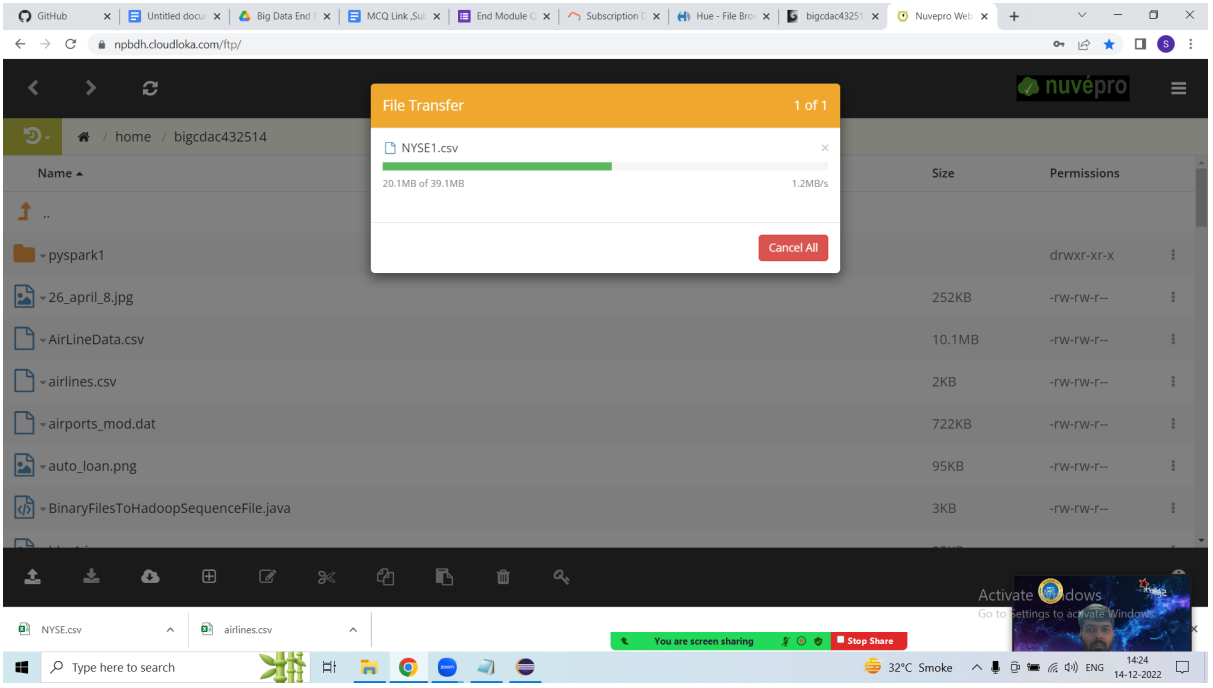
NAME: Sumeet Ramchandra Yadav
PRN NO: 220940325080

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. 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
Find all time High price for each stock

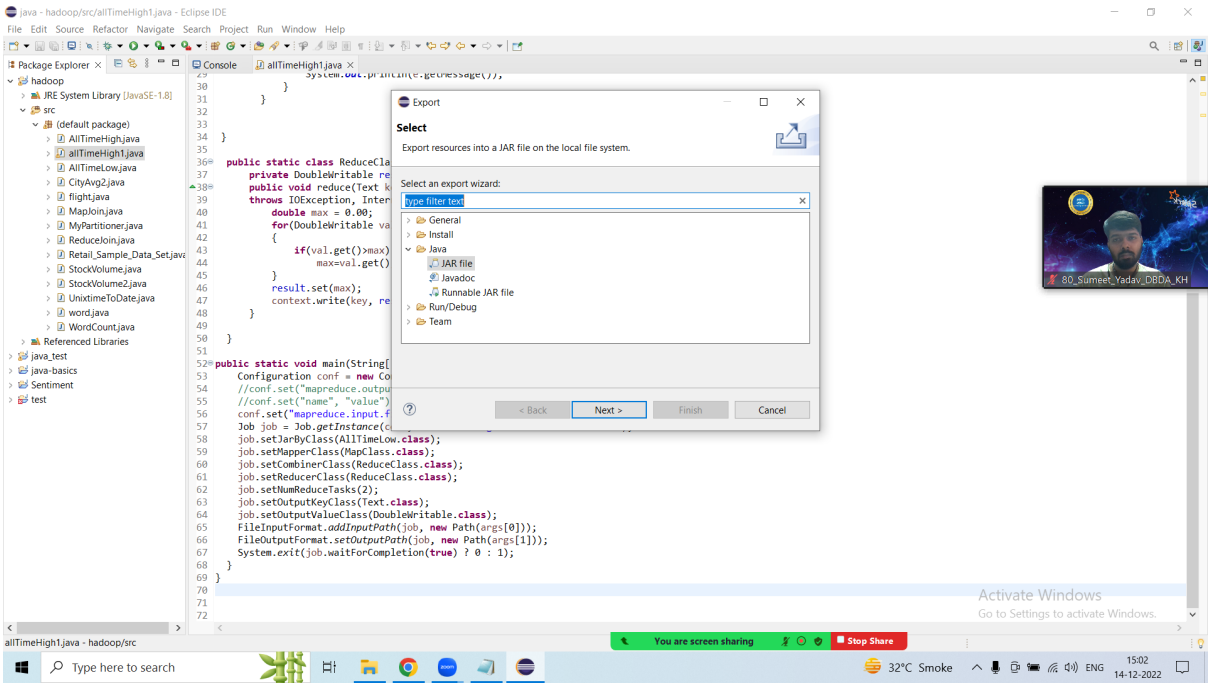
Solution:

