

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.

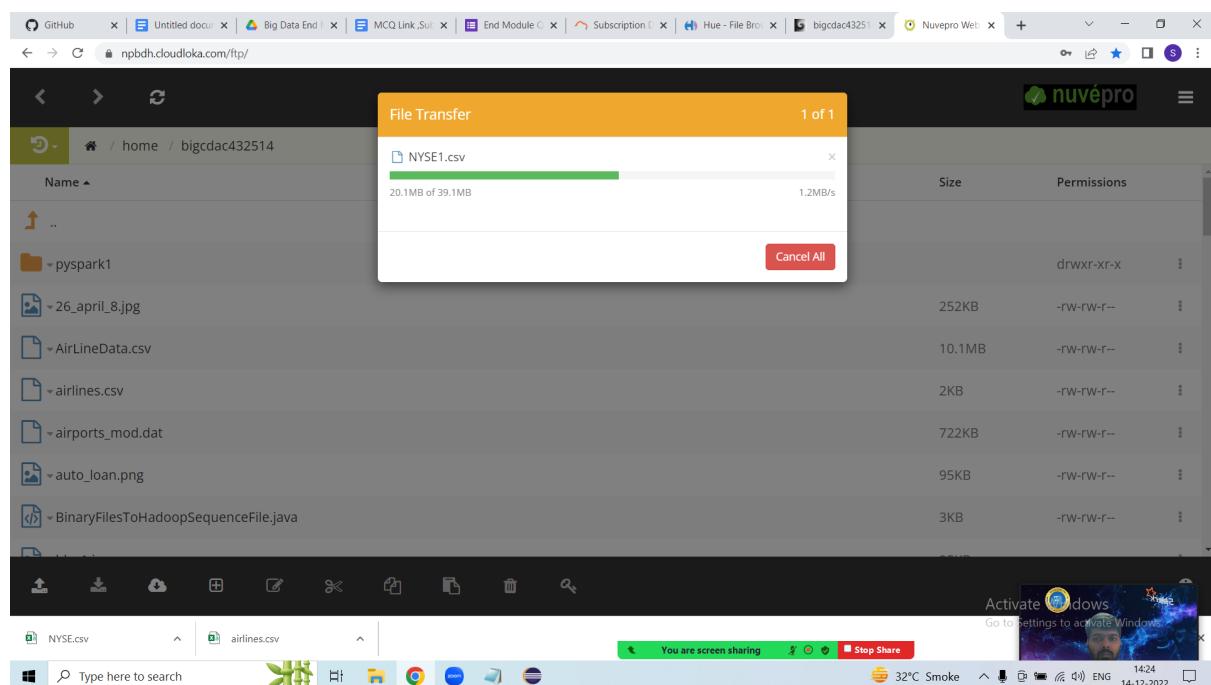
Data Structure

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));

