**BDT EXAM**

**========================================================================**

**Q1. MAP REDUCE**

Mapper class

public class EXAM1 {

public static class mapper extends Mapper<LongWritable, Text, Text, DoubleWritable>{

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

{

try {

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

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

context.write (new Text(str[1]), new DoubleWritable(i));

}

catch(Exception e)

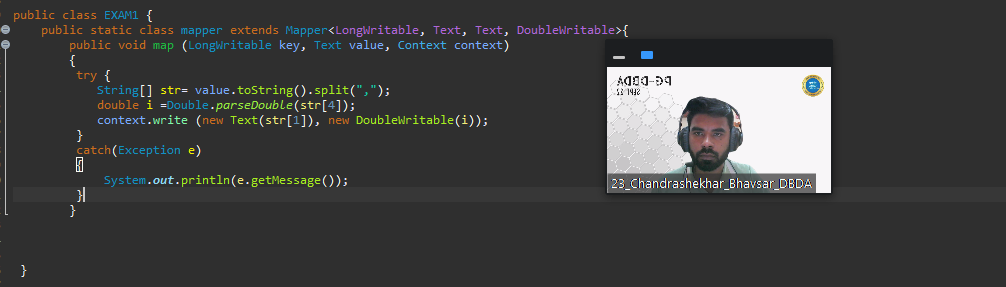
{

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

}

}

**}**

****

**REDUCER CLASS**

public static class reducer 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();

}

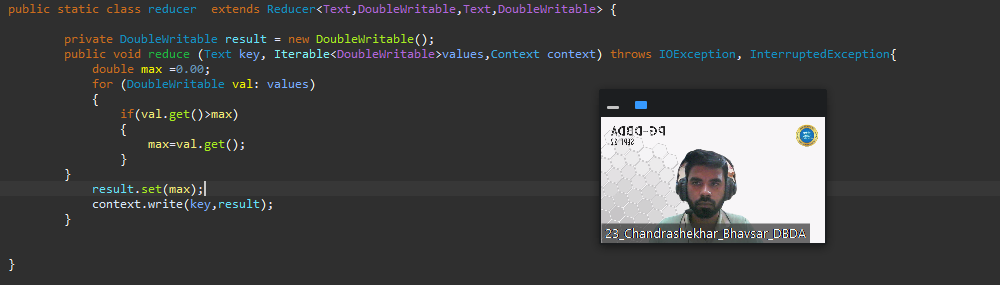
}

result.set(max);

context.write(key,result);

}

**}**

****

**MAIN CLASS**

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

Configuration conf = new Configuration();

Job job = Job.getInstance(conf, " xyz");

job.setJarByClass(EXAM1.class);

job.setMapperClass(mapper.class);

job.setReducerClass(reducer.class);

job.setNumReduceTasks(1);

job.setMapOutputKeyClass(Text.class);

job.setMapOutputValueClass(DoubleWritable.class);

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);

}

}

}

****

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

**—----------------------------------------------------------------------------------------------------------------------------**

**Q2. HIVE CUSTOMER TABLE**

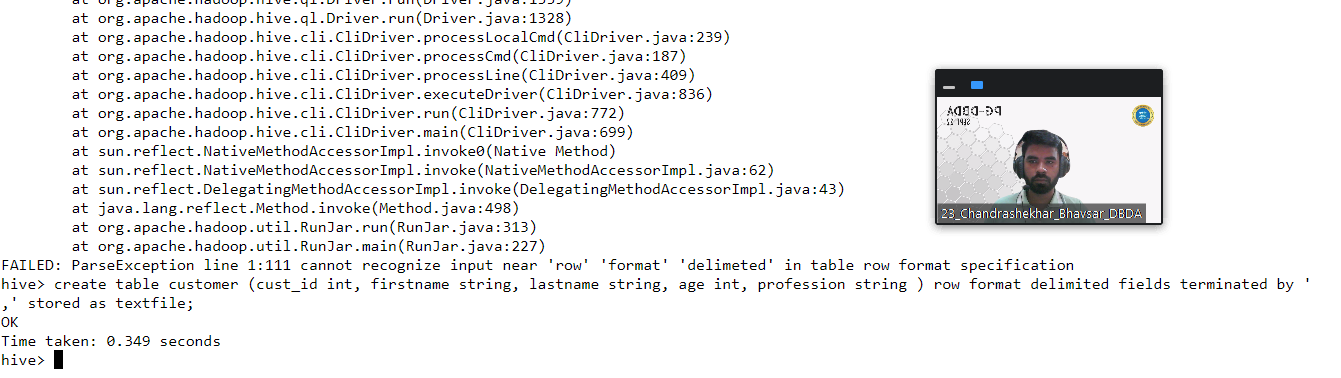
**hive> create table customer (cust\_id int, firstname string, lastname string, age int, profession string ) row format delimited fields terminated by '**