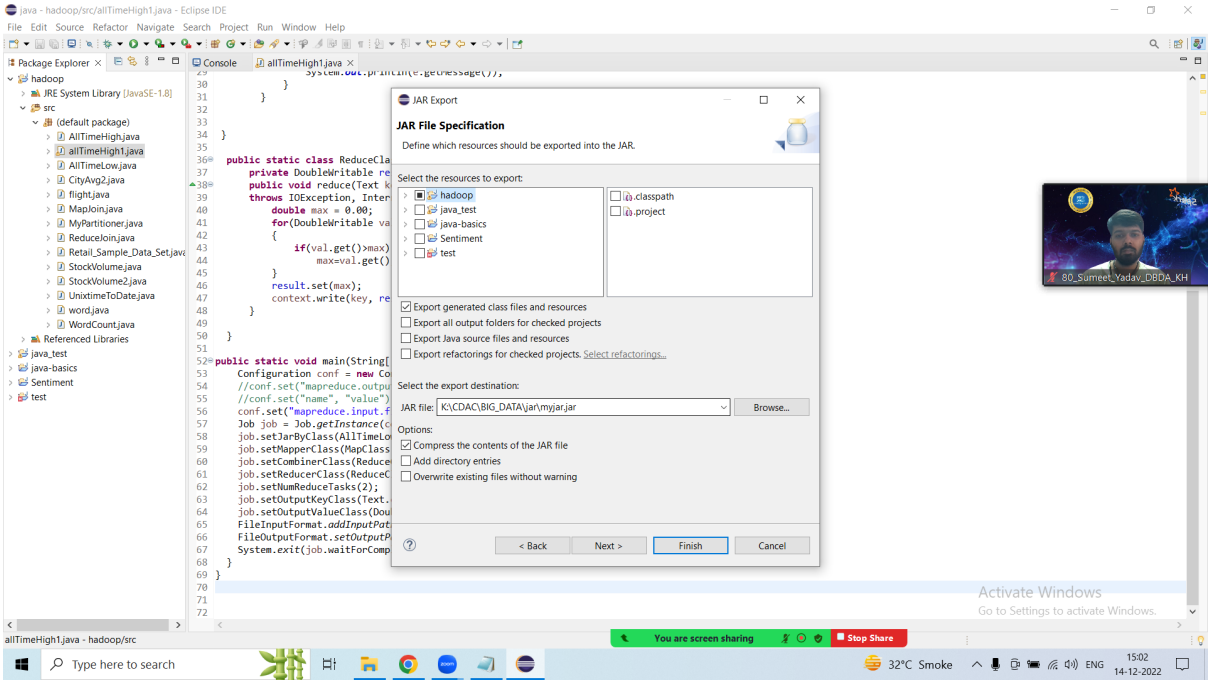


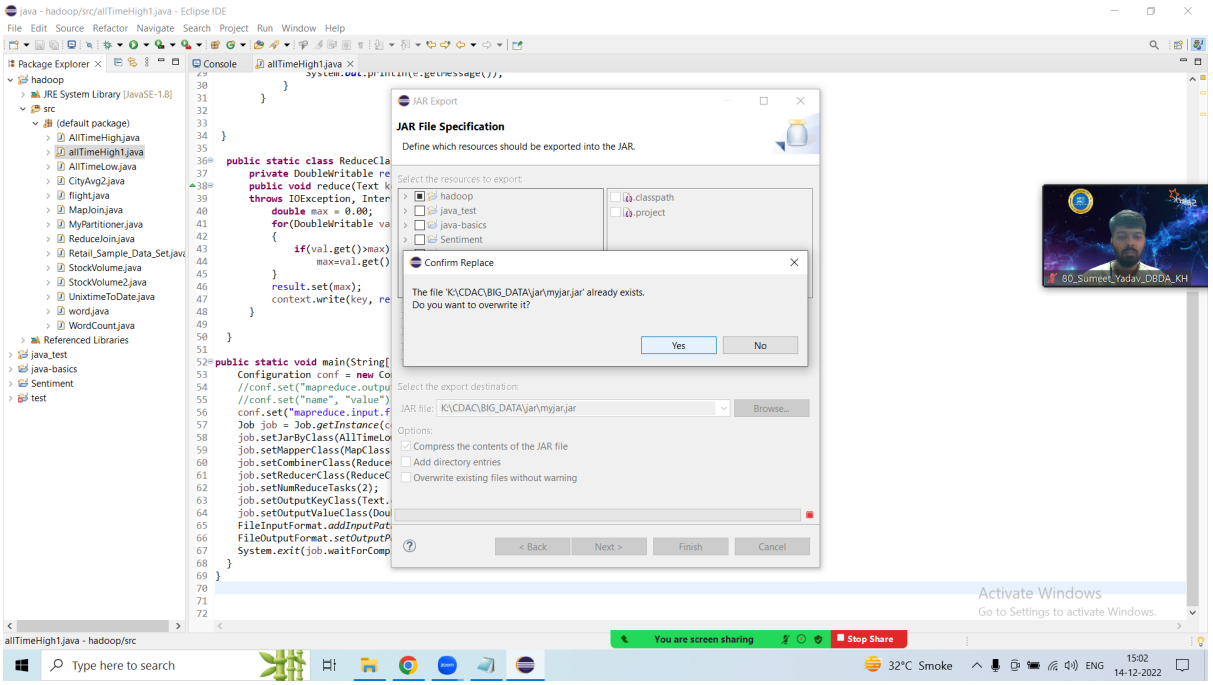
import java.io.*;
import org.apache.hadoop.io.Text;

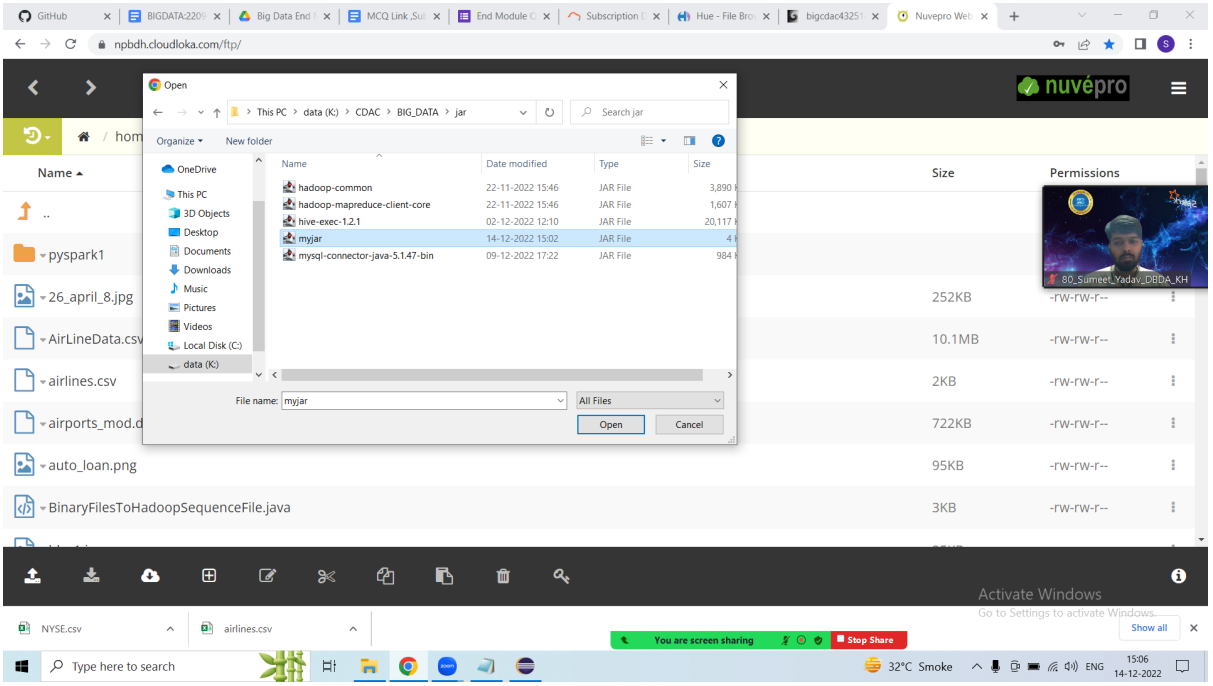
```
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.fs.*;
import org.apache.hadoop.mapreduce.lib.input.*;
import org.apache.hadoop.mapreduce.lib.output.*;
public class allTimeHigh1 {

        public static class MapClass extends
Mapper<LongWritable,Text,Text,DoubleWritable>
          {
            public void map(LongWritable key, Text value, Context context)
            {
              try{
                String[] str = value.toString().split(",");
                Double high  = Double.parseDouble(str[4]);
                context.write(new Text(str[1]),new DoubleWritable(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 max = 0.00;

                        for (DoubleWritable val : values)
                        {
                          if (val.get()>max)
                                    max=val.get();
                        }

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

                    }
```

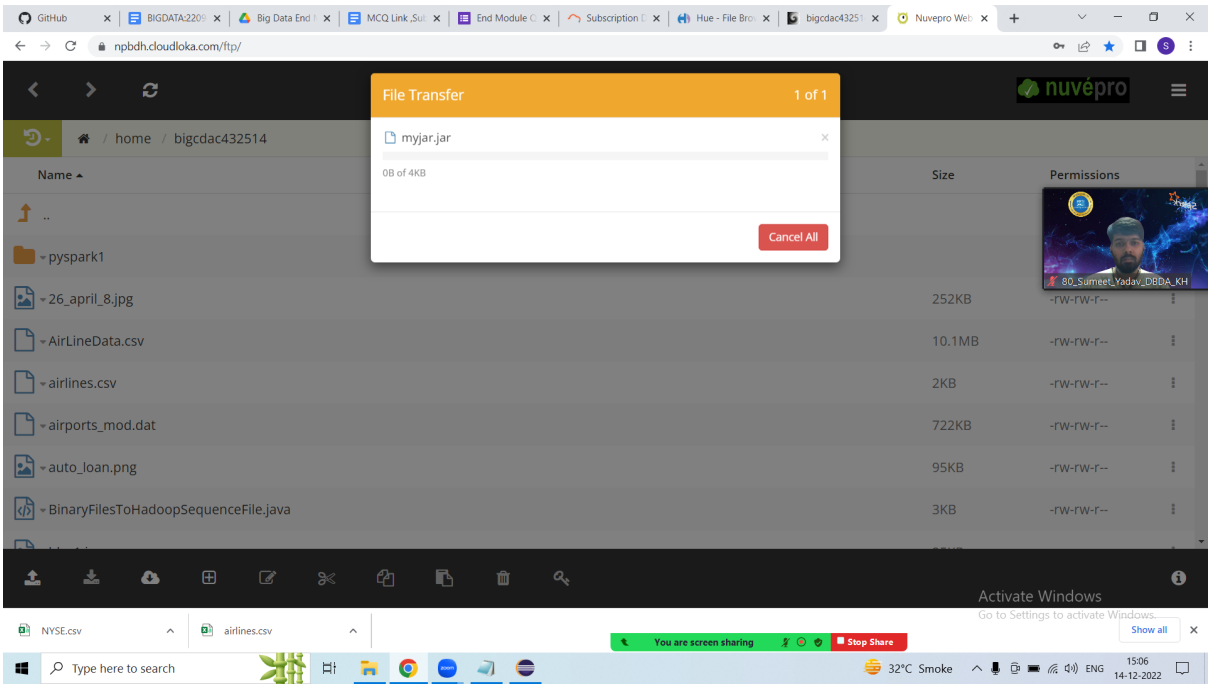
```
        }
        public static void main(String[] args) throws Exception {
                Configuration conf = new Configuration();
                //conf.set("mapreduce.output.textoutputformat.separator",","); to put ,
insted of tab in output
                //conf.set("name", "value")
                conf.set("mapreduce.input.fileinputformat.split.maxsize",
"28311552");
                Job job = Job.getInstance(conf, "All time High Price for each stock");
                job.setJarByClass(allTimeHigh1.class);
                job.setMapperClass(MapClass.class);
                job.setCombinerClass(ReduceClass.class);
                job.setReducerClass(ReduceClass.class);
                job.setNumReduceTasks(2);
                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);
              }
}
```



