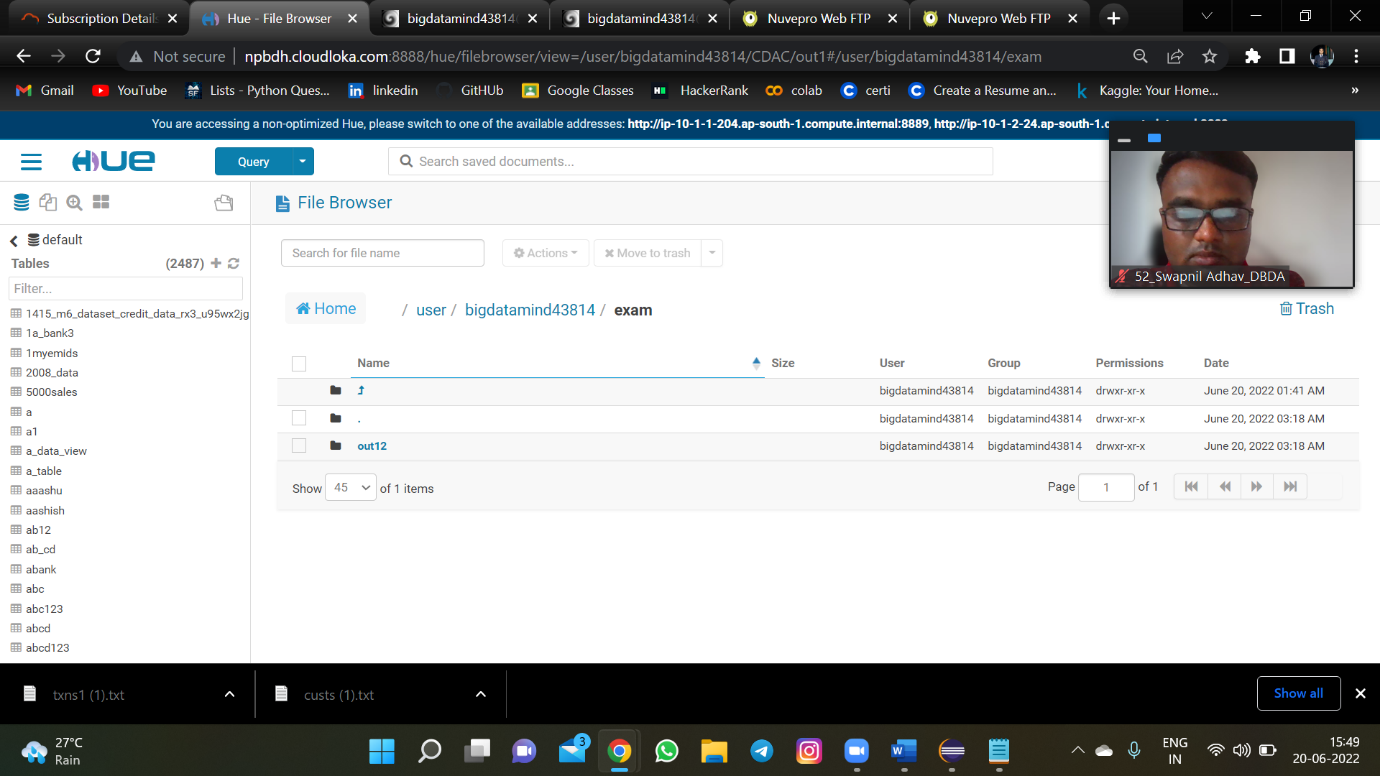
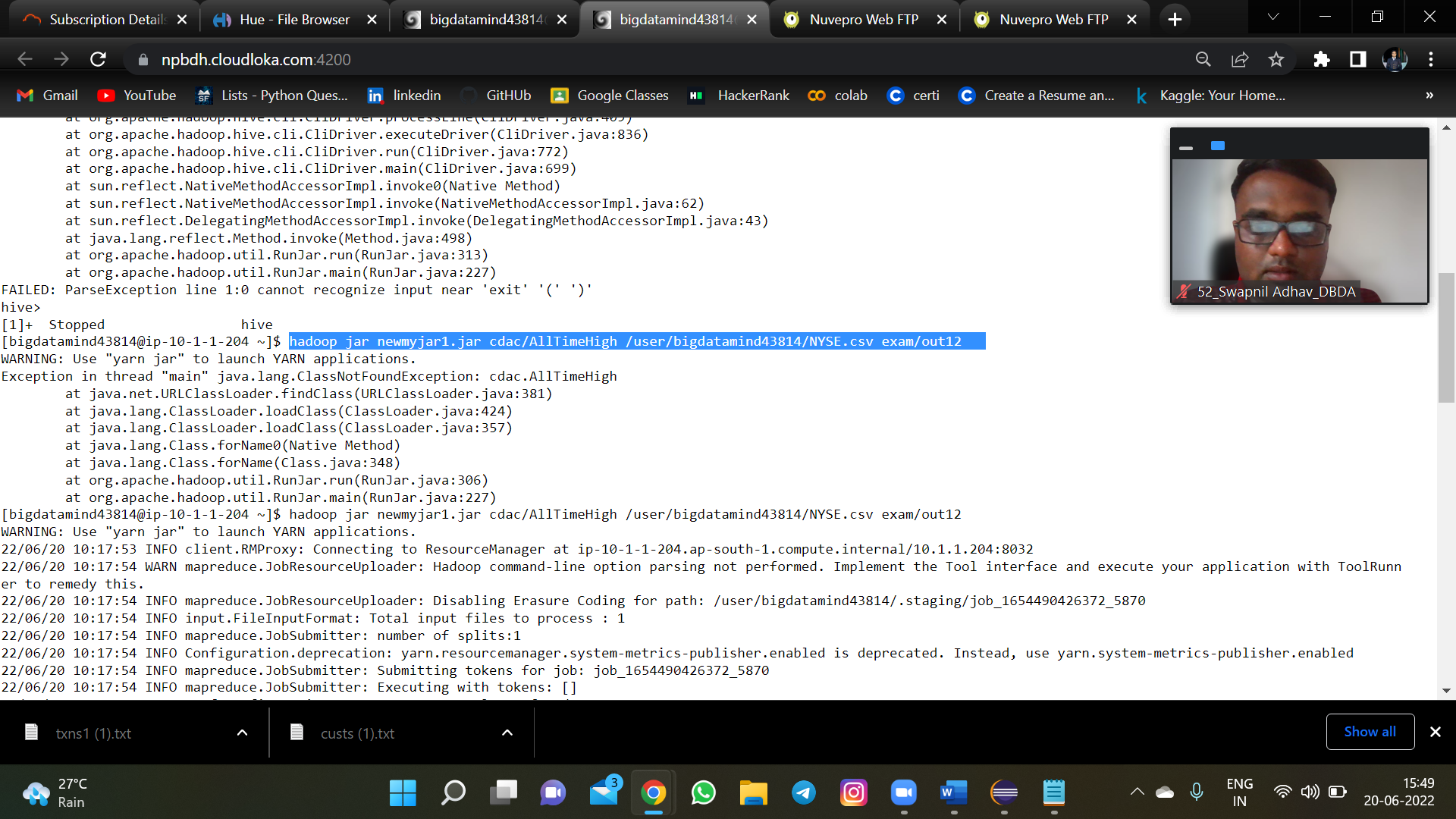
FileInputFormat.addInputPath(job, new Path(args[0]));

