**Question 2 : Find all time High price for each stock (Hadoop Question)**

**hive[15 marks]**

Please find the customer data set.

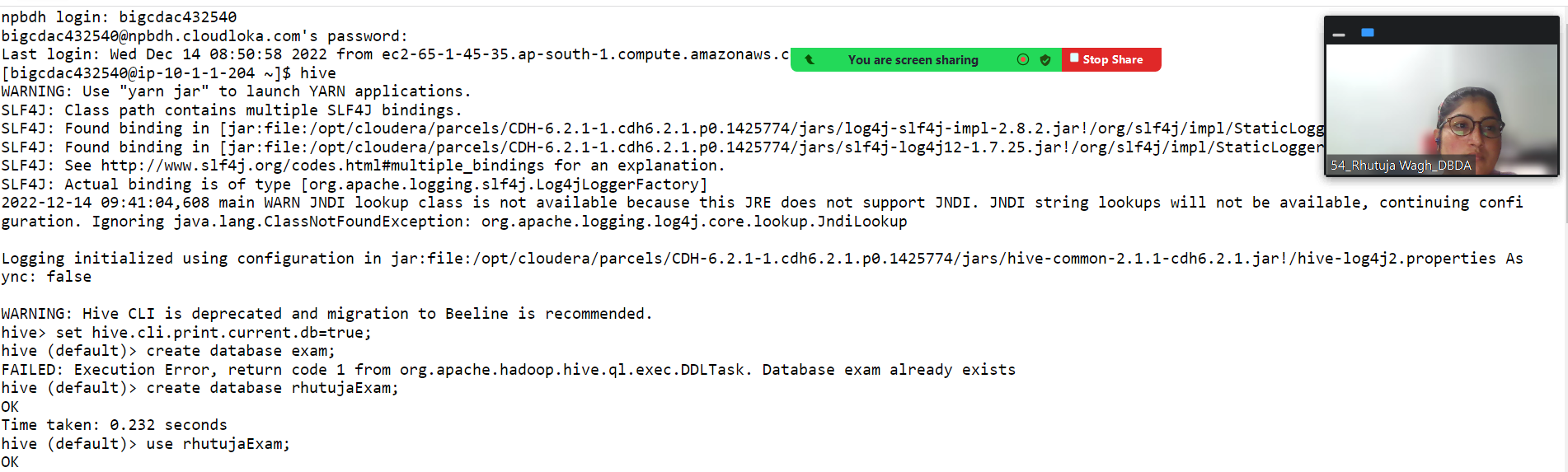
cust id

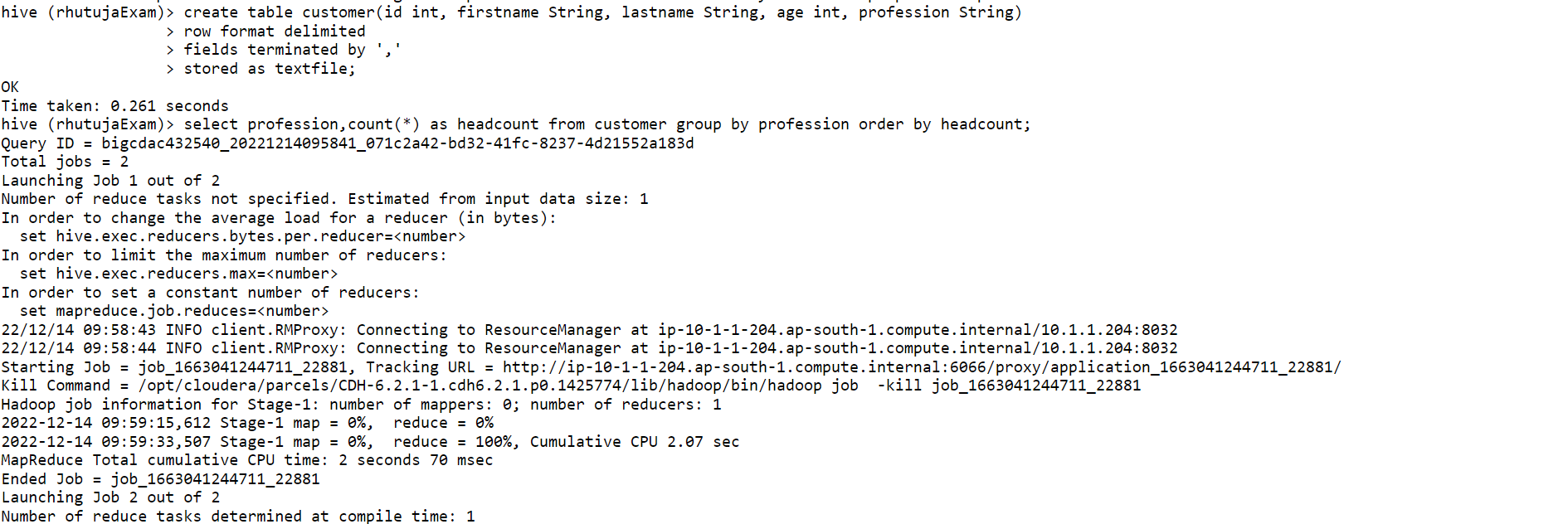
firstname

lastname

age

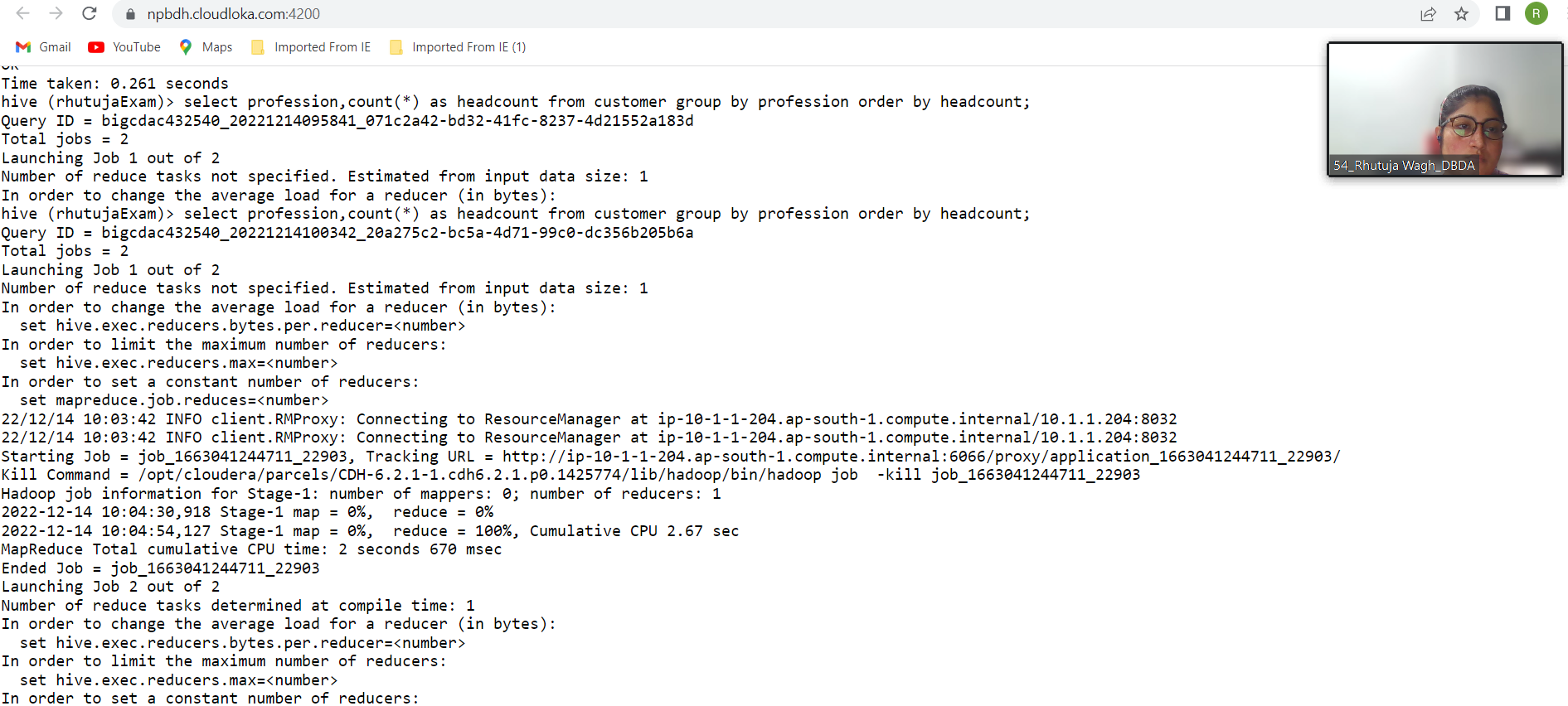
profession

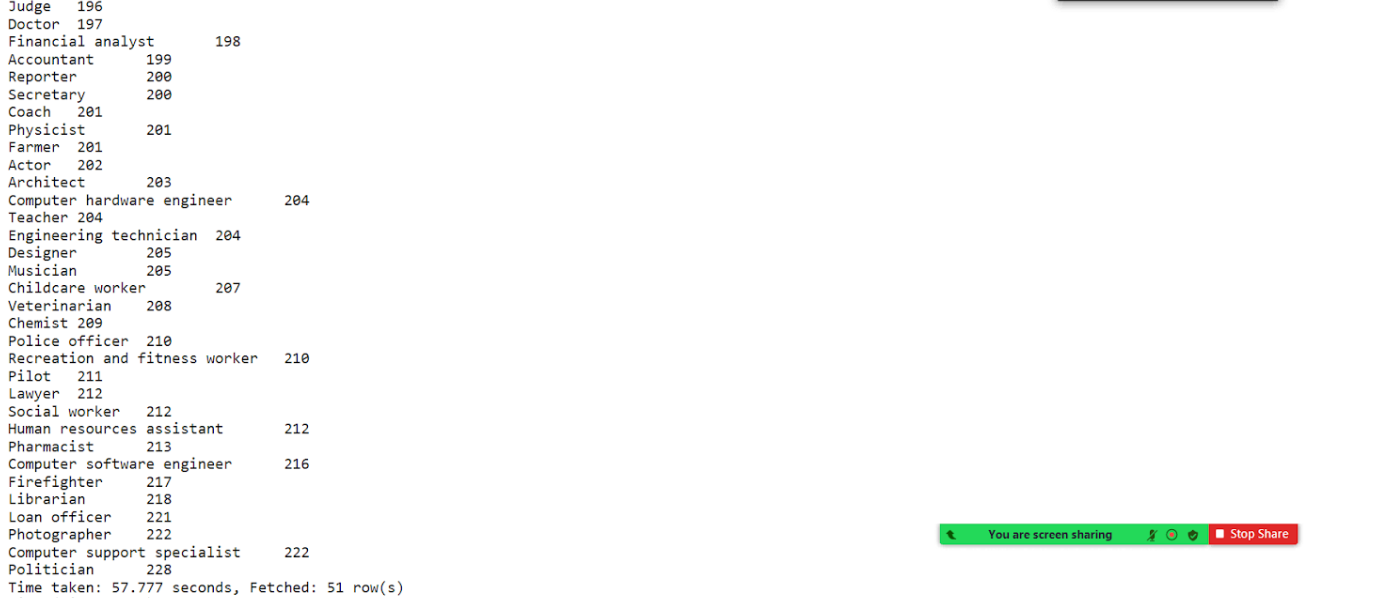




**1) Write a program to find the count of customers for each profession.**

**hive (rhutujaExam)> select profession,count(\*) as headcount from customer group by profession order by headcount;**

****

****

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