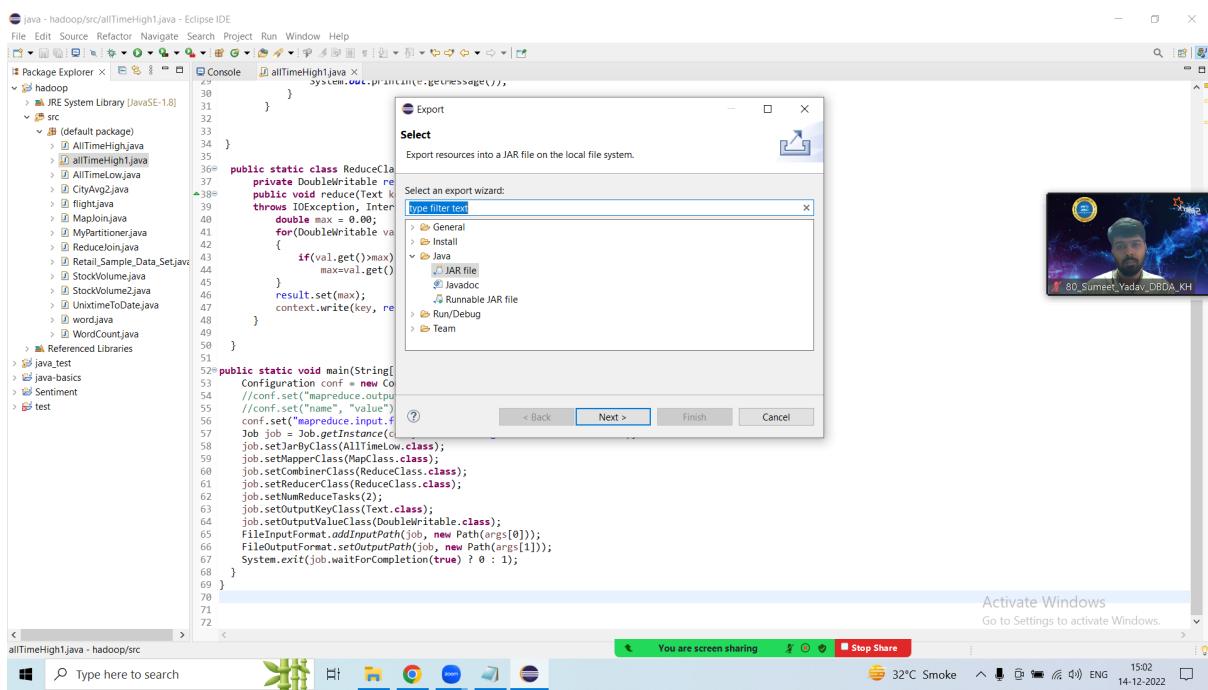
        }
    }
}