**,' stored as textfile;**

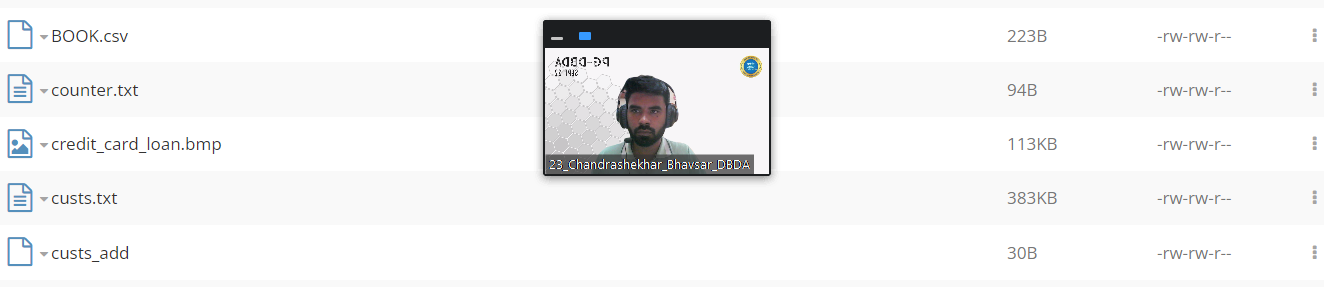
**OK**

**Time taken: 0.349 seconds**

**hive>**

****

**FILE UPLOADED**

****

**MAPPED WITH TABLE**

**hive> load data local inpath "custs.txt" overwrite into table customer;**

**Loading data to table exam\_432571.customer**

**OK**

**Time taken: 1.401 seconds**

**hive>**

**> select \* from customer limit 10;**

**OK**

**4000001 Kristina Chung 55 Pilot**

**4000002 Paige Chen 74 Teacher**

**4000003 Sherri Melton 34 Firefighter**

**4000004 Gretchen Hill 66 Computer hardware engineer**

**4000005 Karen Puckett 74 Lawyer**

**4000006 Patrick Song 42 Veterinarian**

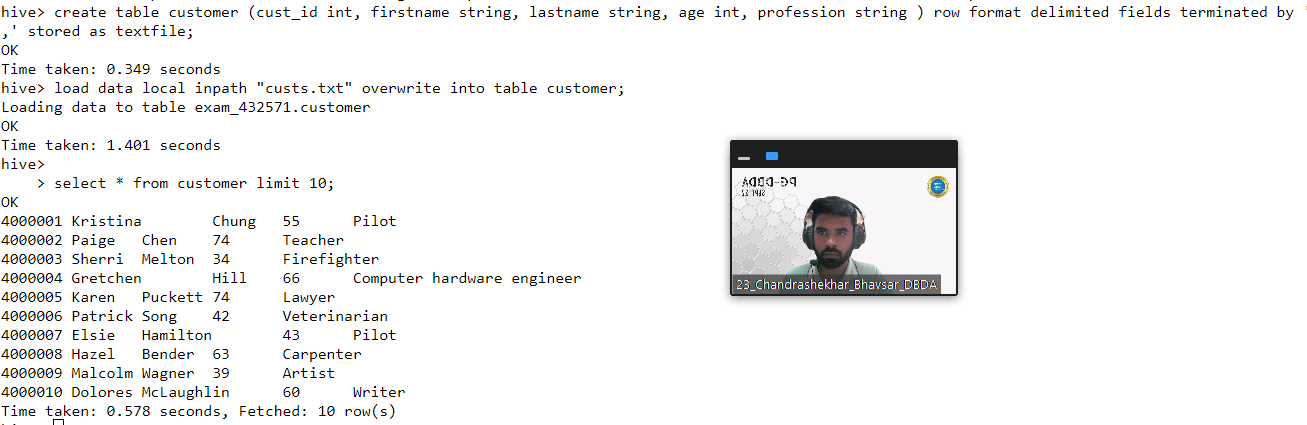
**4000007 Elsie Hamilton 43 Pilot**

**4000008 Hazel Bender 63 Carpenter**

**4000009 Malcolm Wagner 39 Artist**

**4000010 Dolores McLaughlin 60 Writer**

**Time taken: 0.578 seconds, Fetched: 10 row(s)**

****

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

hive> select profession, count(profession) from customer group by profession;

Query ID = bigcdac432571\_20221214090838\_bbb8d6cf-6bbb-40e2-a91e-654b834cb43f

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 09:08:39 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 09:08:39 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032

Starting Job = job\_1663041244711\_22595, Tracking URL = <http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_22595/>

Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job\_1663041244711\_22595

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-12-14 09:08:56,582 Stage-1 map = 0%, reduce = 0%

2022-12-14 09:09:11,712 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.51 sec

2022-12-14 09:09:28,620 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.95 sec

MapReduce Total cumulative CPU time: 6 seconds 950 msec

Ended Job = job\_1663041244711\_22595

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.95 sec HDFS Read: 400602 HDFS Write: 1584 HDFS EC Read: 0 SUCCESS

Total MapReduce CPU Time Spent: 6 seconds 950 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