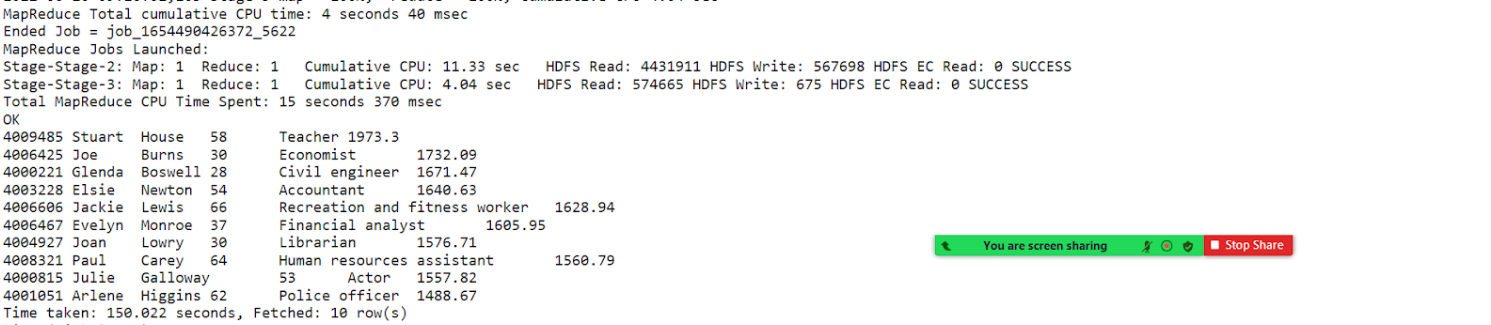
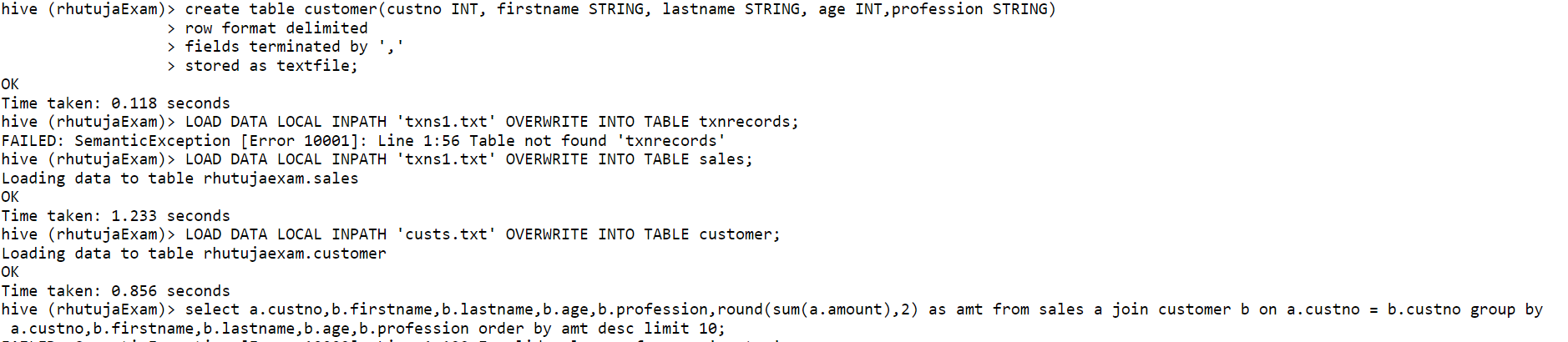
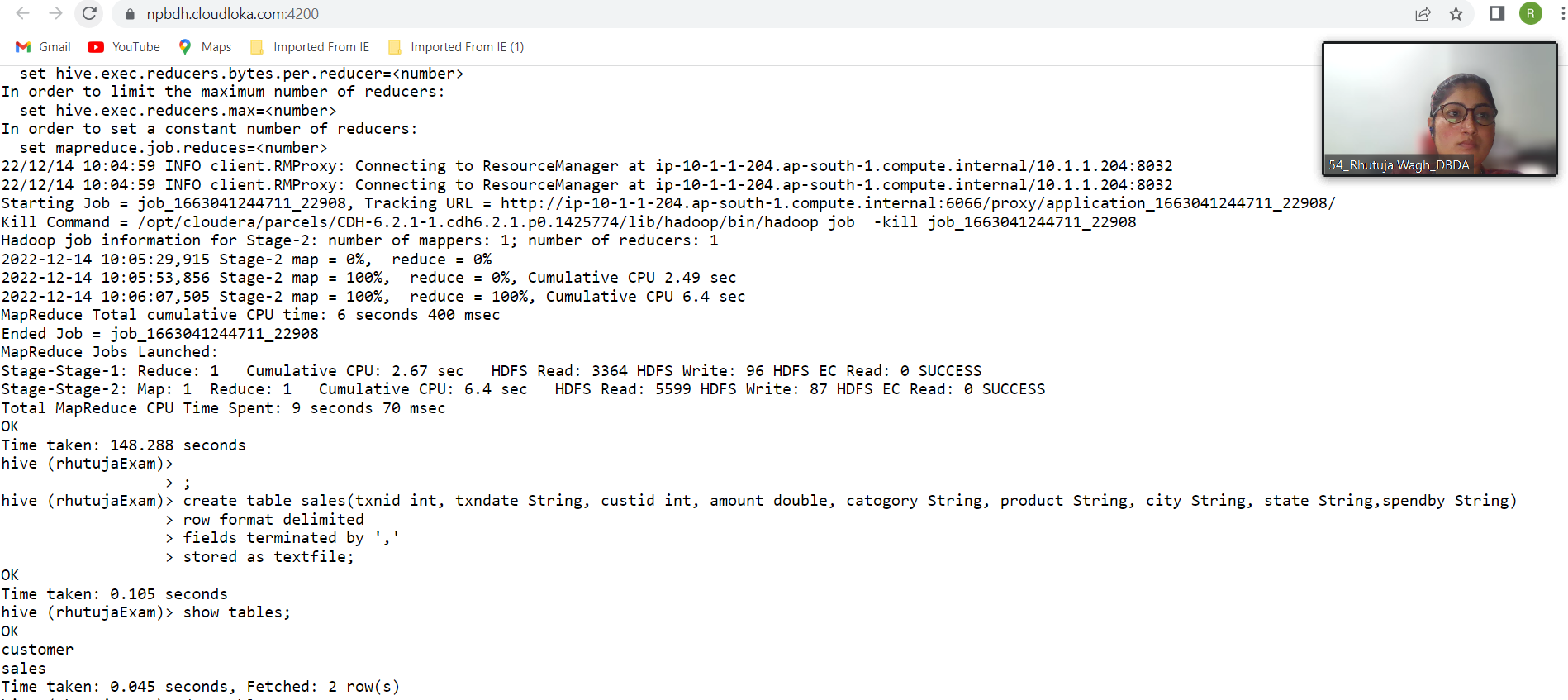
**Write a program to find the top 10 products sales wise**

**create table sales(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;**

****

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

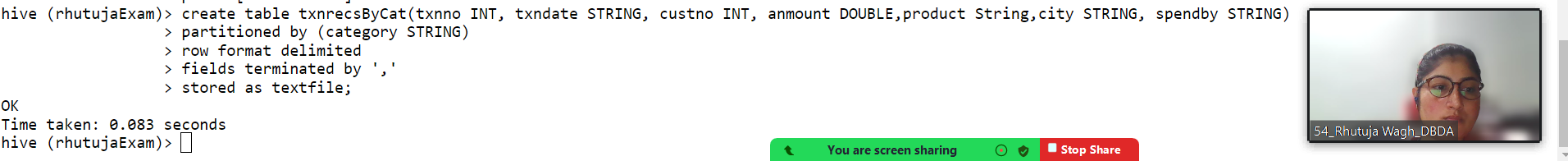
**hive (rhutujaExam)> create table txnrecsByCat(txnno INT, txndate STRING, custno INT, anmount DOUBLE,product String,city STRING, spendby STRING)**

**> partitioned by (category STRING)**

**> row format delimited**

**> fields terminated by ','**

**> stored as textfile;**

****

**QUESTION 3 [15 marks]**

**PySpark**

Please find the AIRLINES data set

Year

Quarter

Average revenue per seat

Total number of booked seats

dataRDD = sc.textFile("hdfs://nameservice1/user/bigcdac432540/projects/airlines.csv")

header = data.first()

data2 = data.filter(lambda a : a != header)

data3 = data2.map(lambda a : a.encode("ascii","ignore"))

data4 = data3.map(lambda a : a.split(","))

data5 = data4.map(lambda a : (a[0]+" "+a[1],float(a[2])\*int(a[3])))

key = RDD5.reduceByKey(lambda a,b : a+b)