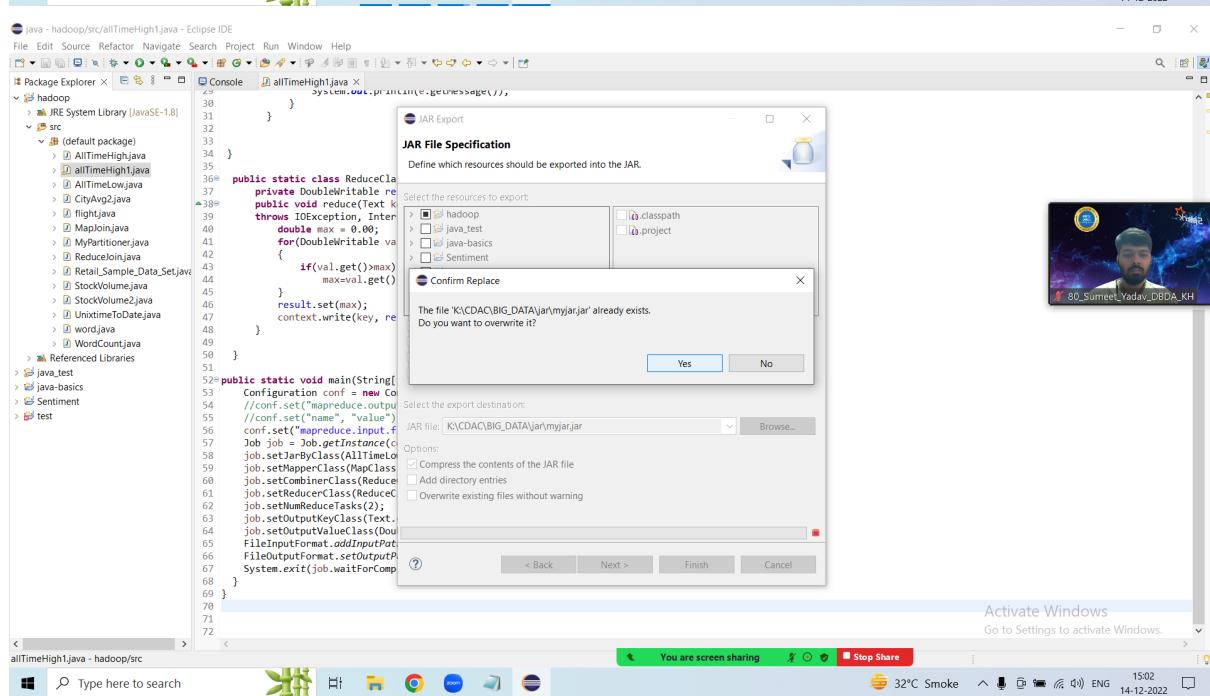
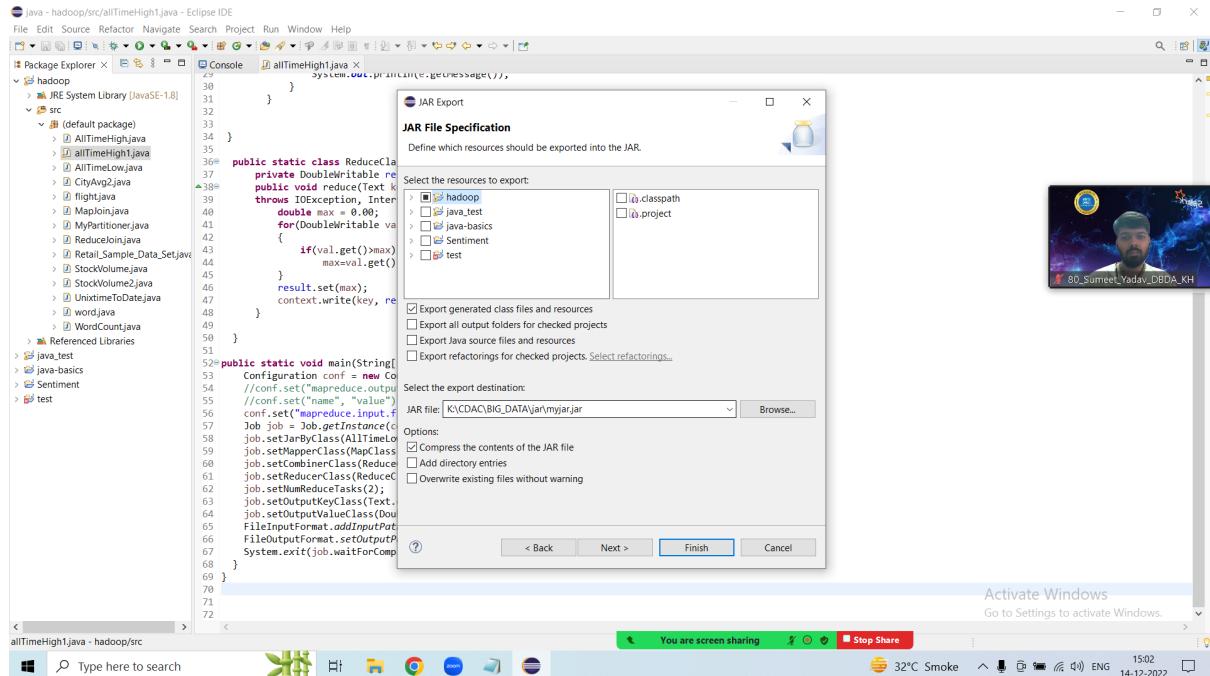
```