FileOutputFormat.setOutputPath(job, new Path(args[1]));

System.exit(job.waitForCompletion(true) ? 0 : 1);

}

}



Q2) Hive

Table creation and data insertion

**CUSTOMER TABLE**

create table cust10 (cust\_id int,first\_name string,last\_name string

,age int , profession string) row format delimited fields

terminated by ',' stored as textfile;

load data inpath '/user/bigdatamind43814/exam/custs.txt'

overwrite into table cust10;

**TRANSECTION TABLE**

create table txn10 (txn\_id int ,txn\_date string, cust\_id int,amount double,

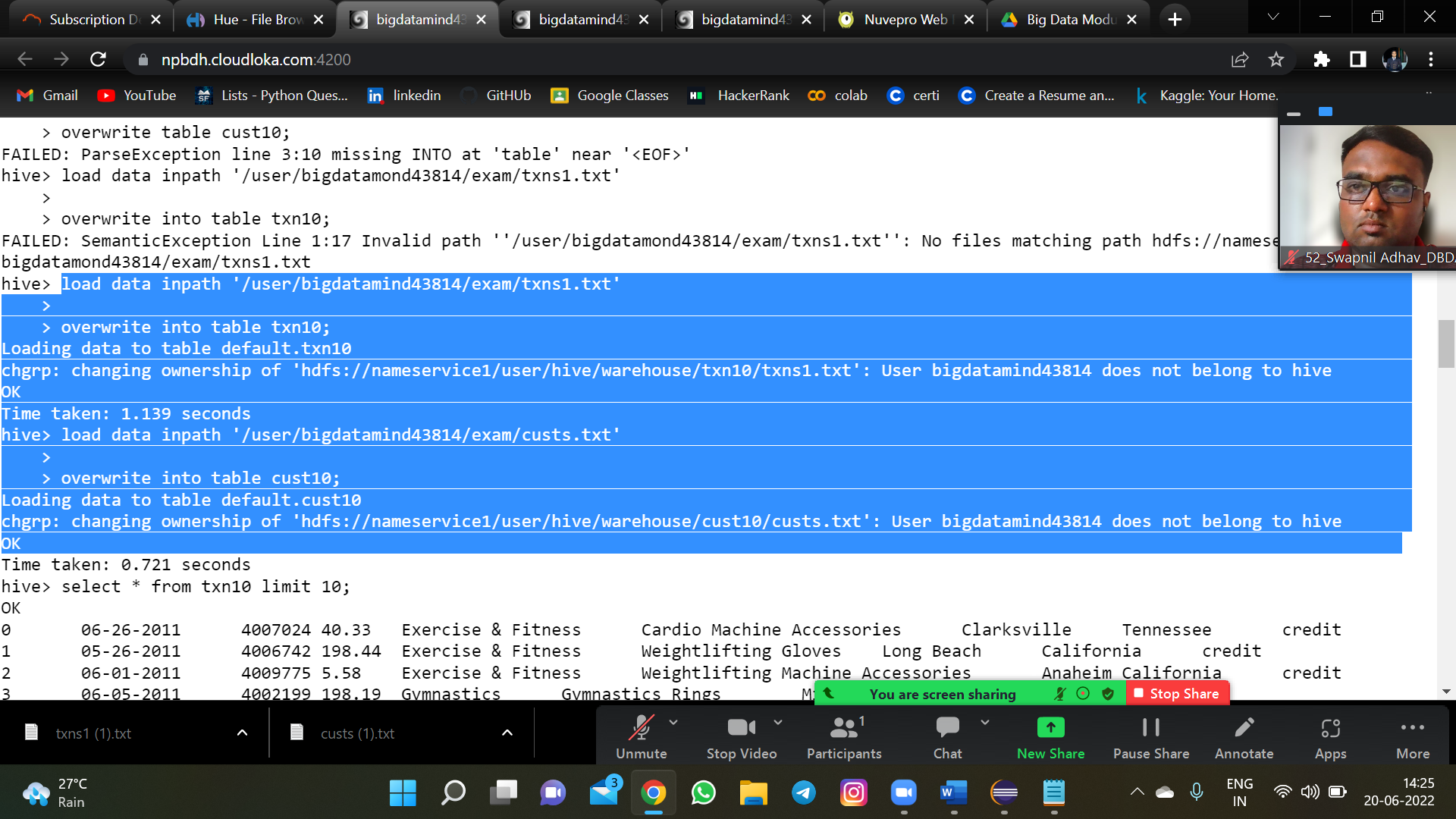
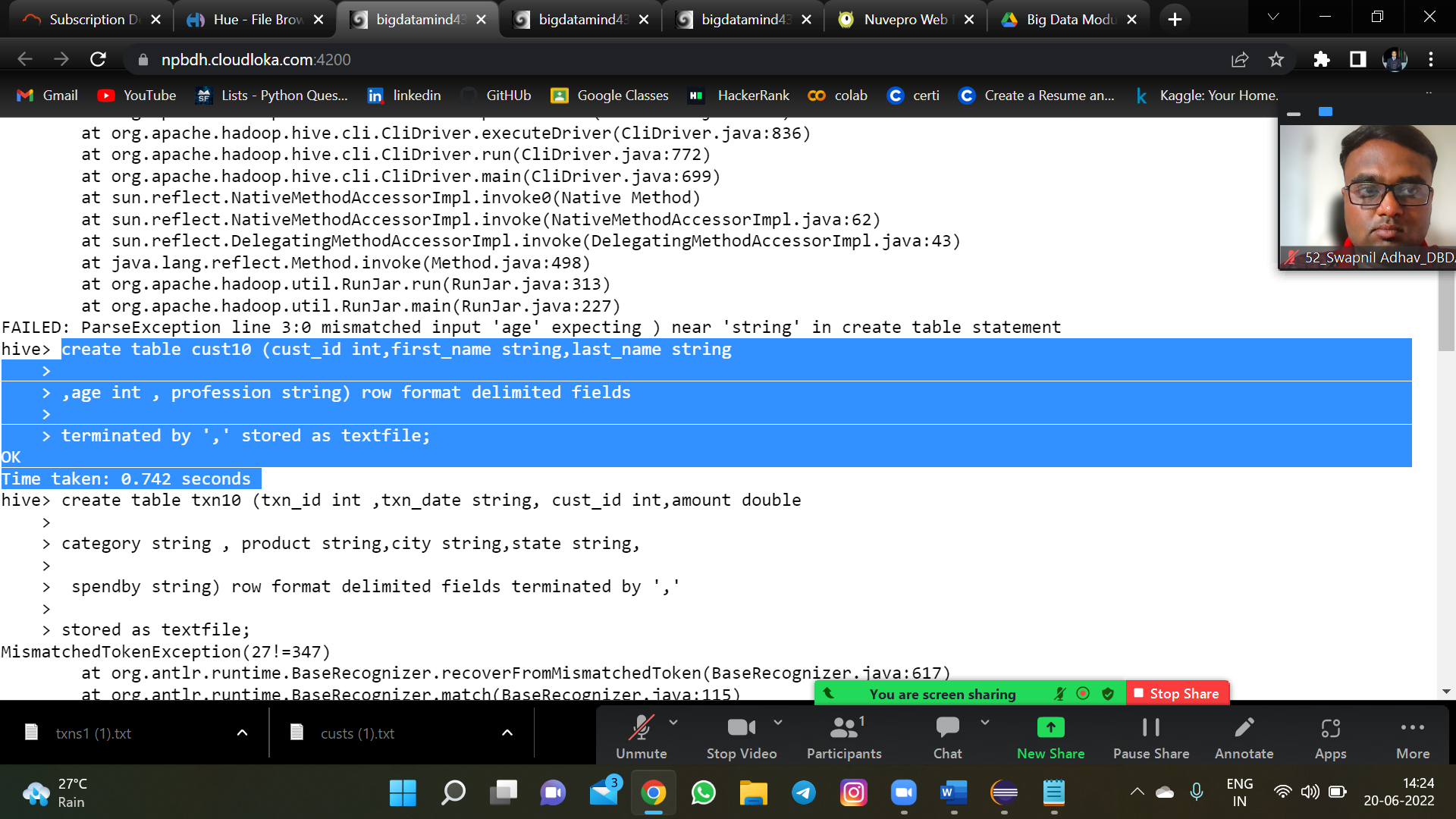
category string , product string,city string,state string,