Photographer 222

Physicist 201

Pilot 211

Police officer 210

Politician 228

Psychologist 194

Real estate agent 191

Recreation and fitness worker 210

Reporter 200

Secretary 200

Social Worker 1

Social worker 212

Statistician 196

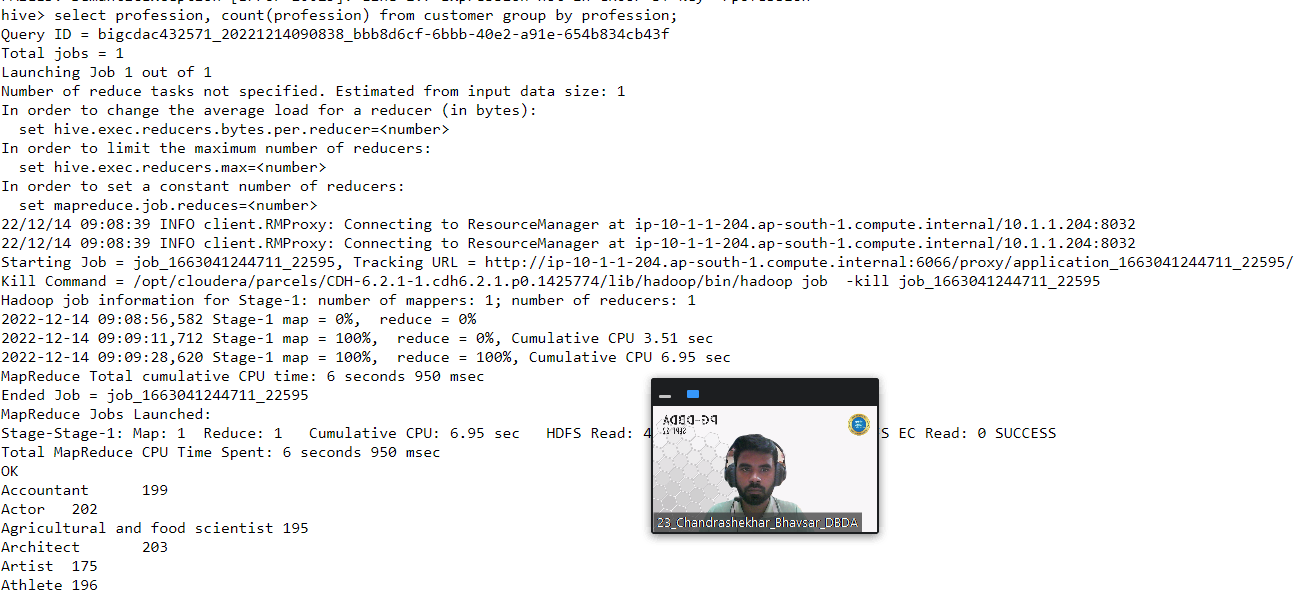
Teacher 204

Therapist 187

Veterinarian 208

Writer 101

Time taken: 53.425 seconds, Fetched: 51 row(s)