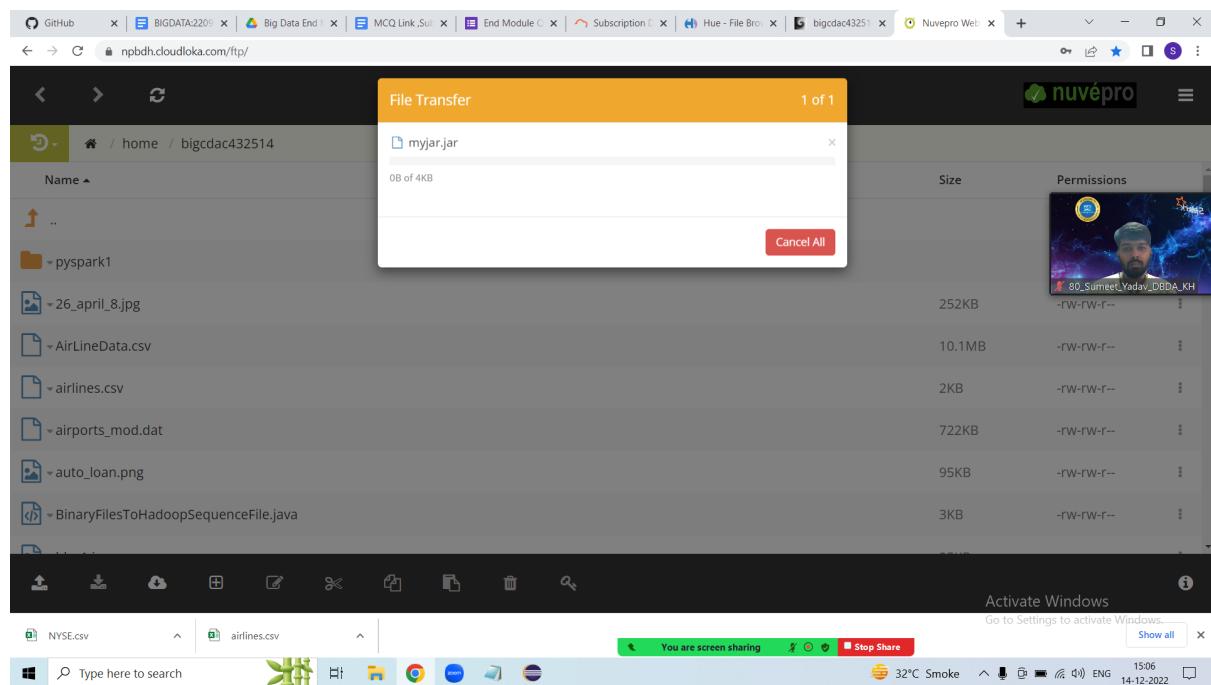
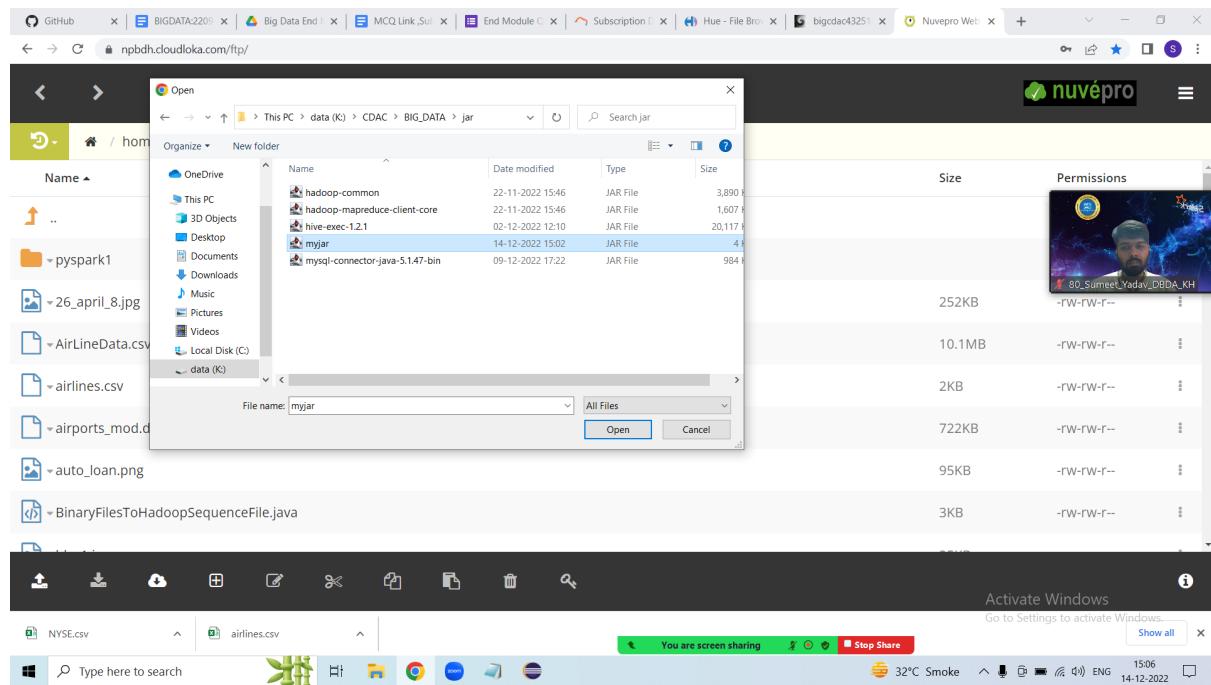
```