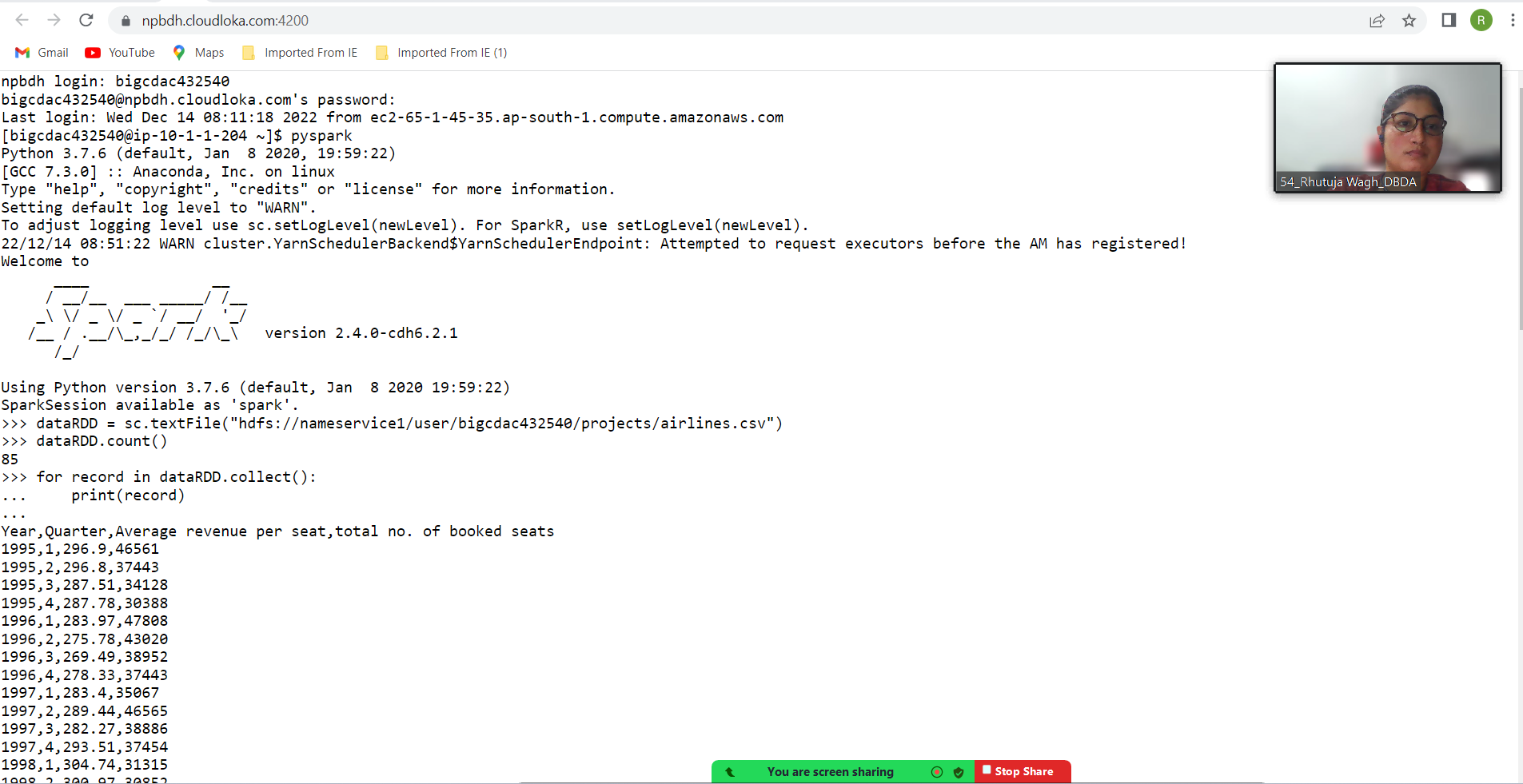
total = counts2.sortBy(lambda a : -a[1])

for i in total.take(1):

... print(i)

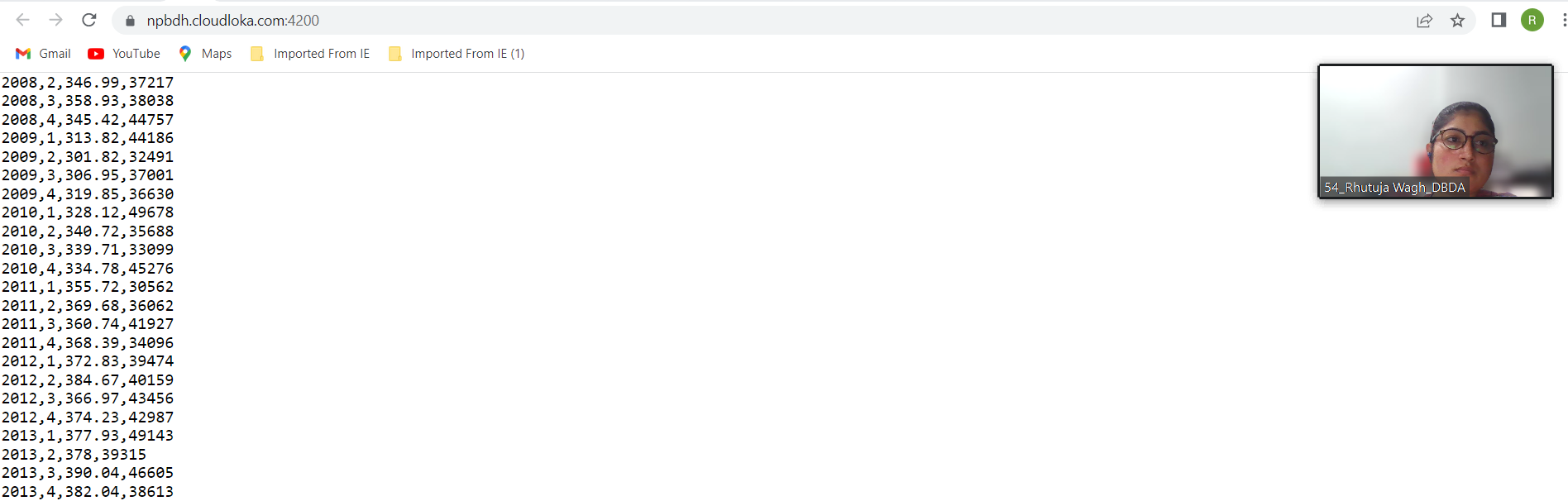
...

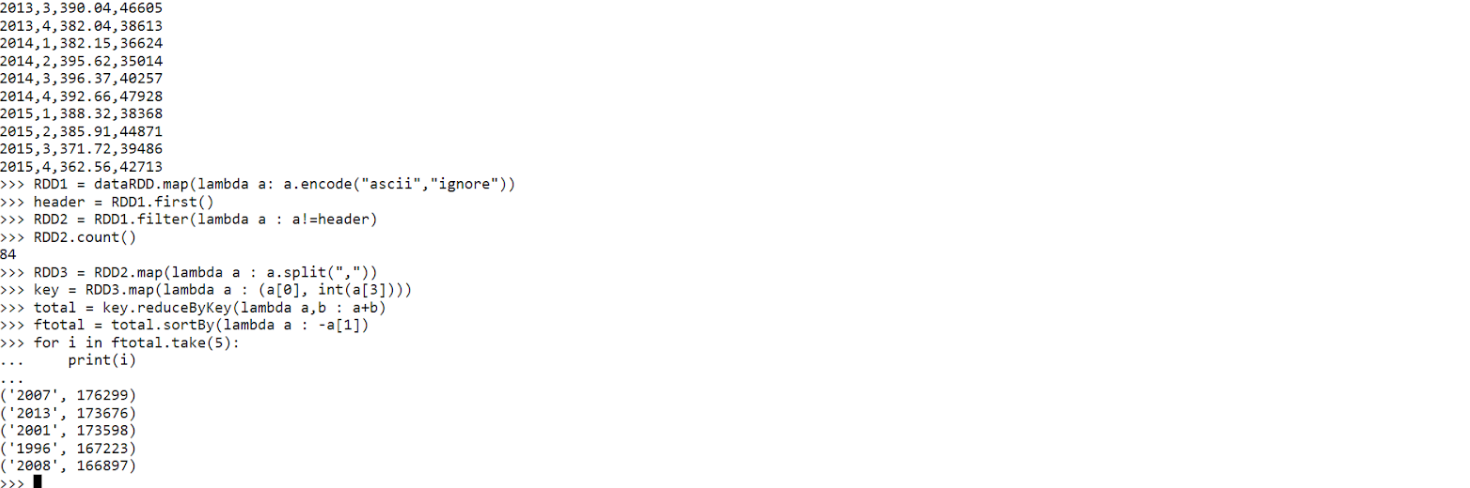
('2014 4', 18819408.48)