**Q2. TRANSACTTION TABLE:**

**hive> create table txn (txn\_id int, txn\_date string, cust\_id int, ammount double, cat string, product string, city string, state string, spendby stri**

**ng) row format delimited fields terminated by ',' stored as textfile;**

**OK**

**Time taken: 0.097 seconds**

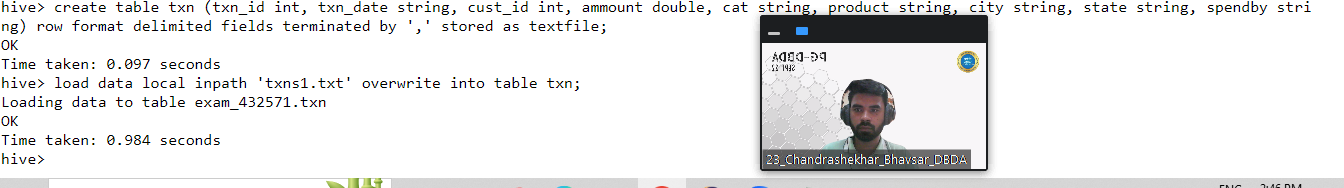
**hive> load data local inpath 'txns1.txt' overwrite into table txn;**

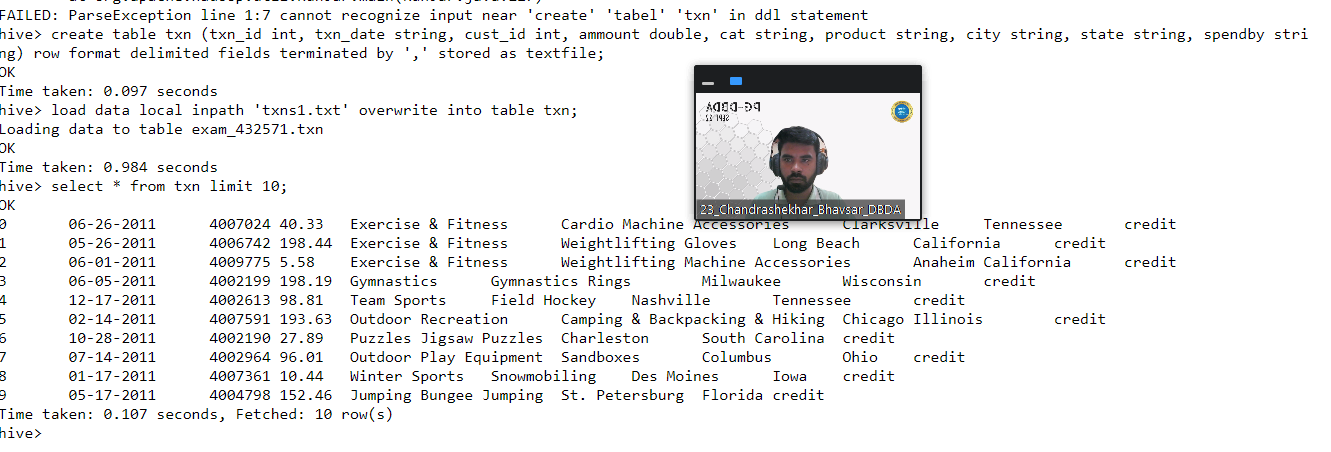
**Loading data to table exam\_432571.txn**