    }
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        //conf.set("mapreduce.output.textoutputformat.separator",","); to put ,
        instead 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);
    }
}

```







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

[bigcdac432514@ip-10-1-1-204 ~]\$ hadoop jar myjar.jar allTimeHigh1 training1/NYSE1.csv training1/opl

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 application 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.JobSubmissionEvent: number of splits:2

22/12/14 09:42:07 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.resourcemanager.metrics.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_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:

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=4105644
- HDFS: Number of bytes written=1998
- HDFS: Number of read operations=16
- HDFS: Number of large read operations=0

Activate Windows
Go to Settings to activate Windows...
Show all

NYSE.csv airlines.csv Type here to search 32°C Smoke 15:13 14-12-2022

[bigcdac432514@ip-10-1-1-204 ~]\$

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

Activate Windows
Go to Settings to activate Windows...
Show all

NYSE.csv airlines.csv Type here to search 32°C Smoke 15:13 14-12-2022

Output:

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>

File Browser

HBase

- 2July
- 2July2
- 2July_cust
- 2Julyritam
- 2Julycustomer
- 2July1
- 2July_Custaman
- 2July_aman2
- 2July_custaman
- 2July_customer
- 2July_customer1
- 2July_customer18
- 2July_customers_silpa
- 2July_shashank_cust
- 2Julyaa
- 2Julyaa_c
- 2Julyaxisbank1
- 2Julyyx
- 2July18
- 2Julysand_customer
- 2Julytable_silpa
- 2Julyvn_customer
- 2uly_muskan
- July

Home / user / bigcdac432514 / training1 / op1

Name	Size	User	Group	Permissions
j		bigcdac432514	bigcdac432514	drwxr-xr-x
.		bigcdac432514	bigcdac432514	drwxr-xr-x
_SUCCESS	0 bytes	bigcdac432514	bigcdac432514	-rw-r--r--
part-r-00000	941 bytes	bigcdac432514	bigcdac432514	-rw-r--r--
part-r-00001	1.0 KB	bigcdac432514	bigcdac432514	-rw-r--r--

Show 30 of 3 items

Page 1 of 1

Activate Windows
Go to Settings to activate Windows...
Show all

File Browser

Home

Edit file

Refresh

Download

Last modified

User

Group

Size

Mode

AAI	57.88	AB	94.94	ABB	33.39	ABD	28.58	ABR	34.45
ABT	93.37	ABV	107.5	ABVT	108.0	ABX	54.74	AC	104.0
ACC	37.0	ACG	12.63	ACI	112.89	ACM	38.25	ACO	42.7
ACE	104.0	ACG	12.63	ACI	112.89	ACM	38.25	ACS	109.55
ACI	112.89	ACM	38.25	ADP	84.31	ADX	48.56	AEC	17.6
ACM	38.25	AE	23.94	AEE	56.77	AEG	148.32	AEG	148.32

Page 1 to 1 of 1

Activate Windows
Go to Settings to activate Windows...
Show all

GitHub | BIGDATA-220 | Big Data End | MCQ Link,Soln | End Module | Subscription | Hue - File Browser | bigdac43251 | Nuvepro Web | +

Not secure | npbdh.cloudloka.com:8888/hue/filebrowser/view=/user/bigdac432514/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>

HUE Query Search saved documents... Jobs bigdac432514

File Browser

HBase

Back Home Edit file Refresh View as binary Download

Last modified 12/14/2022 3:12 PM +05:30 User bigdac432514 Group bigdac432514 Size 1.03 KB Mode 100644

AA	94.62
AAN	35.21
AAP	83.65
AAR	25.25
AAV	24.78
ABA	27.94
ABC	84.35
ABG	38.06
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	88.79
ADY	44.0
AEB	26.5
AED	26.12

Page 1 to 1 of 1

Activate Windows Go to Settings to activate Windows... Show all

NYSE.csv airlines.csv You are screen sharing Stop Share 32°C Smoke 15:14 14-12-2022

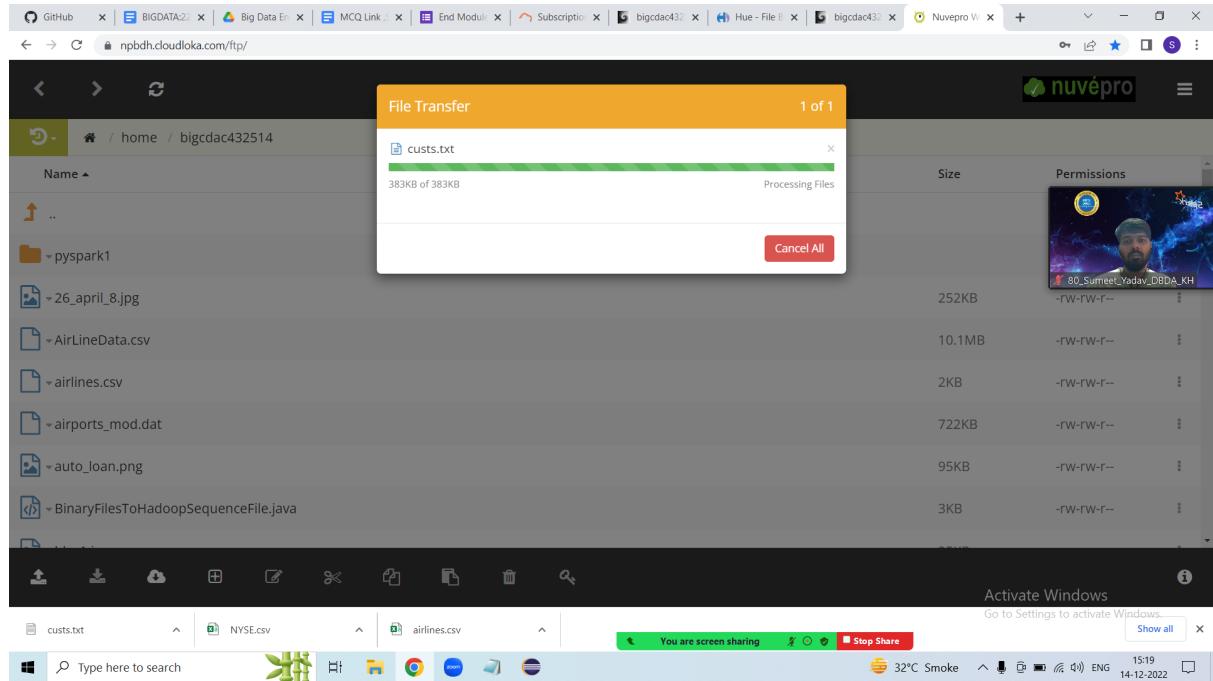
Type here to search

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;
```