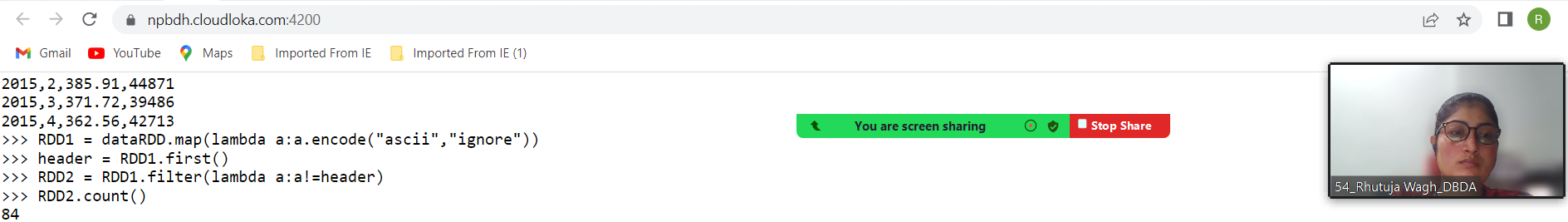
**1) What was the highest number of people travelled in which**

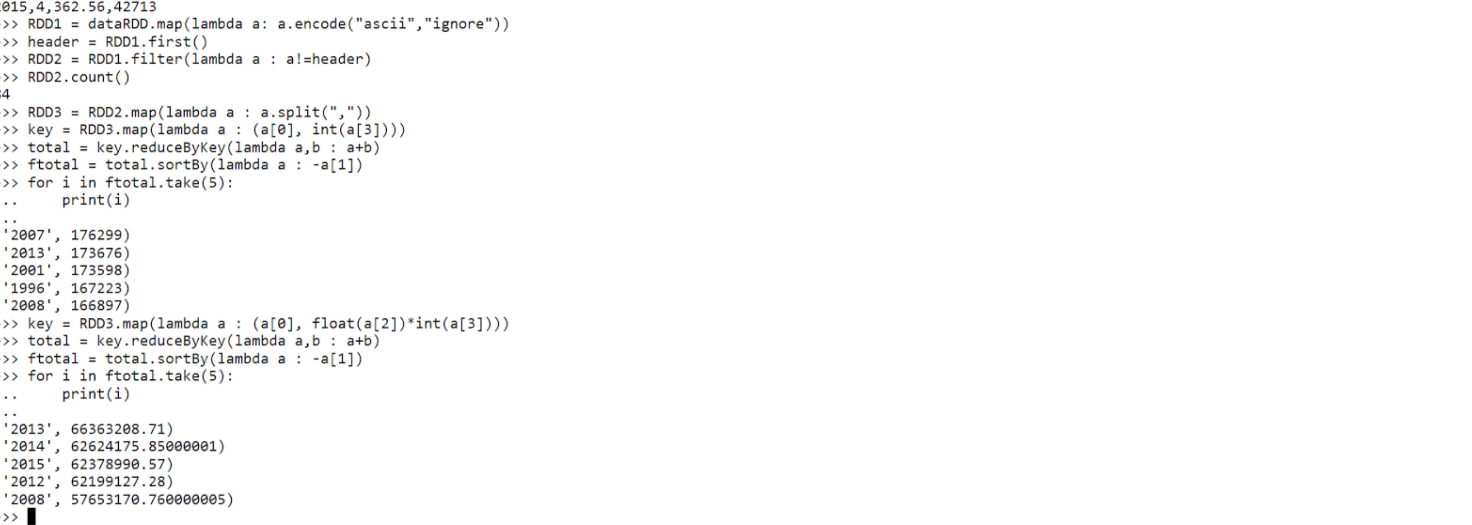
**Year?**

****

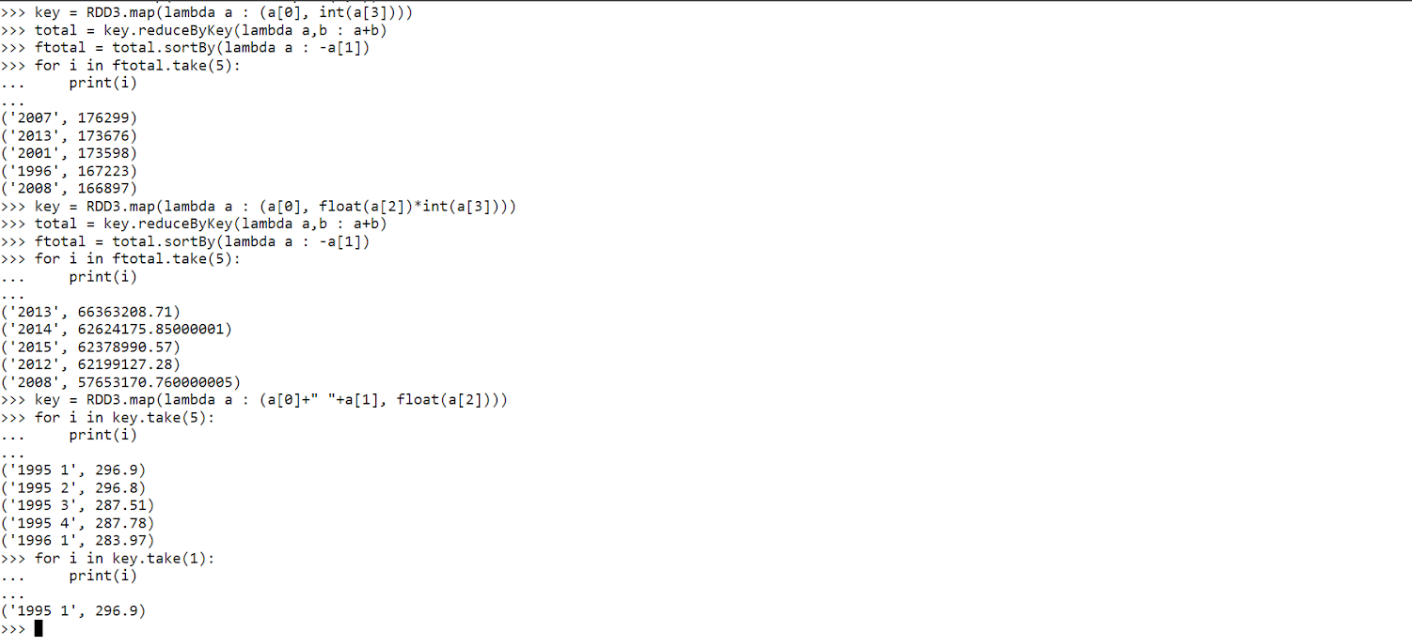
****

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

****

****

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



**Q1.**

**MapReduce**

**Problem Statement [10 marks]**

Here, we have chosen the stock market dataset on which we have performed map-reduce operations. Following is the structure of the data. Kindly Find the solutions to the questions below.