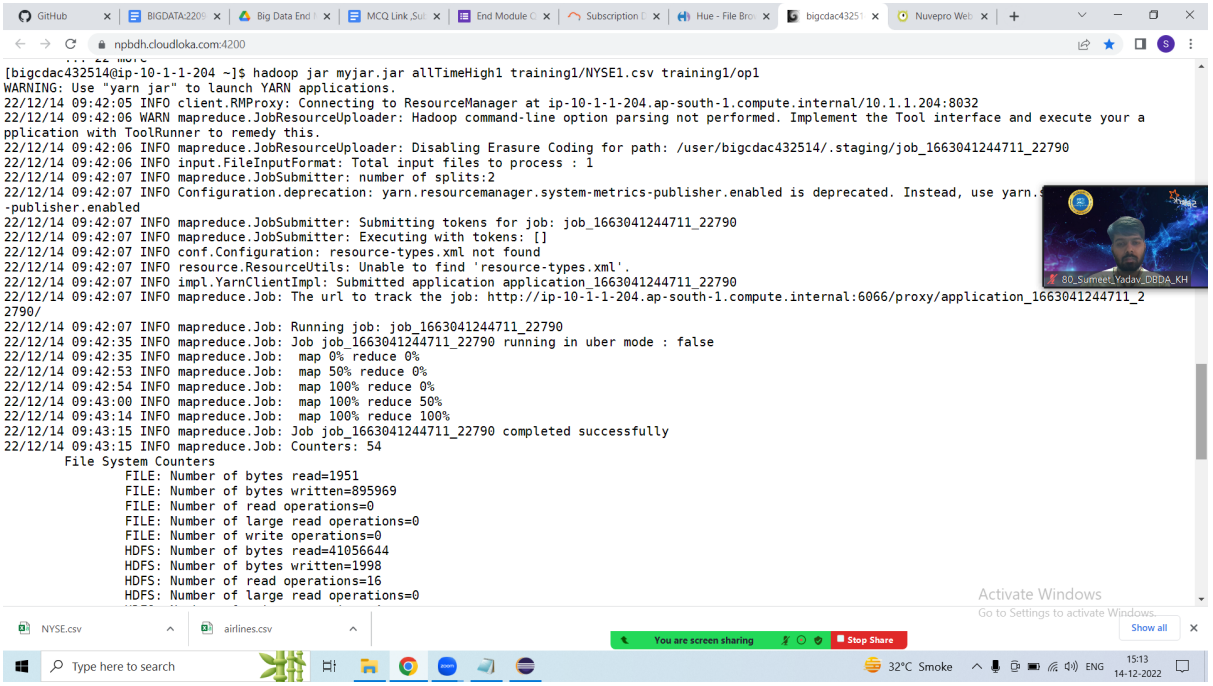

java - hadoop/src/allTimeHigh1.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
Console
allTimeHigh1.java
JAR Export
JAR File Specification
Define which resources should be exported into the JAR.
Select the resources to export:
hadoop
java_test
java-basics
Sentiment
test
.classpath
.project
Export generated class files and resources
Export all output folders for checked projects
Export Java source files and resources
Export refactorings for checked projects. Select refactorings...
Select the export destination:
JAR file: K:\CDAC\BIG_DATA\jar\myjar.jar
Browse...
Options:
Compress the contents of the JAR file
Add directory entries
Overwrite existing files without warning
< Back
Next >
Finish
Cancel
80_Sumeet_Yadav_DBDA_KH
Activate Windows
Go to Settings to activate Windows.
You are screen sharing
Stop Share
allTimeHigh1.java - hadoop/src
Type here to search
32°C Smoke
ENG
15:02
14-12-2022


java - hadoop/src/allTimeHigh1.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
Console
allTimeHigh1.java
JAR Export
JAR File Specification
Define which resources should be exported into the JAR.
Confirm Replace
The file 'K:\CDAC\BIG_DATA\jar\myjar.jar' already exists.
Do you want to overwrite it?
Yes
No
80_Sumeet_Yadav_DBDA_KH
Activate Windows
Go to Settings to activate Windows.
You are screen sharing
Stop Share
allTimeHigh1.java - hadoop/src
Type here to search
32°C Smoke
ENG
15:02
14-12-2022