The screenshot shows a Windows desktop environment. At the top, there is a taskbar with several open windows: GitHub, BIGDATA2, Big Data Env, MCQ Link, End Module, Subscription, bigdac432, Hue - File B, bigdac432, Nuvapro W, and a blank browser tab. The main area of the screen displays a terminal session for Apache Hive. The session starts with:

```
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table customer already exists)
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> describe customers;
OK
custno          int
firstname        string
lastname         string
age              int
profession       string
Time taken: 0.197 seconds, Fetched: 5 row(s)
hive> select profession , count(custno) from customers group by profession;
Query ID = bigcdac432514_20221214100245_ac2120e2-ea27-4665-ac9f-8dal6ebbb89a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
22/12/14 10:02:46 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 10:02:47 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
```

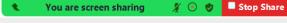
The terminal window has a decorative background image of a person's face. The bottom right corner of the screen shows the Windows system tray with icons for battery, signal, and network status, along with the date and time (14-12-2022).

```

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.63 sec HDFS Read: 400624 HDFS Write: 1584 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 630 msec
Ok
Accountant      199
Actor          202
Agricultural and food scientist 195
Architect        203
Artist          175
Athlete          196
Automotive mechanic    193
Carpenter        181
Chemist          209
Childcare worker   207
Civil engineer    193
Coach            201
Computer hardware engineer 204
Computer software engineer 216
Computer support specialist 222
Dancer           185
Designer         205
Doctor           197
Economist        189
Electrical engineer 192
Electrician       194
Engineering technician 204
Environmental scientist 176
Farmer           201
Financial analyst 198
Firefighter       217
Human resources assistant 212
Judge             196
Lawyer            212
Librarian         218
Loan officer      221
Musician          205
Nurse             192
Pharmacist        213

```

Activate Windows
Go to Settings to activate Windows...
[Show all](#)

Type here to search          

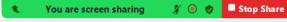
31°C Smoke 15:34 14-12-2022

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

File Transfer 1 of 1

File	Size	Permissions
txns1.txt	113KB	-rW-rW-r--
credit_card_loan.bmp	110B	-rW-rW-r--
cust2.txt	30B	-rW-rW-r--
custs.txt	14.0MB	-rW-rW-r--
custs1.txt	12.8MB	-rW-rW-r--
custs_add	14.4MB	-rW-rW-r--
D01	11.5MB	-rW-rW-r--
D02		
D11		
D12		

Activate Windows
Go to Settings to activate Windows...
[Show all](#)

Type here to search          

31°C Smoke 15:39 14-12-2022

```
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;

```

Nurse 192
Pharmacist 213
Photographer 222
Physicist 201
Pilot 211
Police officer 210
Politician 228
Psychologist 194
Real estate agent 191
Recreation and fitness worker 210
Reporter 209
Secretary 200
Social Worker 1
Social worker 212
Statistician 196
Teacher 204
Therapist 187
Veterinarian 208
Writer 101
Time taken: 109.901 seconds, Fetched: 51 row(s)

hive> describe customers;

OK

custno	int
firstname	string
lastname	string
age	int
profession	string
	Time taken: 0.075 seconds, Fetched: 5 row(s)

hive>

OK

Time taken: 0.112 seconds

hive> LOAD DATA LOCAL INPATH 'txns1.txt' OVERWRITE INTO TABLE sales;

Loading data to table training432514.sales

OK

Time taken: 0.93 seconds

hive>

Activate Windows
Go to Settings to activate Windows...
Show all

txns1.txt custs.txt NYSE.csv airlines.csv You are screen sharing Stop Share 31°C Smoke 15:45 14-12-2022

Ans

```

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

```

```

MapReduce Total cumulative CPU time: 8 seconds 30 msec
Ended Job = job_1663041244711_22963
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
22/12/14 10:20:06 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1
22/12/14 10:20:06 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1
Starting Job = job_1663041244711_22971, Tracking URL = http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/app
Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job_1663041244711
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-12-14 10:20:28,924 Stage-2 map = 0%, reduce = 0%
2022-12-14 10:20:56,644 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 2.42 sec
2022-12-14 10:21:19,901 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 5.9 sec
MapReduce Total cumulative CPU time: 5 seconds 900 msec
Ended Job = job_1663041244711_22971
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.03 sec HDFS Read: 4426662 HDFS Write: 4865 HDFS EC Read: 0 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 5.9 sec HDFS Read: 10530 HDFS Write: 510 HDFS EC Read: 0 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 930 msec
OK
Yoga & Pilates 47804.93999999999
Swing Sets 47204.13999999999
Lawn Games 46828.44
Golf 46577.67999999999
Cardio Machine Accessories 46485.54000000045
Exercise Balls 45143.84
Weightlifting Belts 45111.67999999999
Mahjong 44995.19999999999
Basketball 44954.68000000004
Beach Volleyball 44890.67000000005
Time taken: 194.398 seconds, Fetched: 10 row(s)
hive>

```

3) Write a program to create partitioned table on category

```

set hive.exec.dynamic.partition=true;

set hive.exec.dynamic.partition.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;

```

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>

HUE Query Search saved documents...