Data Structure

1. Exchange Name

2 Stock symbol

3. Transaction date

4. Opening price of the stock

5. Intra day high price of the stock

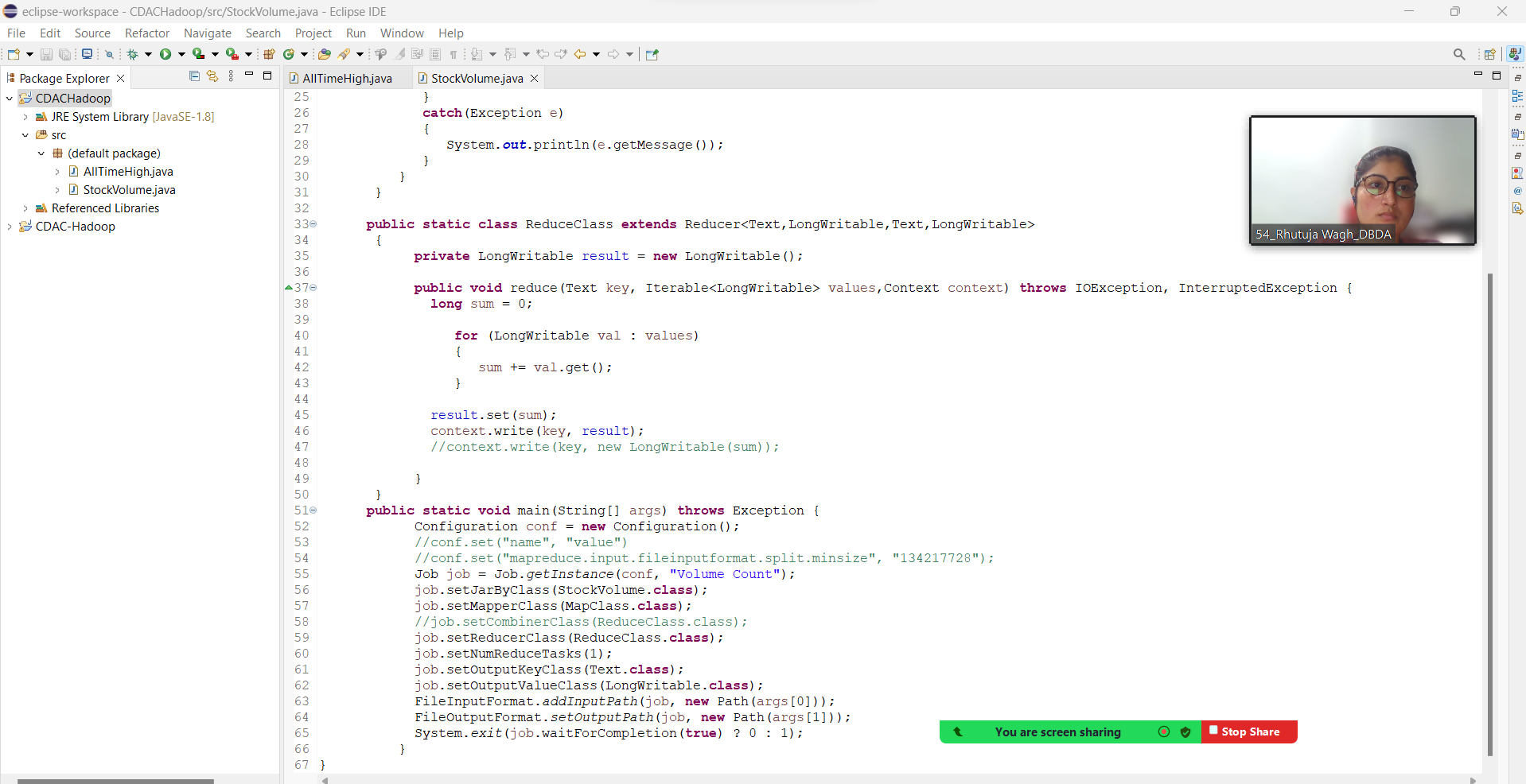
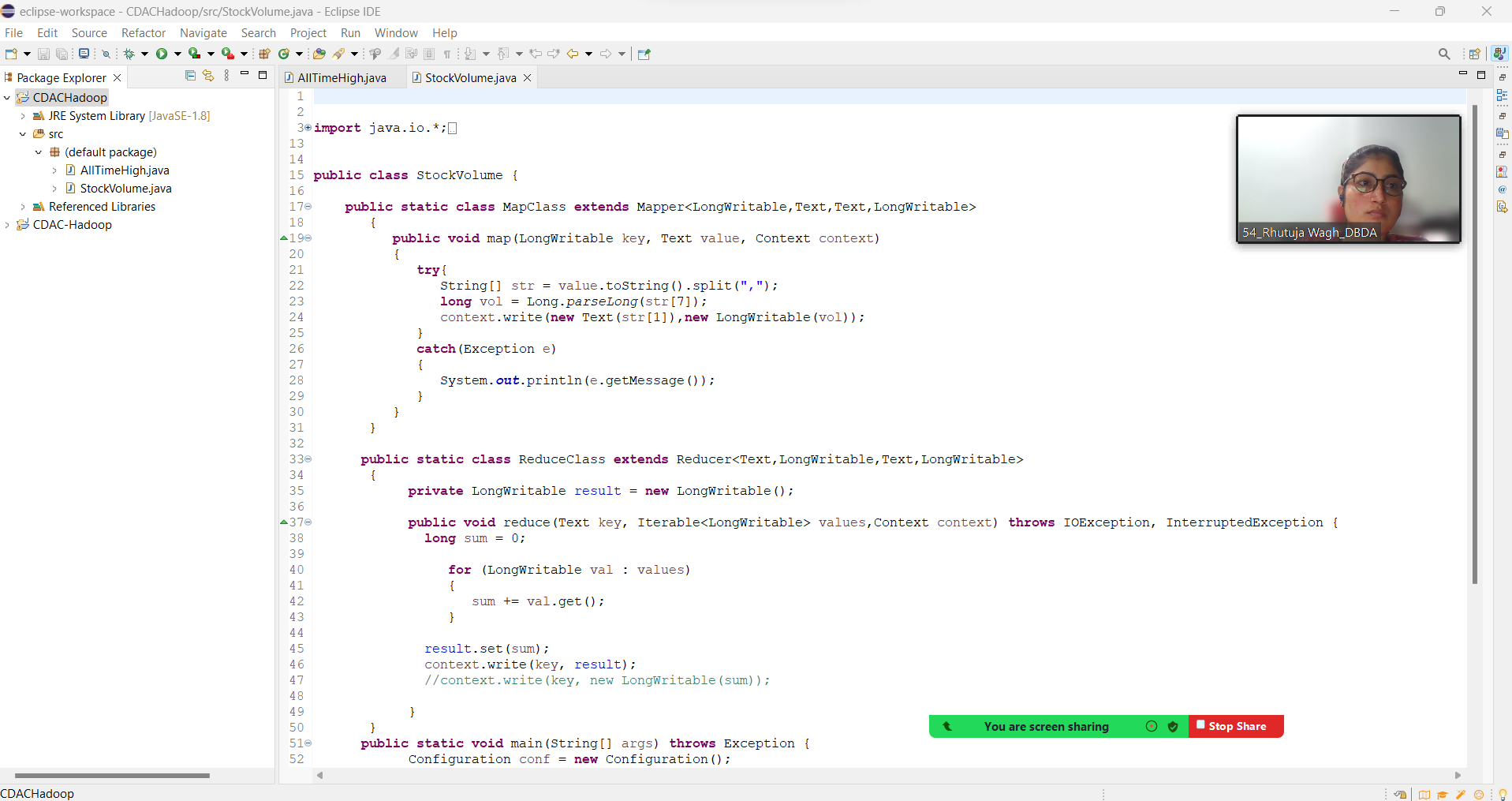
6. Intra day low price of the stock