**OK**

**Time taken: 0.984 seconds**

**hive>**

****

****

**QUERY:**

**select product,sum(ammount) as total from txn group by product order by total desc limit 10;**

**Query ID = bigcdac432571\_20221214092153\_775ea452-65a1-4e32-9fcf-c45e09d06565**

**Total jobs = 2**

**Launching Job 1 out of 2**

**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 09:21:53 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 09:21:53 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032**

**Starting Job = job\_1663041244711\_22659, Tracking URL =** [**http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application\_1663041244711\_22659/**](http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_22659/)

**Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job\_1663041244711\_22659**

**Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1**

**2022-12-14 09:22:04,130 Stage-1 map = 0%, reduce = 0%**

**2022-12-14 09:22:24,845 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.91 sec**

**2022-12-14 09:22:37,374 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.65 sec**

**MapReduce Total cumulative CPU time: 7 seconds 650 msec**

**Ended Job = job\_1663041244711\_22659**

**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 09:22:39 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 09:22:39 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032**

**Starting Job = job\_1663041244711\_22661, Tracking URL =** [**http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application\_1663041244711\_22661/**](http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_22661/)

**Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job\_1663041244711\_22661**

**Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1**

**2022-12-14 09:22:51,799 Stage-2 map = 0%, reduce = 0%**

**2022-12-14 09:23:31,641 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 3.18 sec**

**2022-12-14 09:23:42,088 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 7.1 sec**

**MapReduce Total cumulative CPU time: 7 seconds 100 msec**

**Ended Job = job\_1663041244711\_22661**

**MapReduce Jobs Launched:**

**Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.65 sec HDFS Read: 4426710 HDFS Write: 4865 HDFS EC Read: 0 SUCCESS**

**Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 7.1 sec HDFS Read: 10532 HDFS Write: 510 HDFS EC Read: 0 SUCCESS**

**Total MapReduce CPU Time Spent: 14 seconds 750 msec**

**OK**

**Yoga & Pilates 47804.93999999993**

**Swing Sets 47204.13999999999**

**Lawn Games 46828.44**

**Golf 46577.67999999999**

**Cardio Machine Accessories 46485.540000000045**

**Exercise Balls 45143.84**

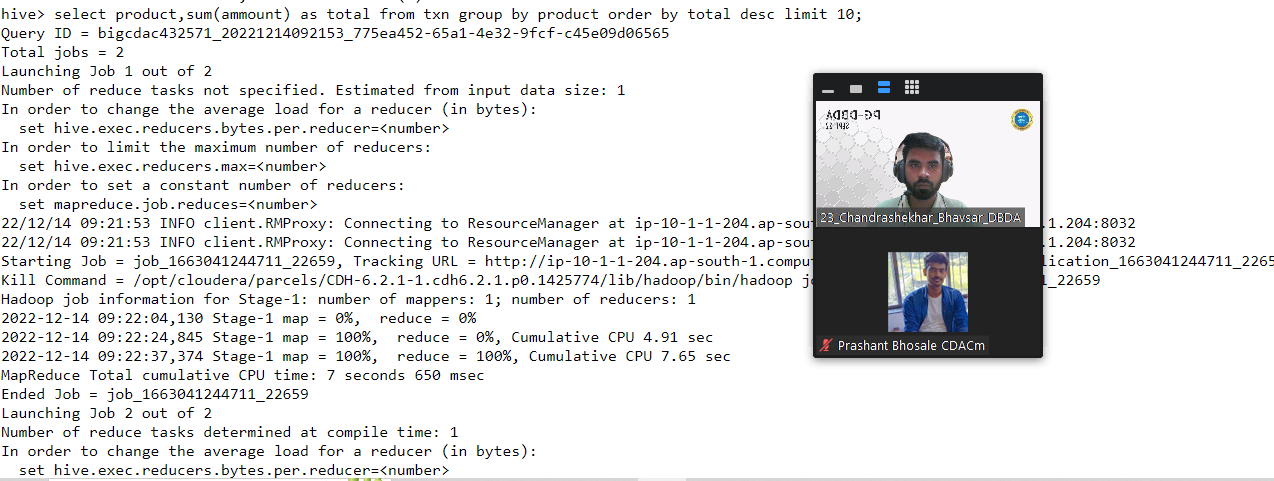
**Weightlifting Belts 45111.67999999996**