File Browser

Search for file name: Actions Move to trash Upload New

Home / user / hive / warehouse / training432514.db / txnrecsbycat

Trash

Name	Size	User	Group	Permissions
category=Air Sports		bigcdac432514	hive	drwxrwxrwx
category=Combat Sports		bigcdac432514	hive	drwxrwxrwx
category=Dancing		bigcdac432514	hive	drwxrwxrwx
category=Exercise & Fitness		bigcdac432514	hive	drwxrwxrwx
category=Games		bigcdac432514	hive	drwxrwxrwx
category=Gymnastics		bigcdac432514	hive	drwxrwxrwx
category=Indoor Games		bigcdac432514	hive	drwxrwxrwx
category=Jumping		bigcdac432514	hive	drwxrwxrwx
category=Outdoor Play Equipment		bigcdac432514	hive	drwxrwxrwx
category=Outdoor Recreation		bigcdac432514	hive	drwxrwxrwx

Activate Windows Go to Settings to activate Windows... Show all

txns1.txt custs.txt NYSE.csv airlines.csv

Type here to search 31°C Smoke 16:09 14-12-2022

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>

HUE Query Search saved documents...

File Browser

Back Edit file Refresh View as binary Download

Last modified 12/02/2022 12:16 AM +05:30

User bigcdac432514

Group hive

Size 64.15 KB

Mode 101777

Home / user / hive / warehouse / training432514.db / txnrecsbycat / category=Air Sports / 000000_0

49912, 11-05-2011, 4002703, 159.28, Parachutes, Orlando, Florida, credit
49843, 11-03-2011, 4004970, 150.28, Parachutes, Paterson, New Jersey, credit
49835, 09-19-2011, 4004967, 108.87, Air Suits, Kansas City, Kansas, credit
49773, 10-22-2011, 4009255, 137.78, Parachutes, Philadelphia, Pennsylvania, credit
49566, 09-30-2011, 4005084, 70.68, Air Suits, Des Moines, Iowa, credit
49568, 09-24-2011, 4002931, 137.24, Parachutes, Berkeley, California, credit
49540, 07-14-2011, 4003384, 56.23, Parachutes, Eugene, Oregon, credit
49457, 10-13-2011, 4001143, 44.38, Parachutes, San Antonio, Texas, cash
49409, 10-12-2011, 4009252, 38.74, Parachutes, Centennial, Colorado, credit
49356, 03-25-2011, 4008577, 73.61, Parachutes, Pasadena, California, credit
49296, 10-31-2011, 4005411, 196.69, Parachutes, Cleveland, Ohio, credit
49263, 11-01-2011, 4005621, 148.89, Parachutes, Charlotte, North Carolina, credit
49161, 07-25-2011, 4005778, 21.2, Parachutes, San Antonio, Texas, cash
49103, 10-26-2011, 4006864, 39.75, Hang Gliding, Chicago, Illinois, credit
49061, 06-23-2011, 4004549, 8.93, Air Suits, Phoenix, Arizona, cash
49015, 07-28-2011, 4001996, 142.94, Hang Gliding, Paterson, New Jersey, credit
49087, 02-19-2011, 4008274, 163.23, Parachutes, Clarksville, Tennessee, credit
48985, 10-02-2011, 4008446, 123.71, Parachutes, St. Petersburg, Florida, credit
48748, 10-23-2011, 4002570, 131.07, Parachutes, Saint Paul, Minnesota, credit
48707, 10-08-2011, 4009243, 61.67, Air Suits, Newark, New Jersey, credit
48689, 05-24-2011, 4006519, 18.97, Air Suits, Cleveland, Ohio, cash