spendby string) row format delimited fields terminated by ','

stored as textfile;

load data inpath '/user/bigdatamind43814/exam/txns1.txt'

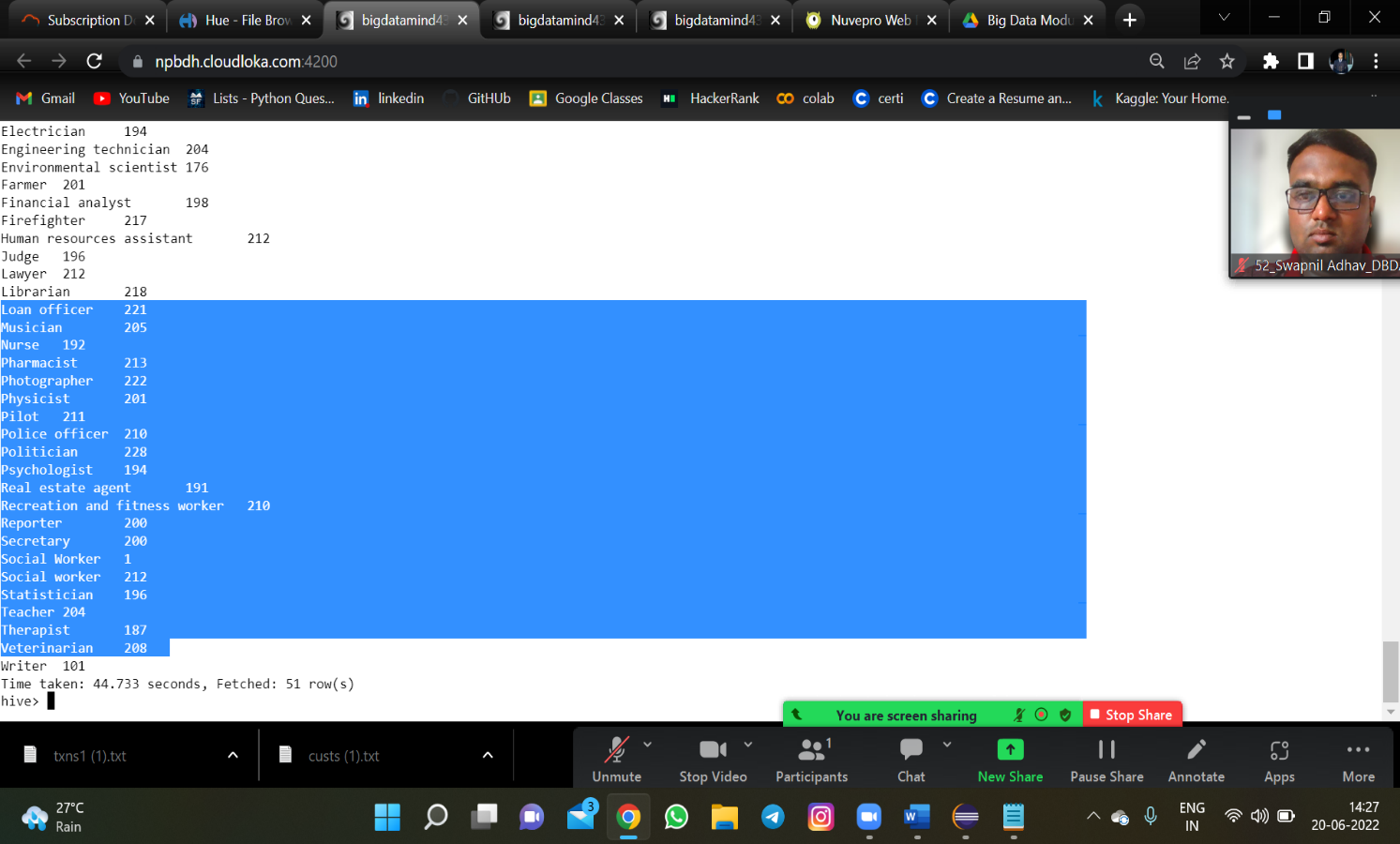
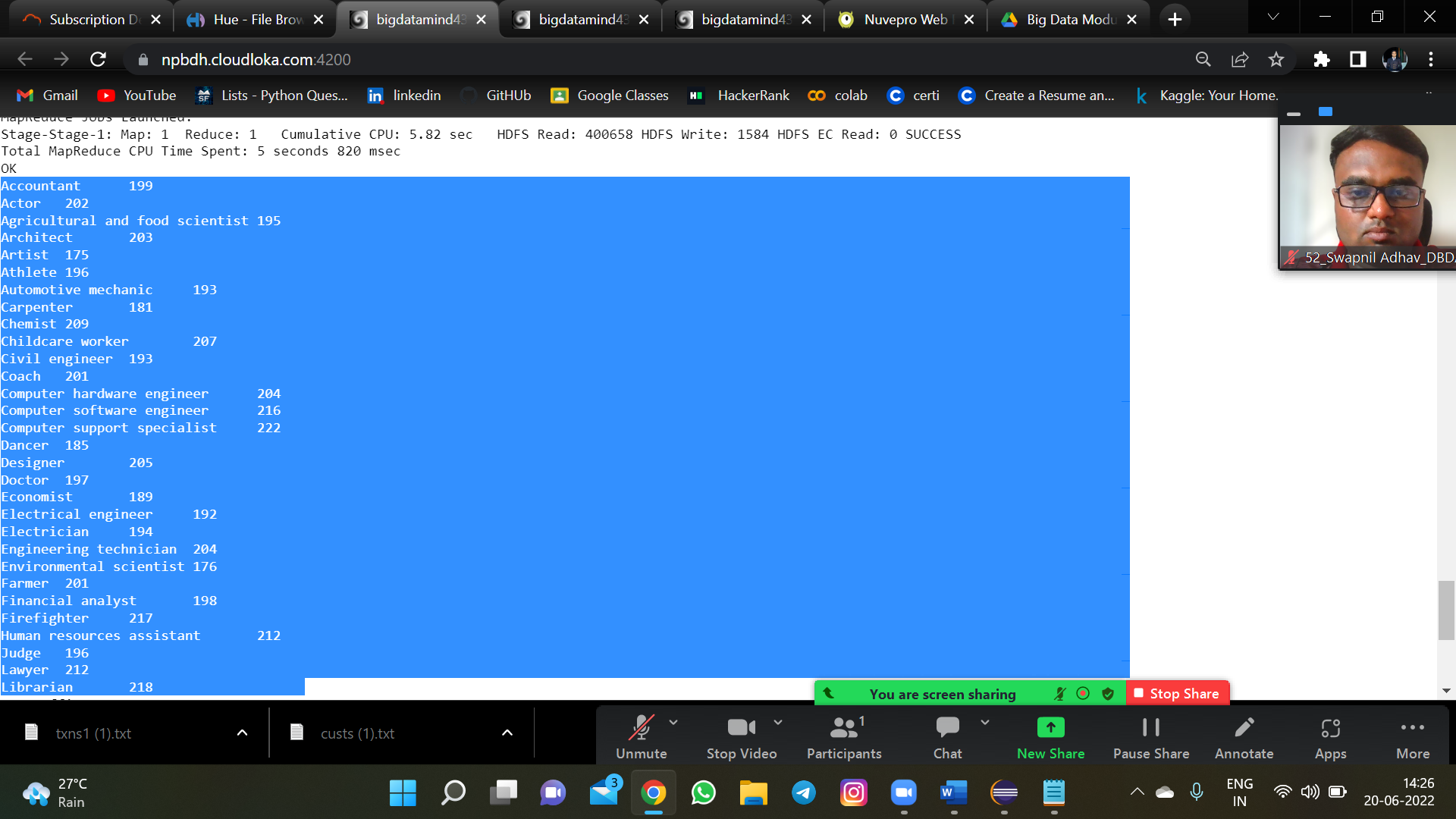
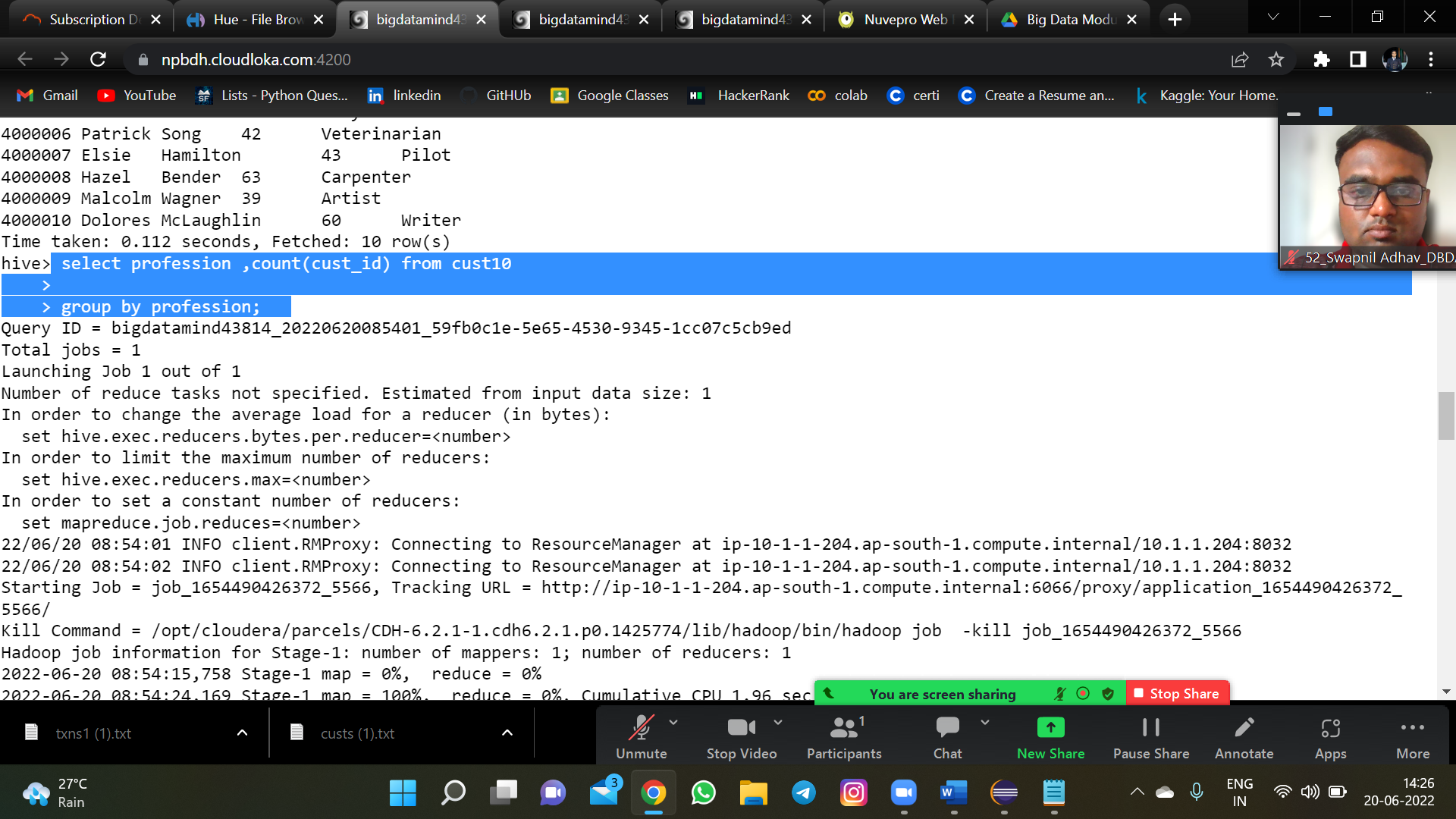
overwrite into table txn10;overwrite into table txn10;