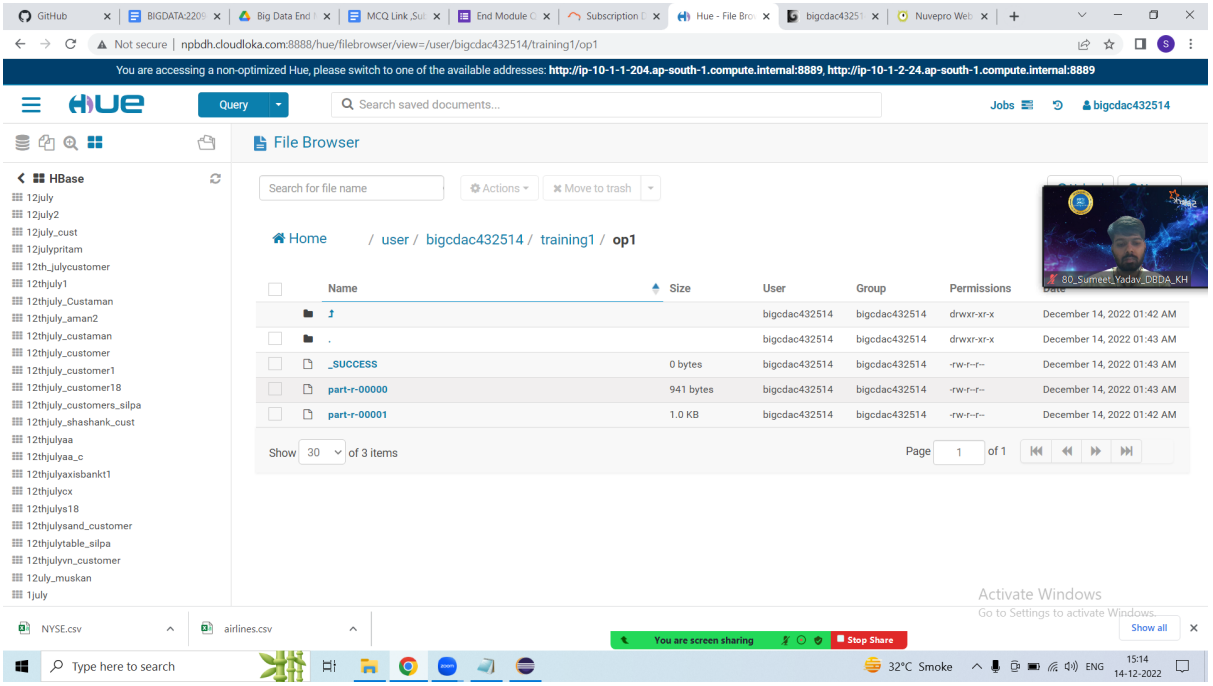


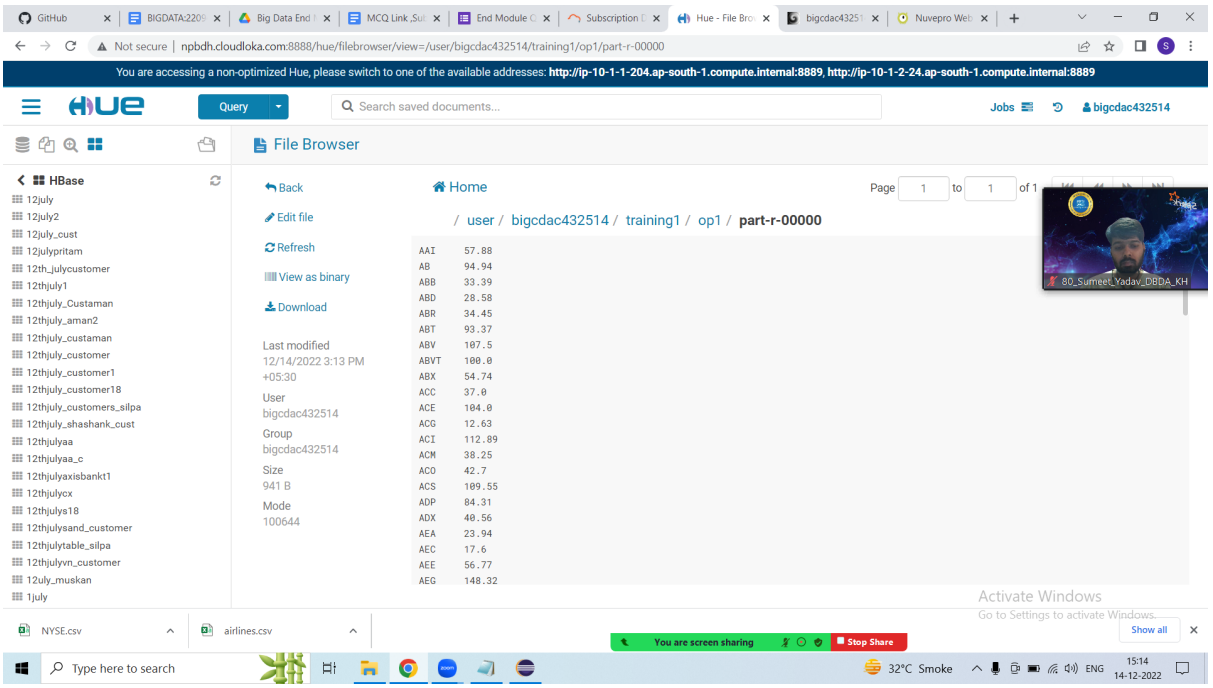

```
hadoop jar myjar.jar allTimeHigh1 training1/NYSE1.csv training1/op1
```

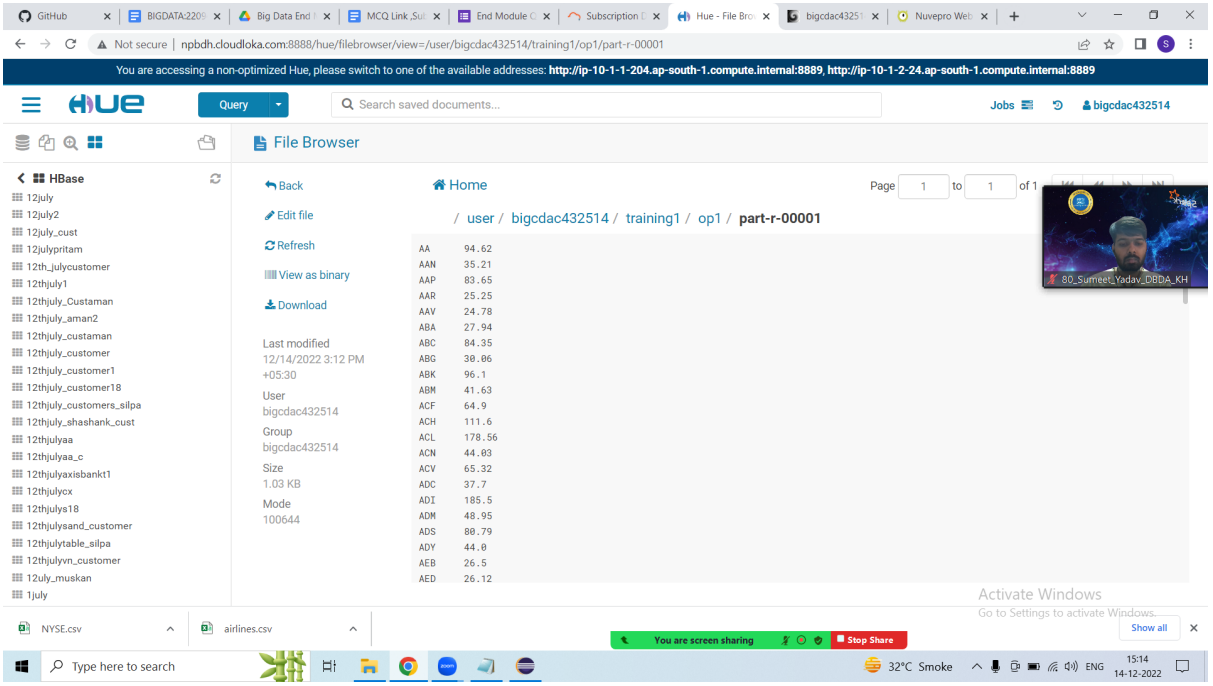

npbdh.cloudloka.com:4200
[bigcdac432514@ip-10-1-1-204 ~]$ hadoop jar myjar.jar allTimeHigh1 training1/NYSE1.csv training1/op1
WARNING: Use "yarn jar" to launch YARN applications.
22/12/14 09:42:05 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
22/12/14 09:42:06 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your a
pplication with ToolRunner to remedy this.
22/12/14 09:42:06 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/bigcdac432514/.staging/job_1663041244711_22790
22/12/14 09:42:06 INFO input.FileInputFormat: Total input files to process : 1
22/12/14 09:42:07 INFO mapreduce.JobSubmitter: number of splits:2
22/12/14 09:42:07 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.
-publisher.enabled
22/12/14 09:42:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1663041244711_22790
22/12/14 09:42:07 INFO mapreduce.JobSubmitter: Executing with tokens: []
22/12/14 09:42:07 INFO conf.Configuration: resource-types.xml not found
22/12/14 09:42:07 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/12/14 09:42:07 INFO impl.YarnClientImpl: Submitted application application_1663041244711_22790
22/12/14 09:42:07 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_2
2790/
22/12/14 09:42:07 INFO mapreduce.Job: Running job: job_1663041244711_22790
22/12/14 09:42:35 INFO mapreduce.Job: Job job_1663041244711_22790 running in uber mode : false
22/12/14 09:42:35 INFO mapreduce.Job: map 0% reduce 0%
22/12/14 09:42:53 INFO mapreduce.Job: map 50% reduce 0%
22/12/14 09:42:54 INFO mapreduce.Job: map 100% reduce 0%
22/12/14 09:43:00 INFO mapreduce.Job: map 100% reduce 50%
22/12/14 09:43:14 INFO mapreduce.Job: map 100% reduce 100%
22/12/14 09:43:15 INFO mapreduce.Job: Job job_1663041244711_22790 completed successfully
22/12/14 09:43:15 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=1951
FILE: Number of bytes written=895969
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=41056644
HDFS: Number of bytes written=1998
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
80_Sumeet_Yadav_DBDA_KH
Activate Windows
Go to Settings to activate Windows.
NYSE.csv
airlines.csv
You are screen sharing
Stop Share
Show all
Type here to search
32°C Smoke
ENG
15:13
14-12-2022


npbdh.cloudloka.com:4200
Map input records=735026
Map output records=735026
Map output bytes=8781587
Map output materialized bytes=2132
Input split bytes=246
Combine input records=735026
Combine output records=204
Reduce input groups=203
Reduce shuffle bytes=2132
Reduce input records=204
Reduce output records=203
Spilled Records=408
Shuffled Maps =4
Failed Shuffles=0
Merged Map outputs=4
GC time elapsed (ms)=936
CPU time spent (ms)=10680
Physical memory (bytes) snapshot=1733894144
Virtual memory (bytes) snapshot=10373488640
Total committed heap usage (bytes)=1977614336
Peak Map Physical memory (bytes)=637186048
Peak Map Virtual memory (bytes)=2587807744
Peak Reduce Physical memory (bytes)=267534336
Peak Reduce Virtual memory (bytes)=2606125056
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=41056398
File Output Format Counters
Bytes Written=1998
[bigcdac432514@ip-10-1-1-204 ~]$
80_Sumeet_Yadav_DBDA_KH
Activate Windows
Go to Settings to activate Windows.
NYSE.csv
airlines.csv
You are screen sharing
Stop Share
Show all
Type here to search
32°C Smoke
ENG
15:13
14-12-2022

Output:


Not secure | npbdh.cloudloka.com:8888/hue/filebrowser/view=/user/bigcdac432514/training1/op1
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-1-1-204.ap-south-1.compute.internal:8889, http://ip-10-1-2-24.ap-south-1.compute.internal:8889
Query
Search saved documents...
Jobs
bigcdac432514
File Browser
HBase
Search for file name
Actions
Move to trash
Home / user / bigcdac432514 / training1 / op1
Name Size User Group Permissions
_SUCCESS 0 bytes bigcdac432514 bigcdac432514 -rw-r--r-- December 14, 2022 01:43 AM
part-r-00000 941 bytes bigcdac432514 bigcdac432514 -rw-r--r-- December 14, 2022 01:43 AM
part-r-00001 1.0 KB bigcdac432514 bigcdac432514 -rw-r--r-- December 14, 2022 01:42 AM
Show 30 of 3 items
Page 1 of 1
Activate Windows
You are screen sharing
Stop Share

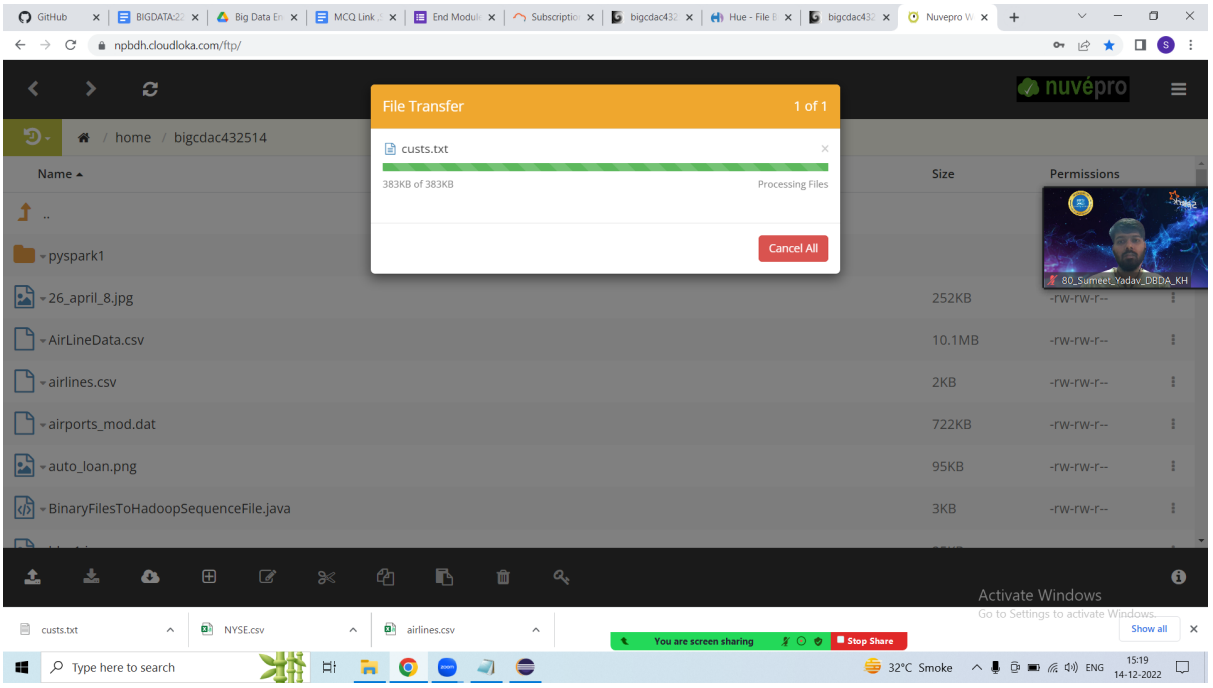

Not secure | npbdh.cloudloka.com:8888/hue/filebrowser/view=/user/bigcdac432514/training1/op1/part-r-00000
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-1-1-204.ap-south-1.compute.internal:8889, http://ip-10-1-2-24.ap-south-1.compute.internal:8889
Query
Search saved documents...
Jobs
bigcdac432514
File Browser
HBase
Back
Edit file
Refresh
View as binary
Download
Last modified 12/14/2022 3:13 PM +05:30
User bigcdac432514
Group bigcdac432514
Size 941 B
Mode 100644
Home / user / bigcdac432514 / training1 / op1 / part-r-00000
Page 1 to 1 of 1
AAI 57.88
AB 94.94
ABB 33.39
ABD 28.58
ABR 34.45
ABT 93.37
ABV 107.5
ABVT 100.0
ABX 54.74
ACC 37.0
ACE 104.0
ACG 12.63
ACI 112.89
ACM 38.25
ACO 42.7
ACS 109.55
ADP 84.31
ADX 40.56
AEA 23.94
AEC 17.6
AEE 56.77
AEG 148.32
Activate Windows
You are screen sharing
Stop Share


GitHub
BIGDATA:220
Big Data End
MCQ Link ,Su
End Module
Subscription
Hue - File Bro
bigcdac4325
Nuvepro Web
Not secure | npbdh.cloudloka.com:8888/hue/filebrowser/view=/user/bigcdac432514/training1/op1/part-r-00001
You are accessing a non-optimized Hue, please switch to one of the available addresses: http://ip-10-1-1-204.ap-south-1.compute.internal:8889, http://ip-10-1-2-24.ap-south-1.compute.internal:8889
Query
Search saved documents...
Jobs
bigcdac432514
File Browser
HBase
12july
12july2
12july_cust
12julypritam
12th_julycustomer
12thjuly1
12thjuly_Custaman
12thjuly_aman2
12thjuly_custaman
12thjuly_customer
12thjuly_customer1
12thjuly_customer18
12thjuly_customers_silpa
12thjuly_shashank_cust
12thjulyaa
12thjulyaa_c
12thjulyaxisbankt1
12thjulycx
12thjulys18
12thjulysand_customer
12thjulytable_silpa
12thjulyvn_customer
12uly_muskan
1july
Back
Edit file
Refresh
View as binary
Download
Last modified
12/14/2022 3:12 PM
+05:30
User
bigcdac432514
Group
bigcdac432514
Size
1.03 KB
Mode
100644
Home
/ user / bigcdac432514 / training1 / op1 / part-r-00001
Page 1 to 1 of 1
AA 94.62
AAN 35.21
AAP 83.65
AAR 25.25
AAV 24.78
ABA 27.94
ABC 84.35
ABG 30.06
ABK 96.1
ABM 41.63
ACF 64.9
ACH 111.6
ACL 178.56
ACN 44.03
ACV 65.32
ADC 37.7
ADI 185.5
ADM 48.95
ADS 80.79
ADY 44.0
AEB 26.5
AED 26.12
80_Sumeet_Yadav_DBDA_KH
Activate Windows
Go to Settings to activate Windows.
NYSE.csv
airlines.csv
You are screen sharing
Stop Share
Show all
Type here to search
32°C Smoke
ENG
15:14
14-12-2022

Question 2 :