Activate Windows Go to Settings to activate Windows... Show all

txns1.txt custs.txt NYSE.csv airlines.csv

Type here to search 31°C Smoke 16:08 14-12-2022

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
```

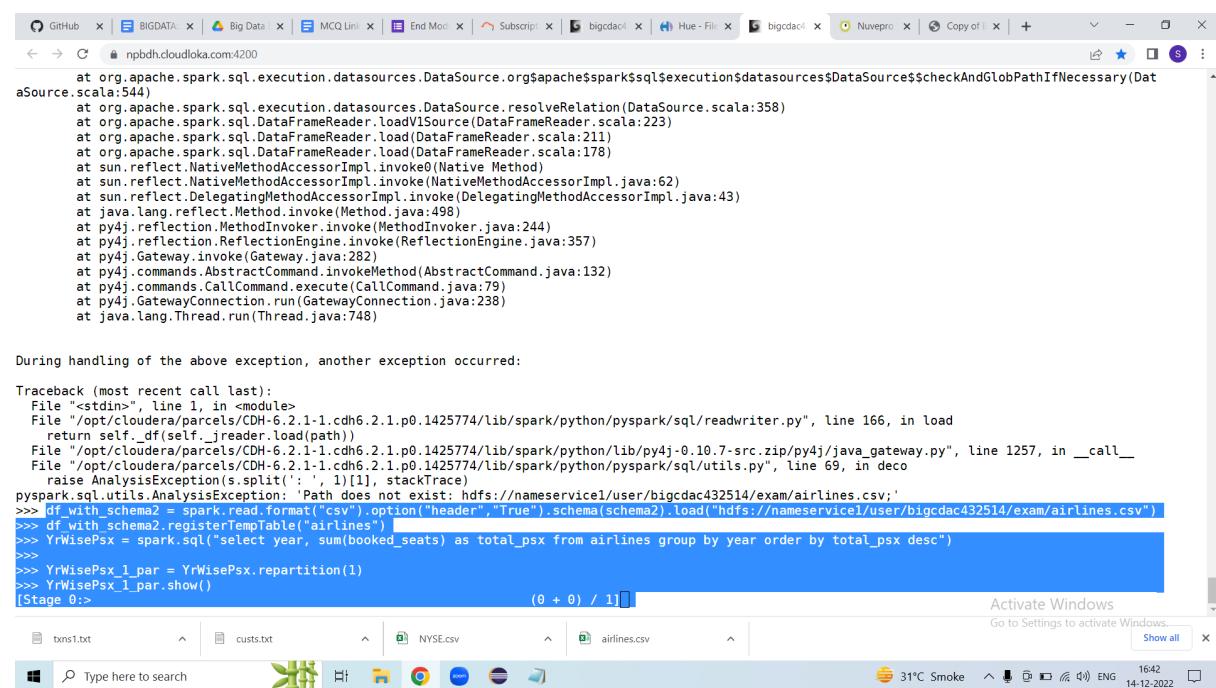
```
Create schema  
schema2 =  
StructType().add("Year", StringType(), True).add("Quarter", StringType(), True).add("ARPS", DoubleType(), True).add("Booked_seats", IntegerType(), True)
```

```
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")
```



```
at org.apache.spark.sql.execution.datasources.DataSource.org$apache$spark$sql$execution$dataSources$DataSource$$checkAndGlobPathIfNecessary(DataSource.scala:544)
at org.apache.spark.sql.execution.datasources.DataSource.resolveRelation(DataSource.scala:358)
at org.apache.spark.sql.DataFrameReader.loadV1Source(DataFrameReader.scala:223)
at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:211)
at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:178)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
at py4j.Gateway.invoke(Gateway.java:282)
at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:132)
at py4j.commands.CallCommand.execute(CallCommand.java:79)
at py4j.GatewayConnection.run(GatewayConnection.java:238)
at java.lang.Thread.run(Thread.java:748)

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
    File "/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/pyspark/sql/readwriter.py", line 166, in load
      return self._df(self._jreader.load(path))
    File "/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1257, in __call__
      File "/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/pyspark/sql/utils.py", line 69, in deco
        raise AnalysisException(s.split(': ', 1)[1], stackTrace)
pyspark.sql.utils.AnalysisException: Path does not exist: hdfs://nameservice1/user/bigcdac432514/exam/airlines.csv'
>>> df_with_schema2 = spark.read.format("csv").option("header","True").schema(schema2).load("hdfs://nameservice1/user/bigcdac432514/exam/airlines.csv")
>>> df_with_schema2.registerTempTable("airlines")
>>> 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_1_par.show()
[Stage 0:>
```

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)
```

During handling of the above exception, another exception occurred:

```

Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
    File "/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/pyspark/sql/readwriter.py", line 166, in load
    return self._df(self._jreader.load(path))
  File "/opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1257, in __call__
    raise AnalysisException(s.split('; ', 1)[1], stackTrace)
pyspark.sql.utils.AnalysisException: Path does not exist: hdfs://nameservice1/user/bigcdac432514/exam/airlines.csv'
>>> df_with_schema2 = spark.read.format("csv").option("header","True").schema(schema2).load("hdfs://nameservice1/user/bigcdac432514/exam/airlines.csv")
>>> df_with_schema2.registerTempTable("airlines")
>>> 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_1_par.show()
[Stage 0:                                     (0 + 0) / 1]22/12/14 11:12:05 WARN cluster.YarnScheduler: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
22/12/14 11:12:20 WARN cluster.YarnScheduler: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources
22/12/14 11:12:35 WARN cluster.YarnScheduler: Initial job has not accepted any resources; check your cluster UI to ensure that workers are registered and have sufficient resources

```

Activate Windows
Go to Settings to activate Windows...
Show all

txns1.txt custs.txt NYSE.csv airlines.csv

Type here to search 31°C Smoke 16:42 14-12-2022

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)