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

**ANS:-** select profession ,count(cust\_id) from cust10

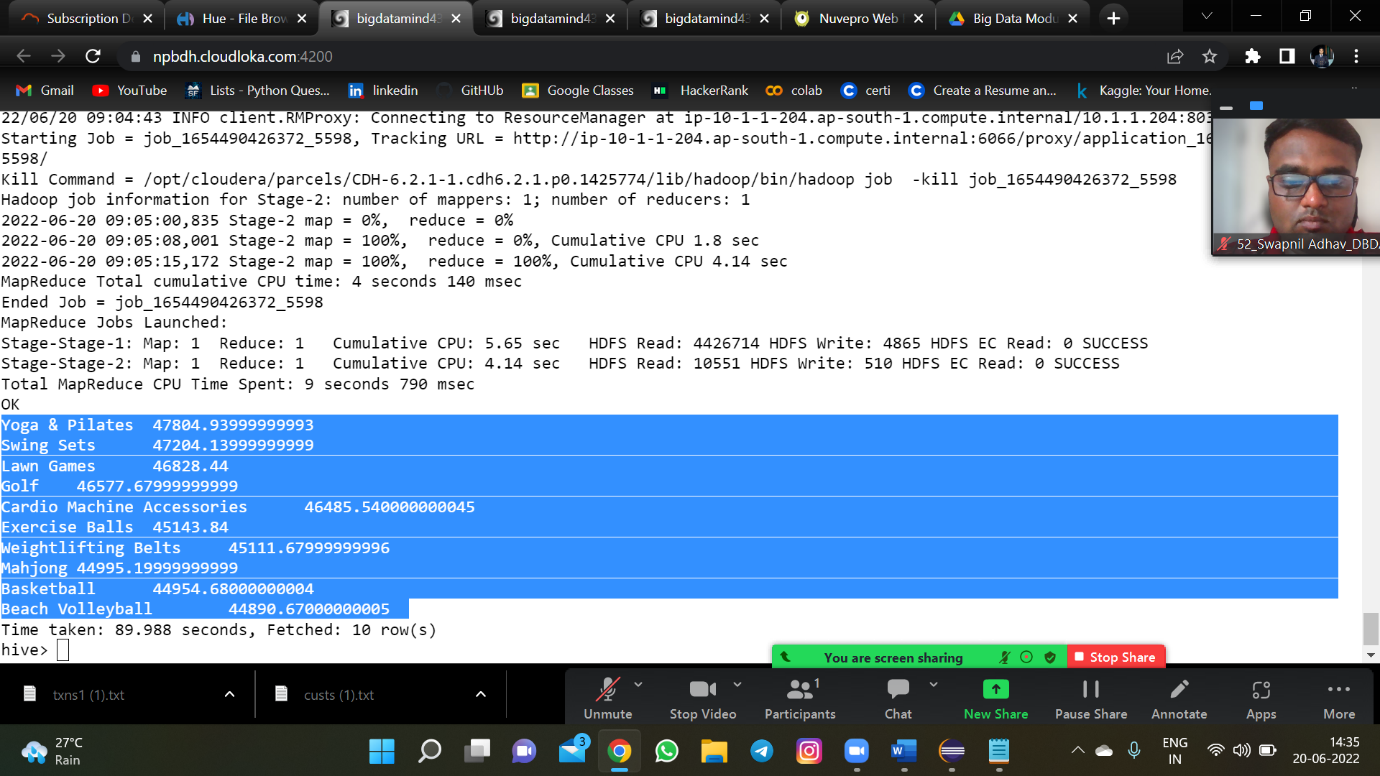
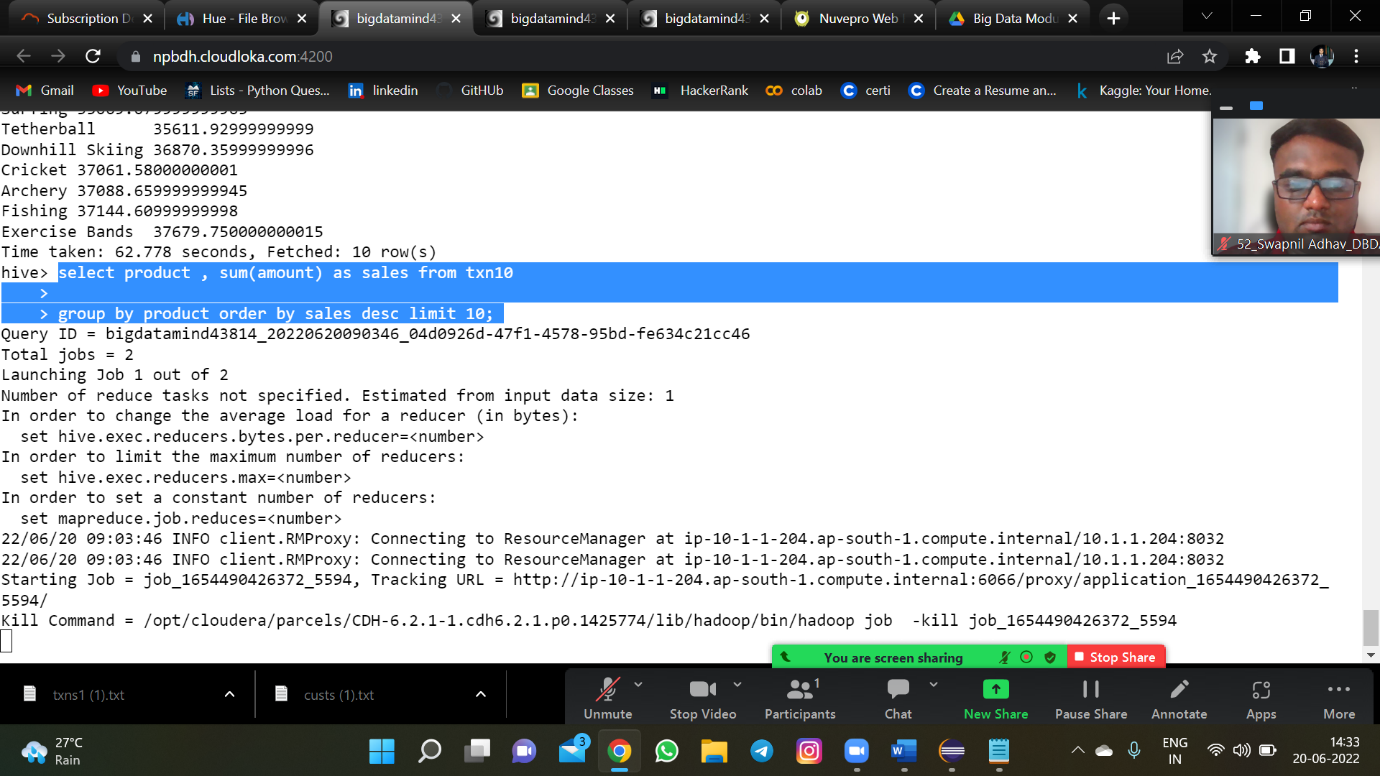
group by profession;