[15 marks]
Log in to hive using hive cmd

```
use training432514;
```



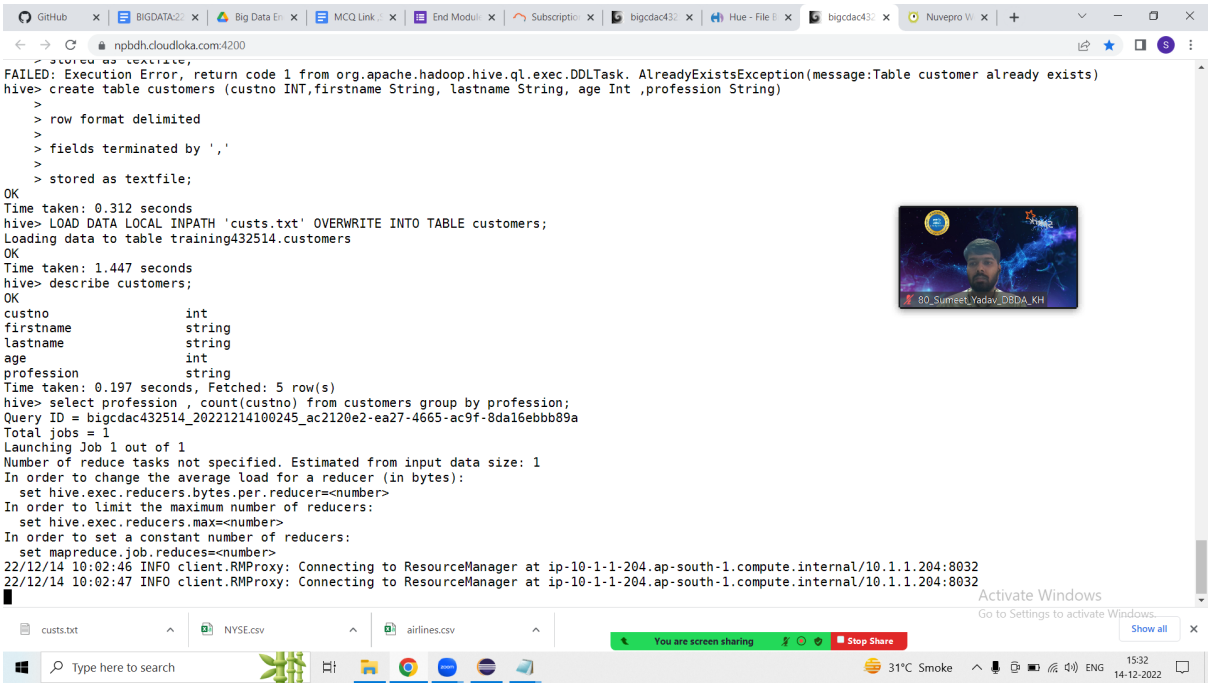

```
hive> create table customers (custno INT,firstname String, lastname
String, age Int ,profession String)
    >
    > row format delimited
    >
    > fields terminated by ','
    >
    > stored as textfile;
OK
Time taken: 0.312 seconds

hive> LOAD DATA LOCAL INPATH 'custs.txt' OVERWRITE INTO TABLE
customers;
Loading data to table training432514.customers
OK
Time taken: 1.447 seconds
```

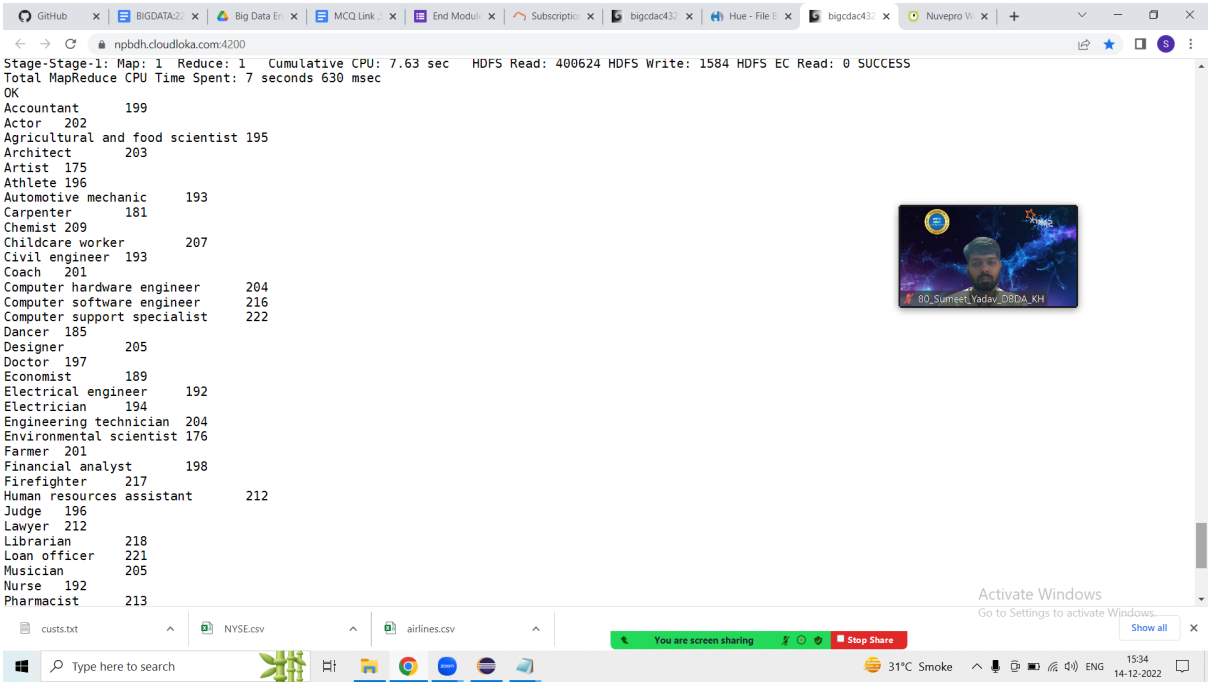
Hive
Please find the customer data set.
cust id
firstname
lastname
age
profession

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

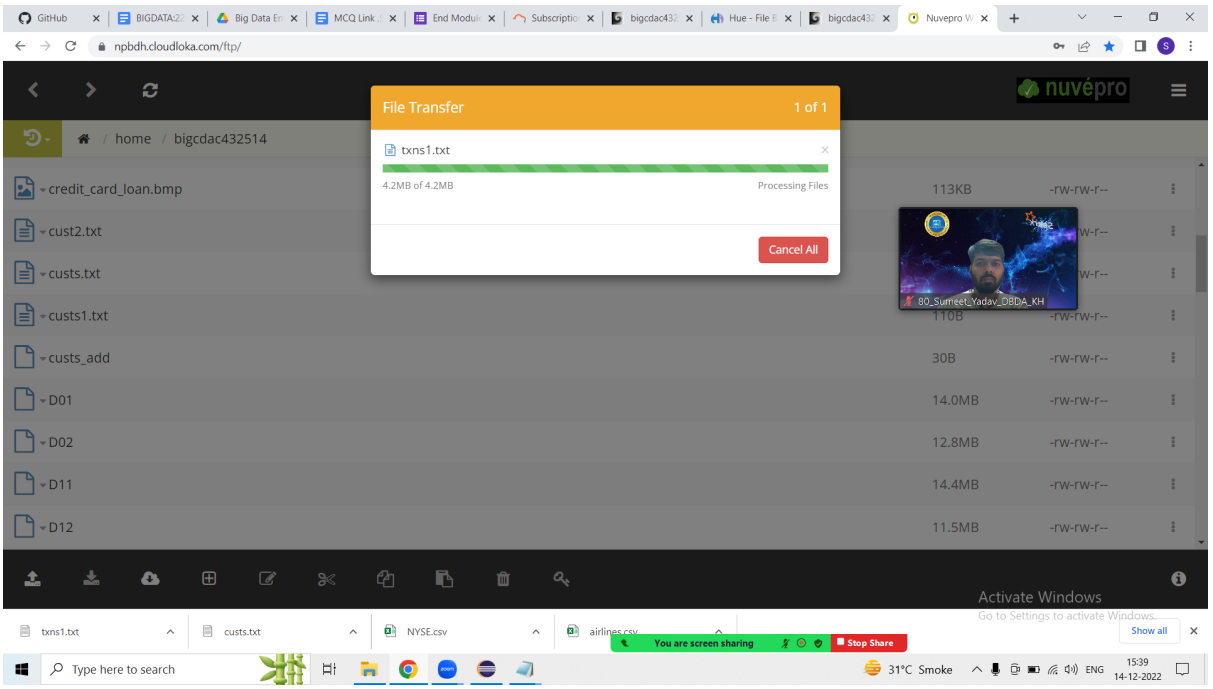
Please find the sales data set.
txn id
txn date
cust id
amount
category
product
city
state
Spendby

Ans
select profession , count(custno) from customers group by profession;