7. Closing price of the stock

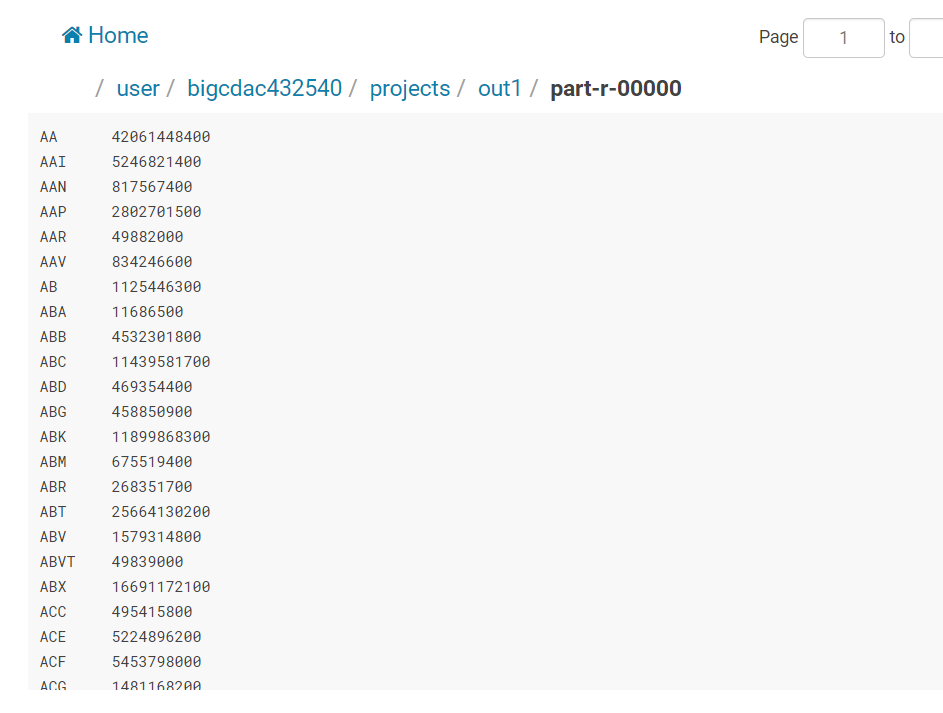
8. Total Volume of the stock on the particular day