**QUERY2) 2)** Write a program to find the top 10 products sales wise

**Ans:** select product , sum(amount) as sales from txn10

group by product order by sales desc limit 10;



Q3) 3) Write a program to create partiioned table on category

Ans:-

create table txn10\_part (txn\_id int ,txn\_date string,

cust\_id int,amount double, product string,city string,state string,

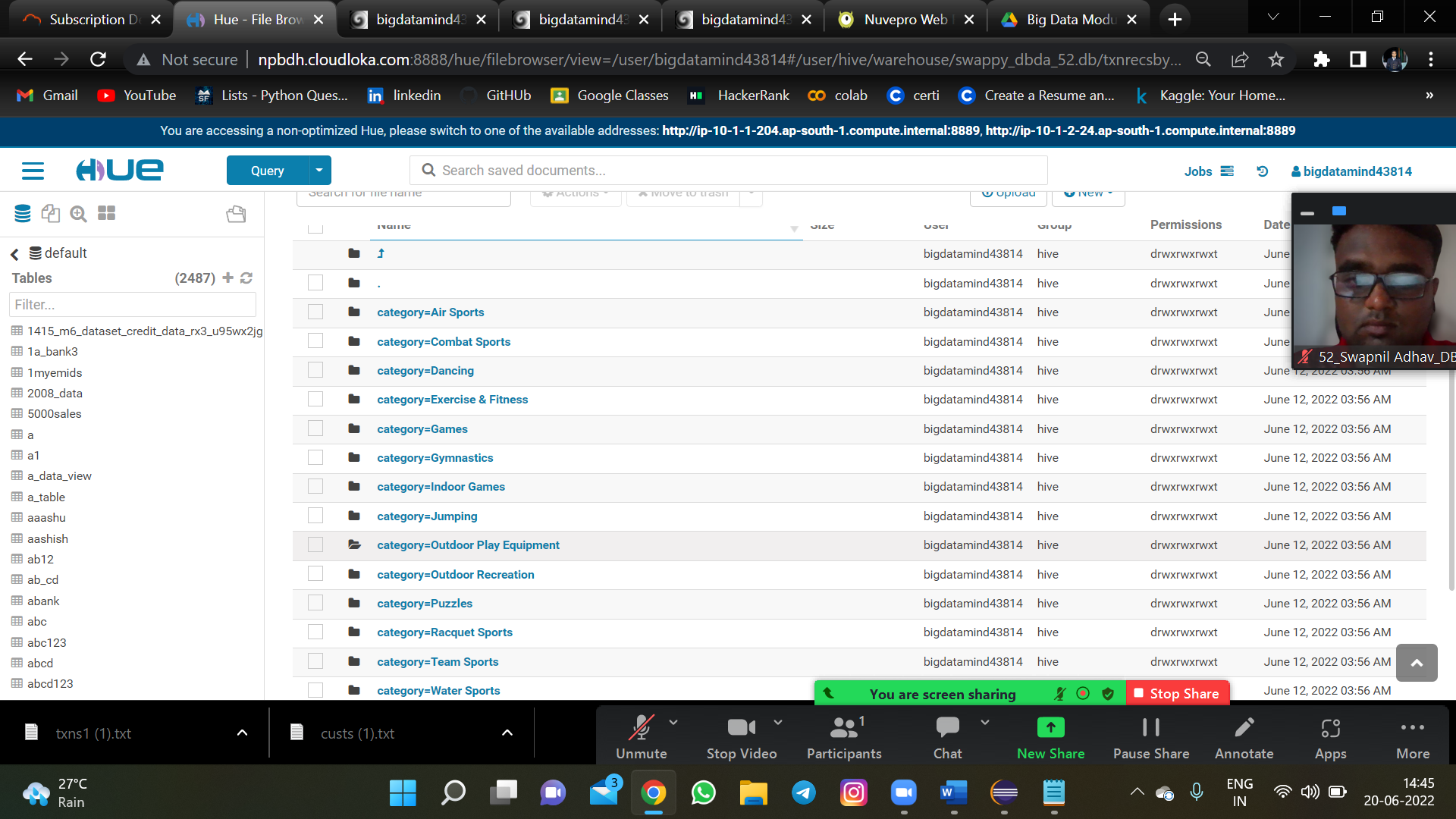
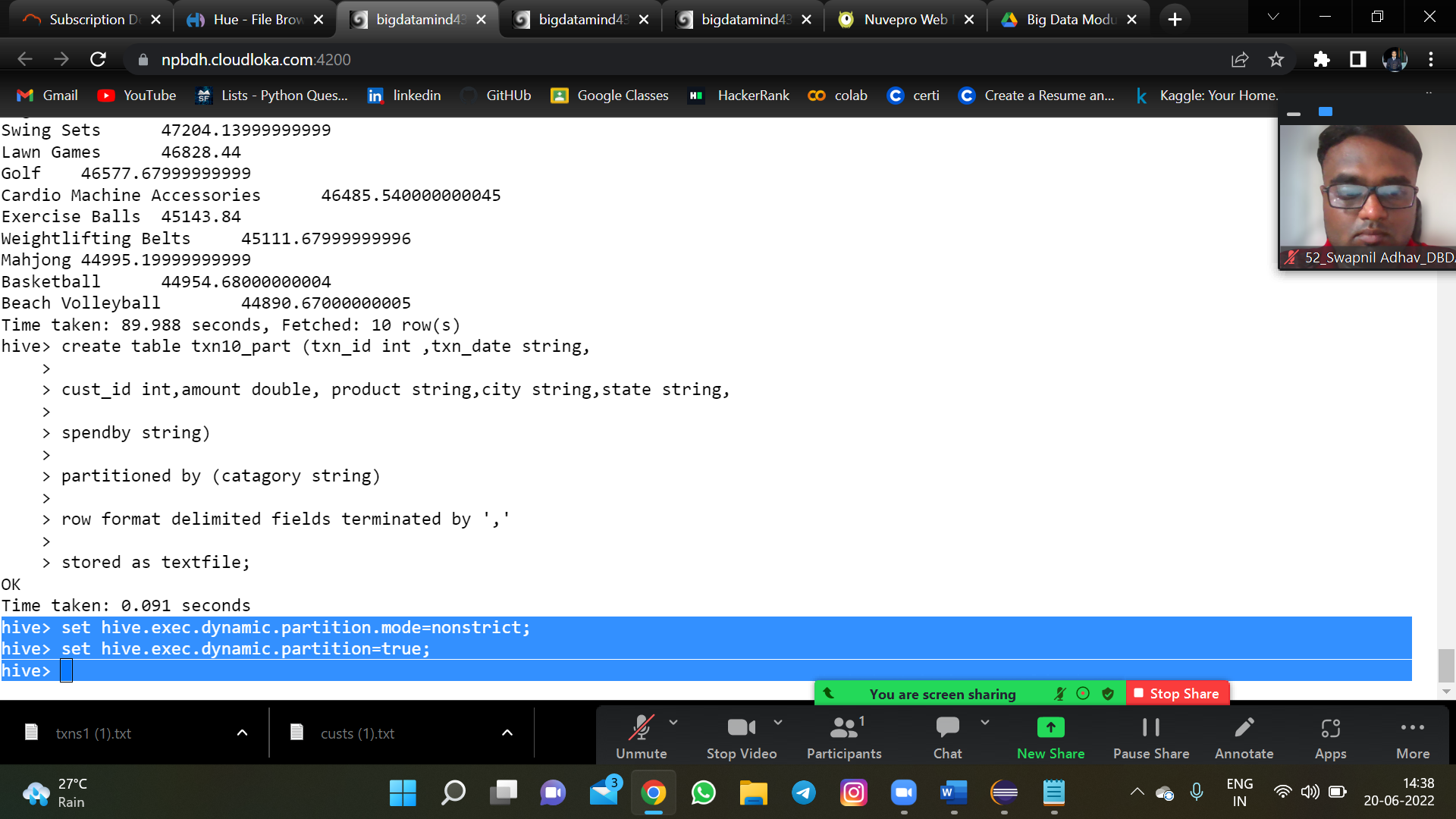
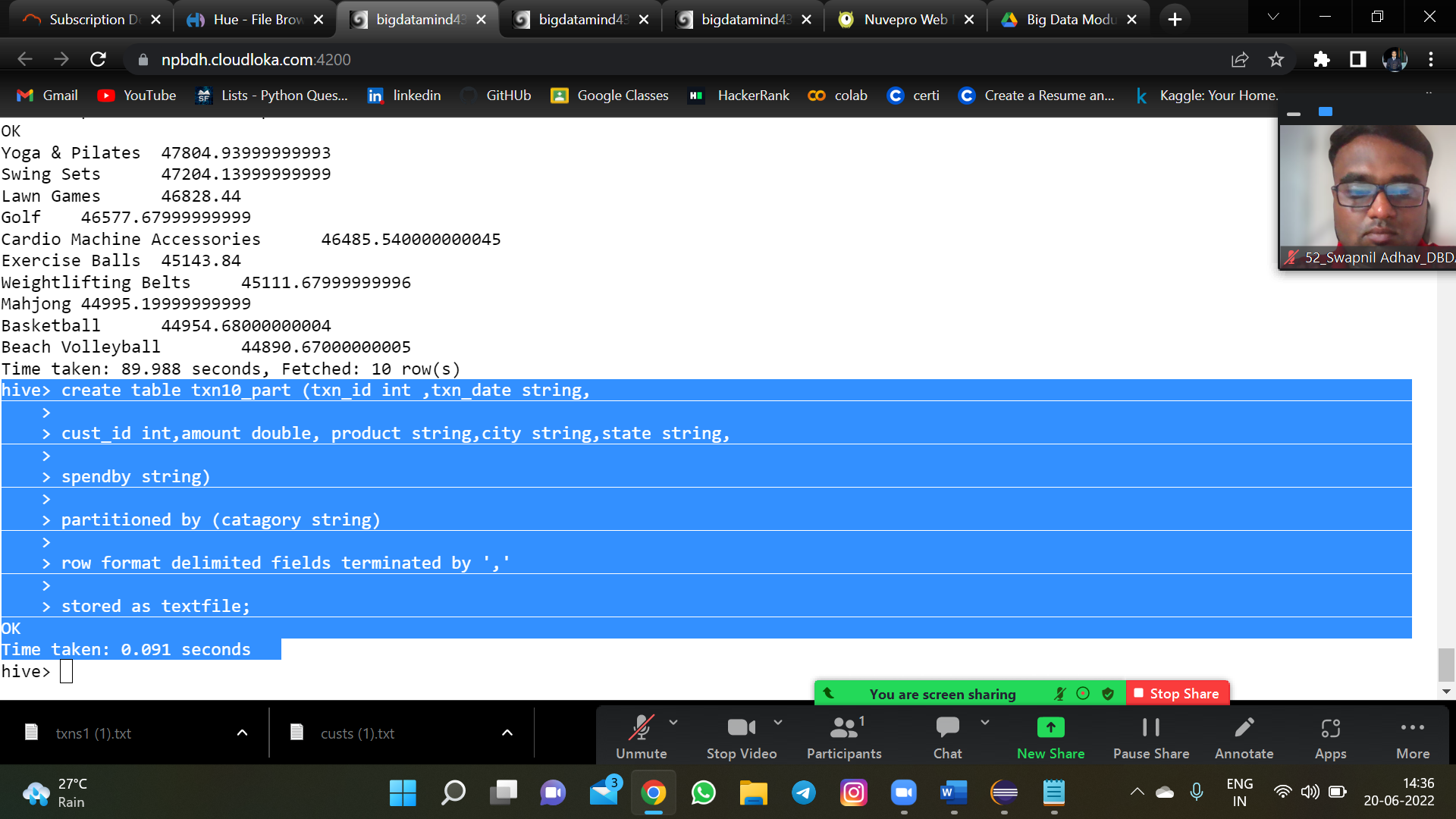
spendby string)