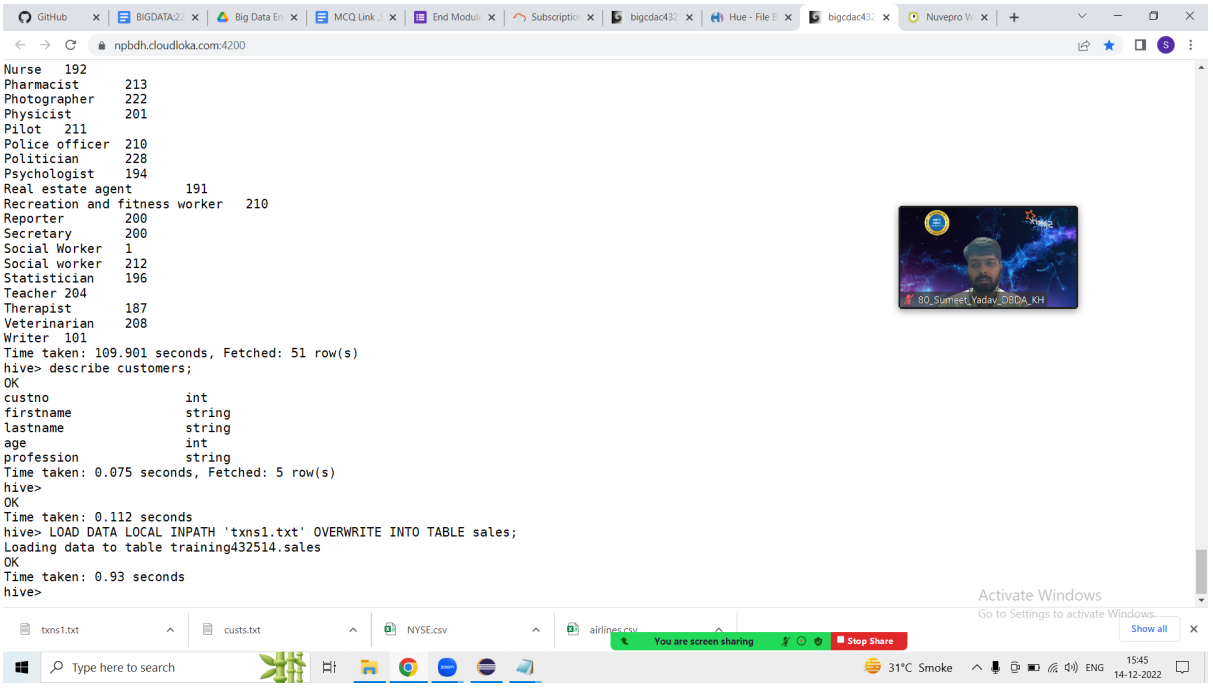


2) Write a program to find the top 10 products sales wise



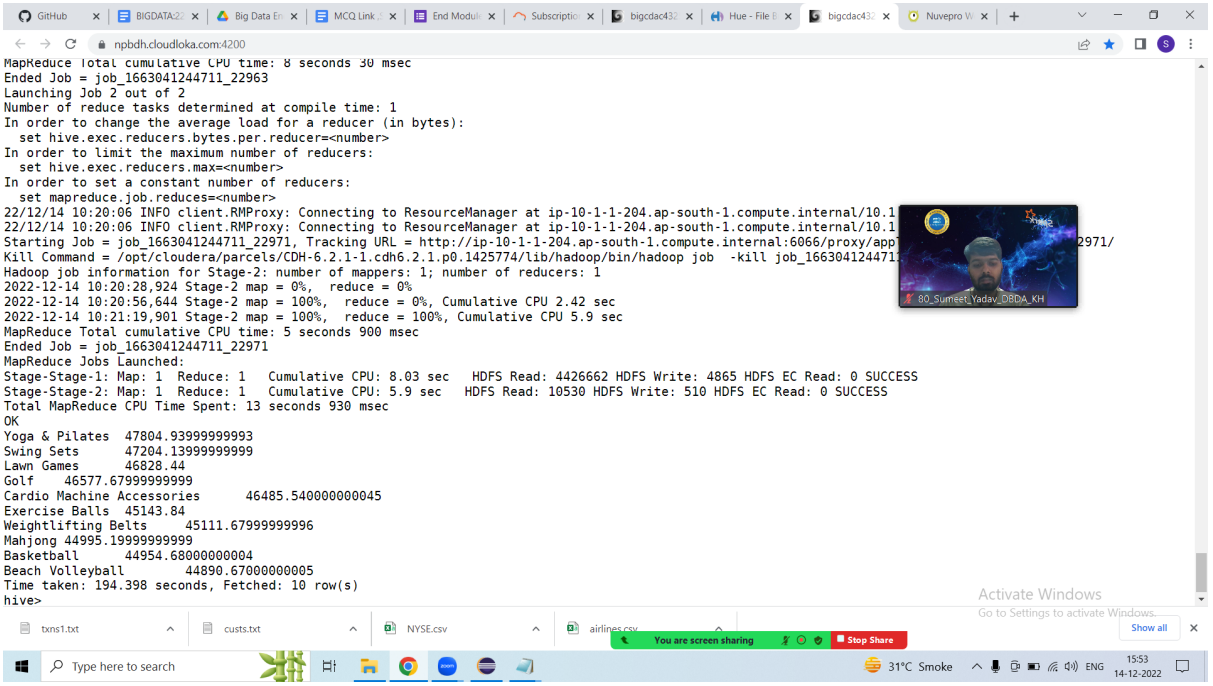

```
hadoop fs -put txns1.txt exam
```

```
create table sales(txnno INT , txndate STRING ,custno 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 sales;
```