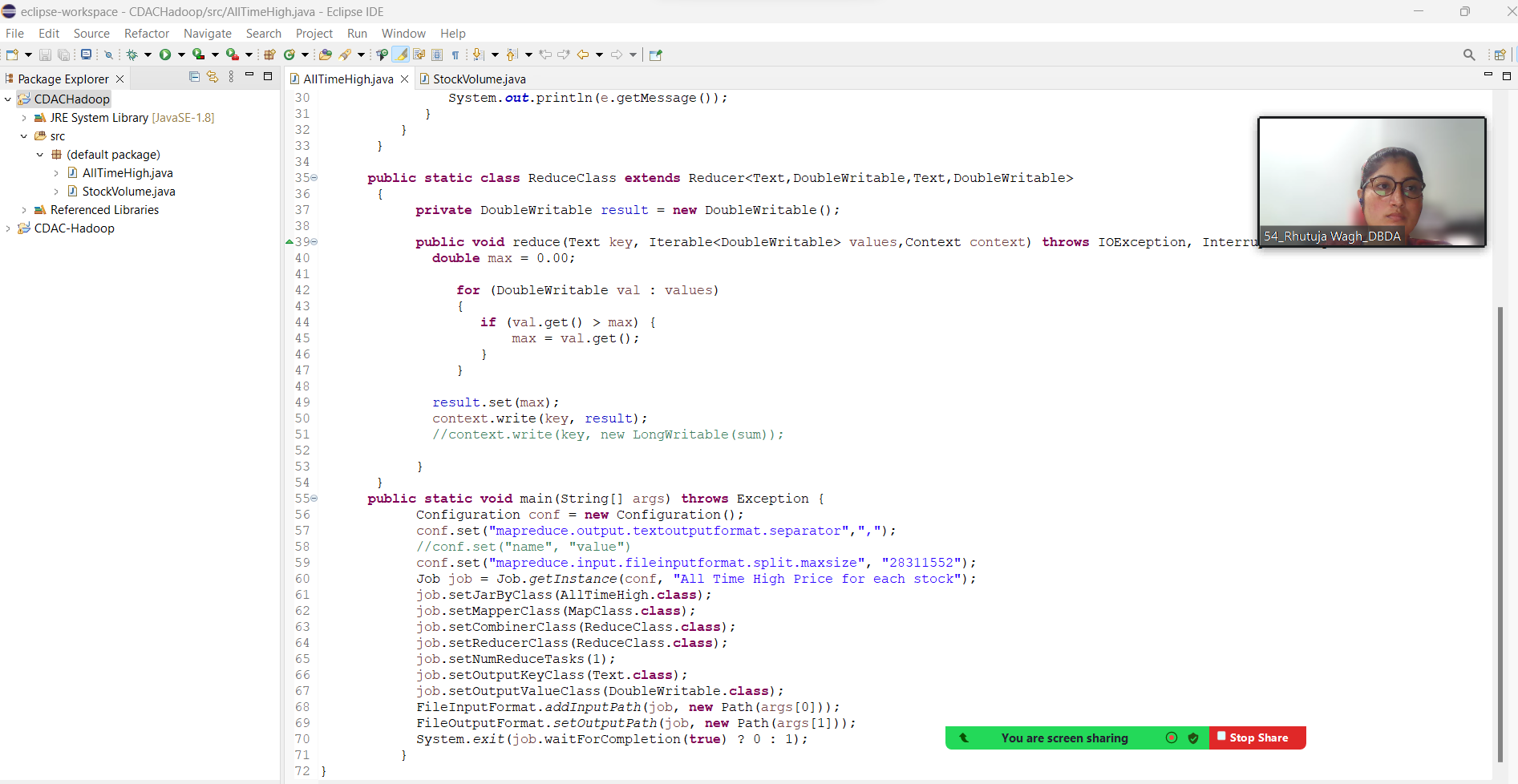
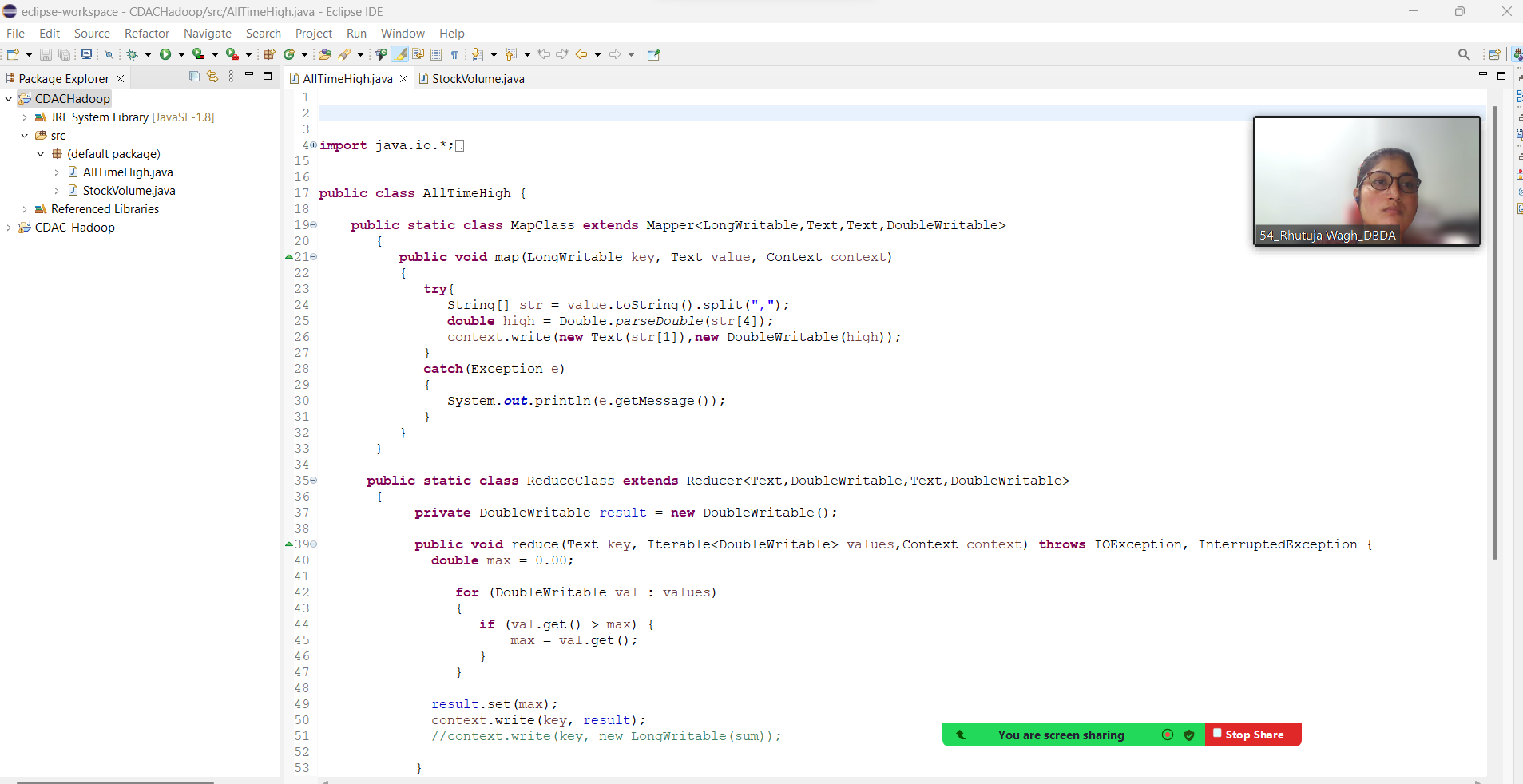
9. Adjustment Closing price of the stock

Field Separator – comma

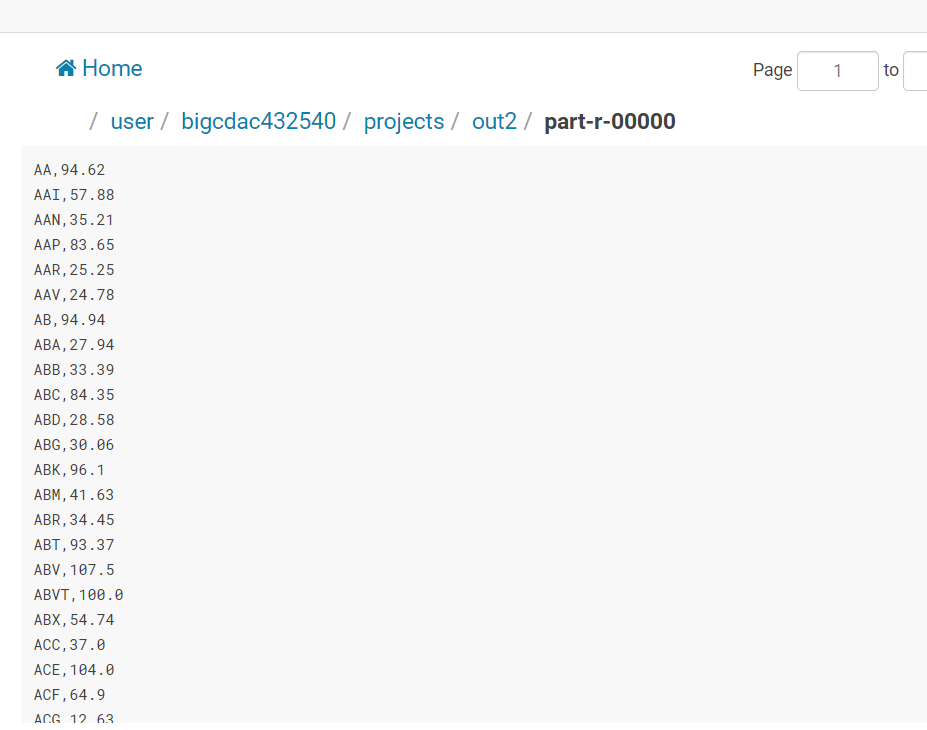


**[bigcdac432540@ip-10-1-1-204 ~]$ hadoop jar myjar.jar StockVolume projects/NYSE.csv projects/out1**





**[bigcdac432540@ip-10-1-1-204 ~]$ hadoop jar myjar.jar AllTimeHigh projects/NYSE.csv projects/out2**

****