partitioned by (catagory string)

row format delimited fields terminated by ','

stored as textfile;

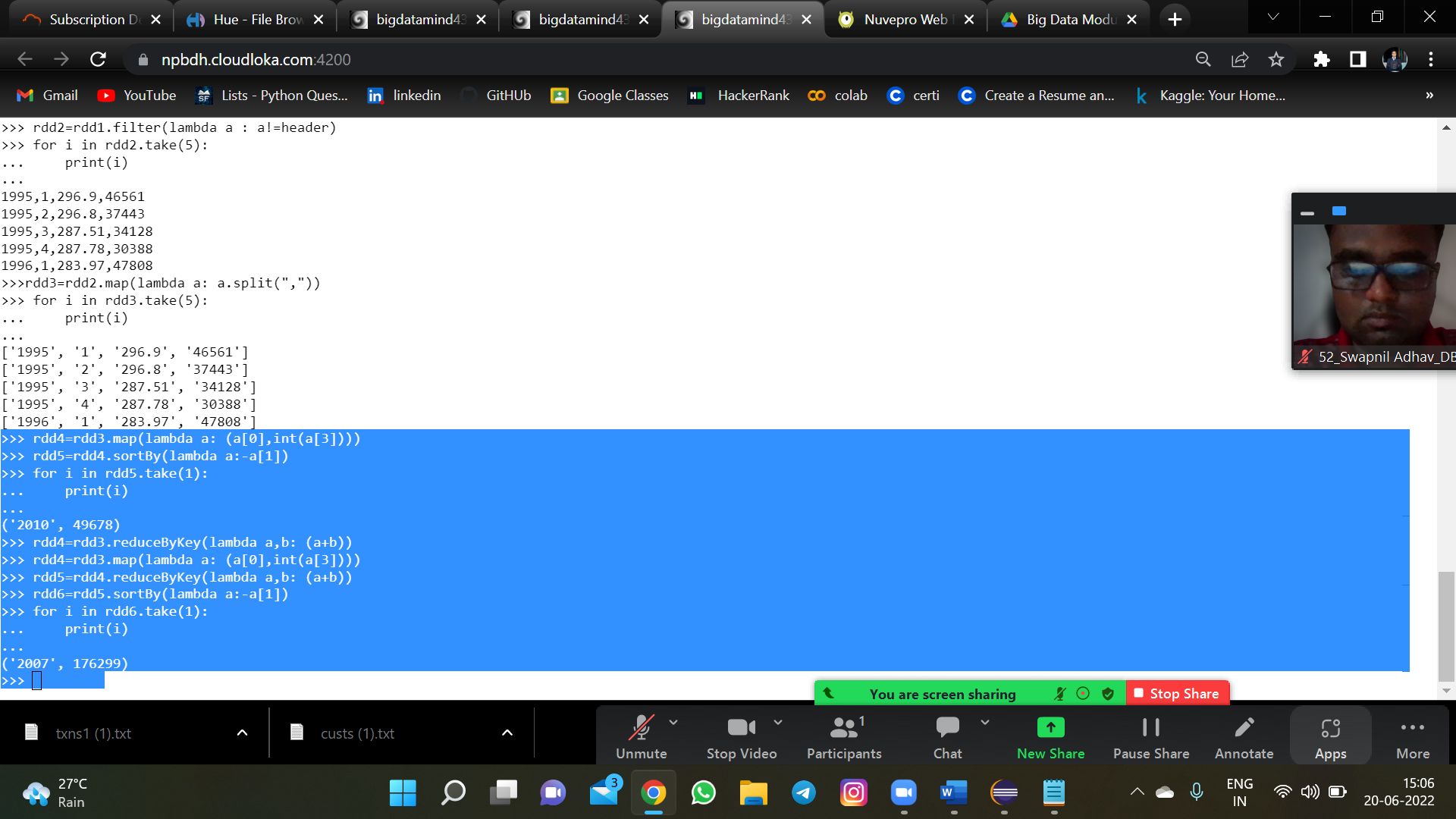
insert overwrite table txn10\_part partition(category) select t.txn\_id,t.txn\_date,t.cust\_id,t.amount,

t.product,t.city,t.state,t.spendby , t.category from txn10 t distribute by categ

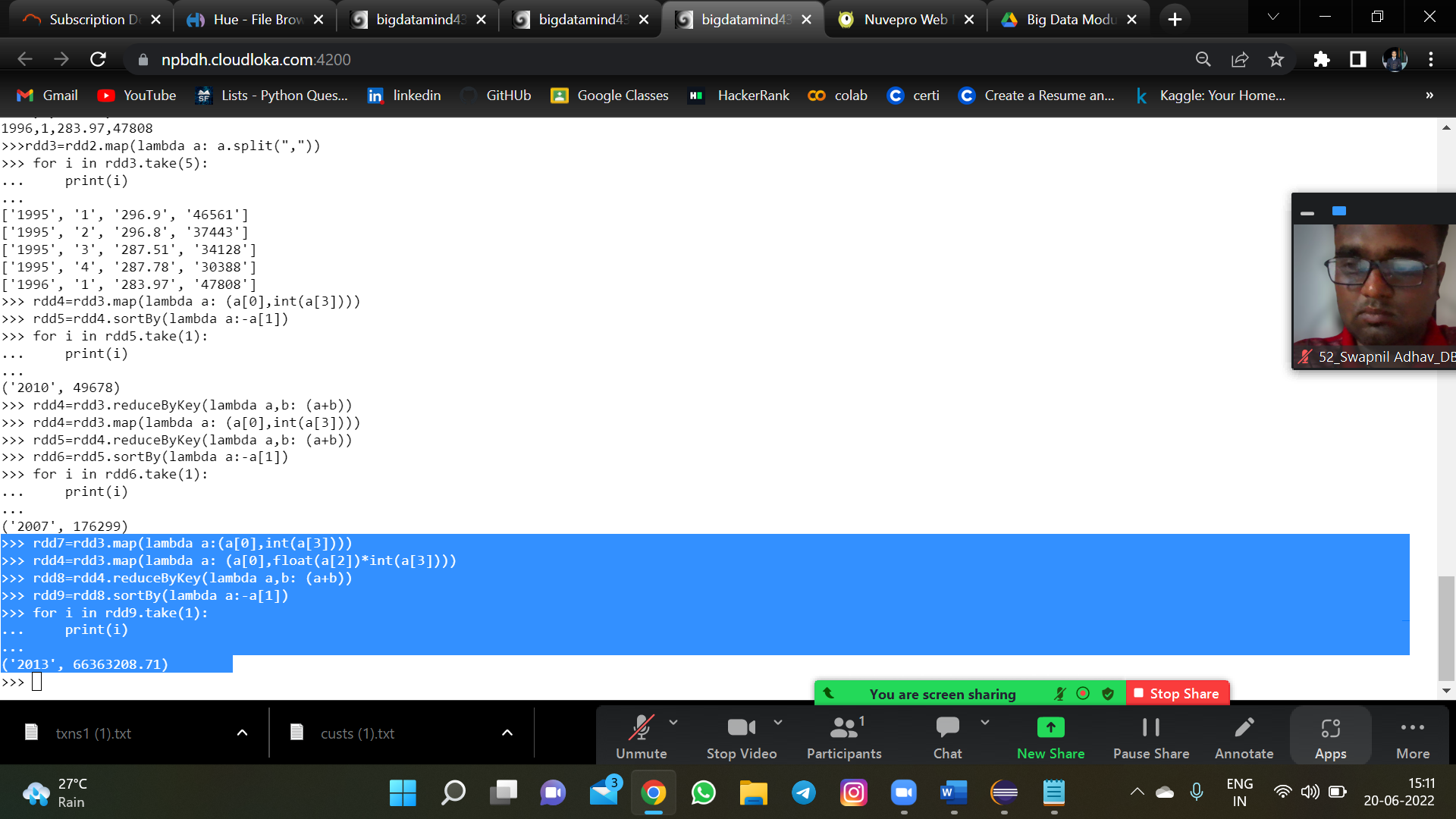
**Q3)** **PySpark**

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

**Ans :-2007 (176299 people travelled)**

****

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

**Ans :- 2013 (66363208.71 revenu)**

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

**group)**

**ans:- 2014 4 (18819408.48 in 4th quarter 2014)**