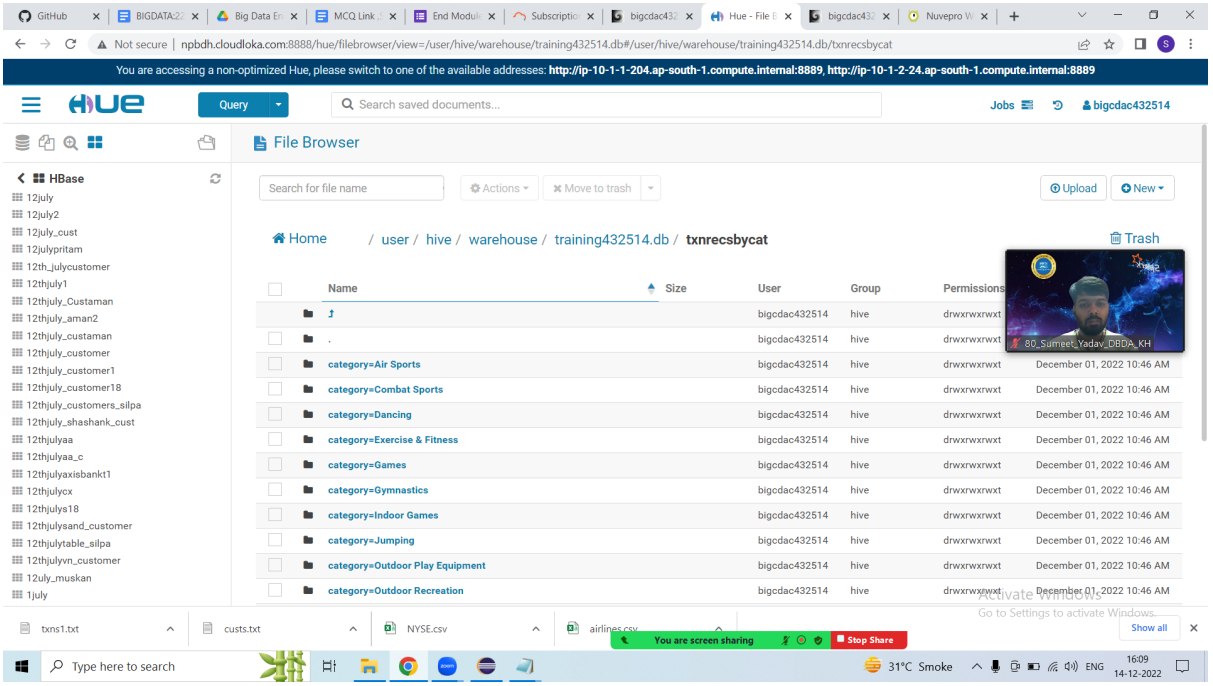
Ans

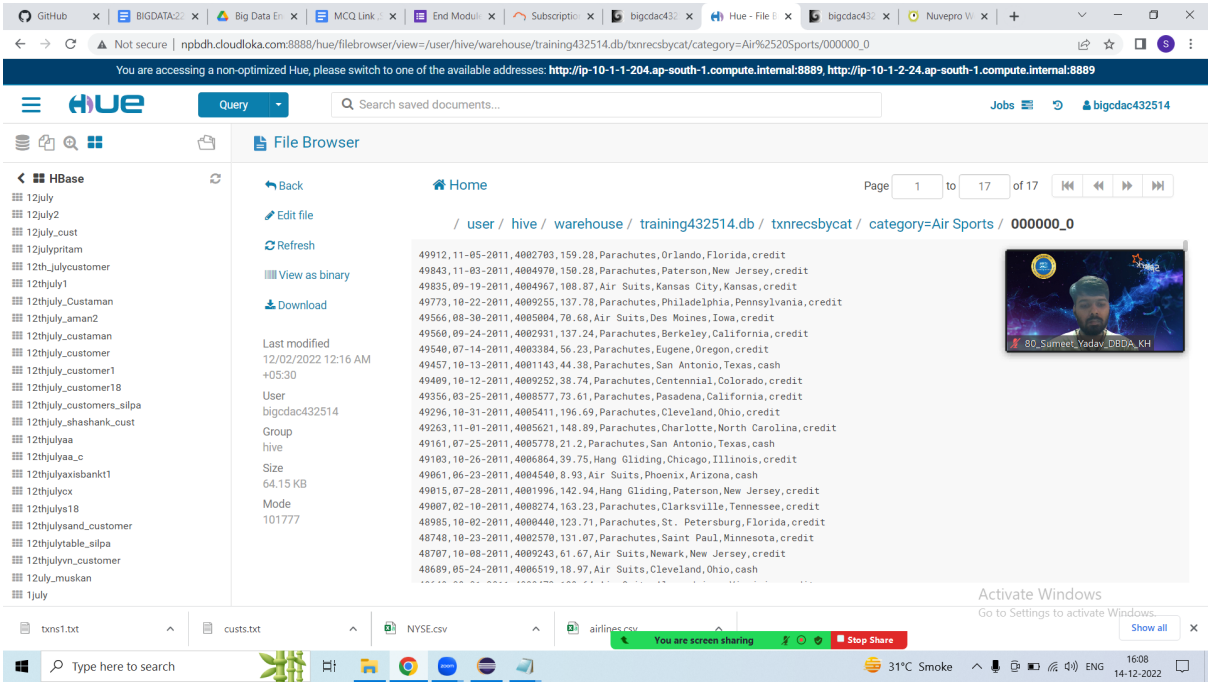
```
select product, sum(amount) as total from sales  group by product
order by total desc limit 10;
```



3) Write a program to create partiioned table on category

```
set hive.exec.dynamic.partation=true;

set hive.exec.dynamic.partation.mode=nonstrict;

create table salesbyCat(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;

INSERT OVERWRITE TABLE txnbybyCat PARTITION(category) select
txn.txnno,txn.txndate,txn.custno,
txn.amount,txn.product,txn.city,txn.state, txn.spendby, txn.category
from txnrecords txn 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

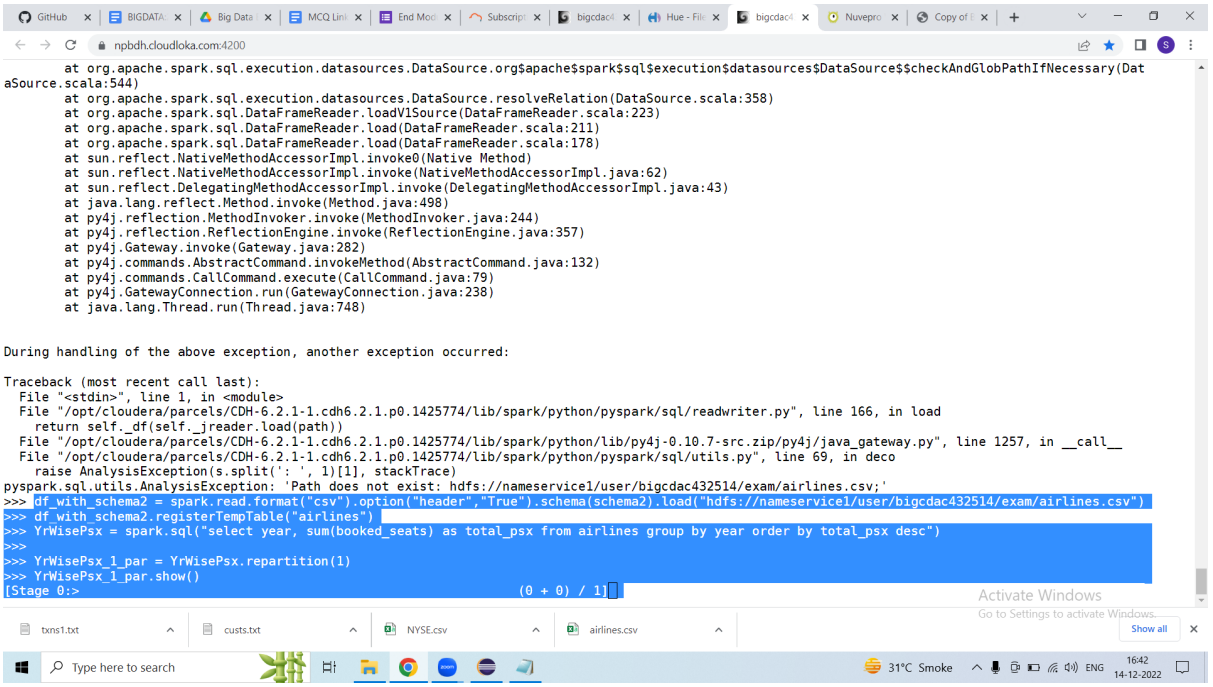
Log in to pyspark using cmd
Pyspark

Create schema
```
schema2 =
StructType().add("Year",StringType(),True).add("Quarter",StringType(
),True).a
dd("ARPS",DoubleType(),True).add("Booked_seats",IntegerType(),True)
```

Load data into the schema

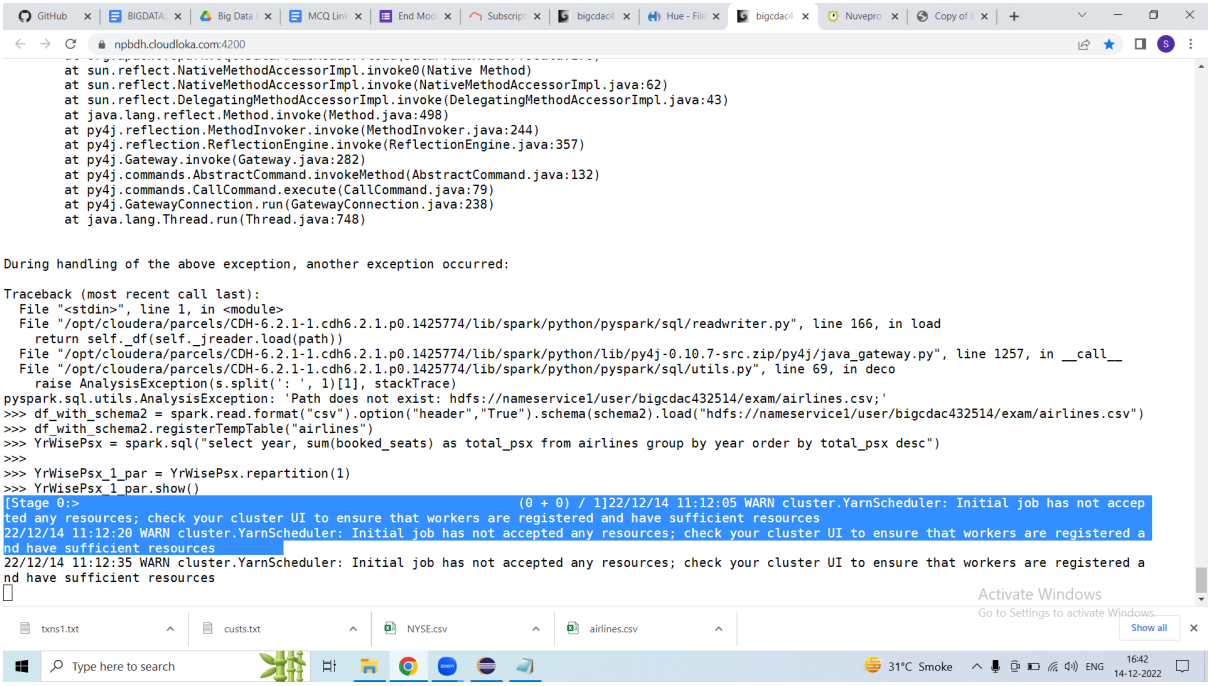
```
df_with_schema2 = spark.read.format("csv").option("header",
"True").schema(schema2).load("hdfs://nameservice1/user/bigcdac432514
/exam/airlines
.csv")
```

Register the table
```
df_with_schema2.registerTempTable("airlines")
```



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

```
YrWisePsx = spark.sql("select year, sum(booked_seats) as total_psx
from airlines group by year order by total_psx desc")
YrWisePsx_1_par = YrWisePsx.repartition(1)
```



YrWisePsx.show()

2) Identifying the highest revenue generation for which year

```
YrWiseRev = spark.sql("select year,
round(sum(arps*booked_seats)/1000000,2) as total_in_mill from
airlines group by year order by total_in_mill desc limit 1")
YrWiseRev_1_par = YrWiseRev.repartition(1)

YrWiseRev.show()
```

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