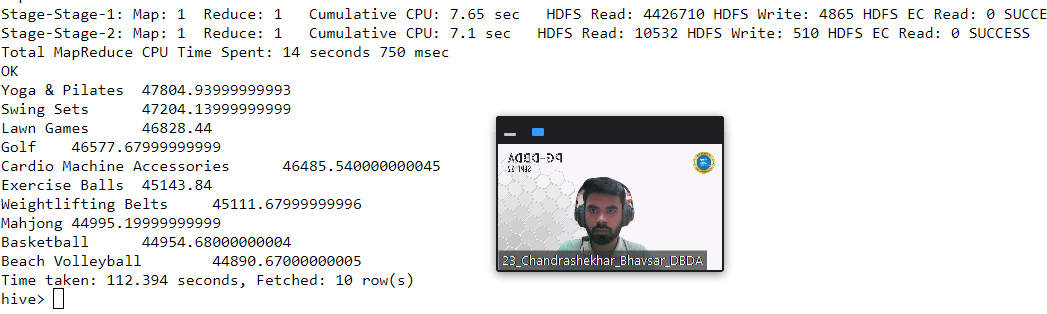
**Mahjong 44995.19999999999**

**Basketball 44954.68000000004**

**Beach Volleyball 44890.67000000005**

**Time taken: 112.394 seconds, Fetched: 10 row(s)**

****

****

**3) Write a program to create partiioned table on category QUESTION 3 [15 marks]**

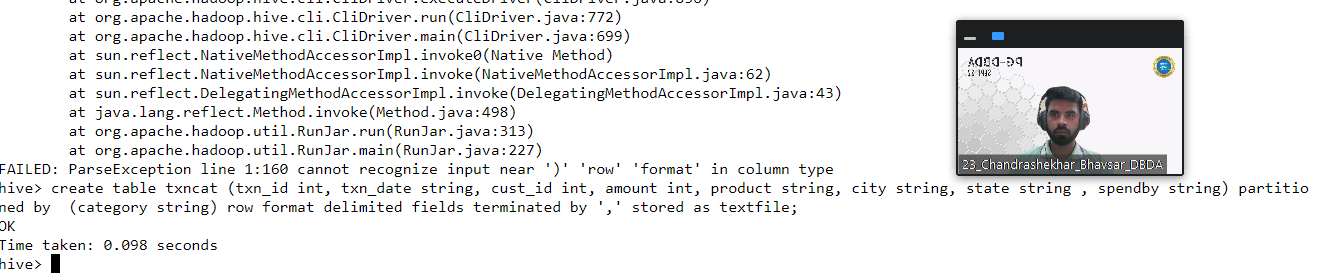
**hive> create table txncat (txn\_id int, txn\_date string, cust\_id int, amount int, product string, city string, state string , spendby string) partitio**

**ned by (category string) row format delimited fields terminated by ',' stored as textfile;**

**OK**

**Time taken: 0.098 seconds**

**hive>**

****

**hive> insert overwrite table txncat partition (category) select txn.txn\_id, txn.txn\_date, txn.cust\_id,txn.ammount,txn.product,txn.city,txn.state,txn.**

**spendby,txn.cat from txn distribute by cat;**

**Query ID = bigcdac432571\_20221214093912\_6c907758-e279-4453-909e-3ceec696d113**

**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 09:39:13 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 09:39:14 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032**

**Starting Job = job\_1663041244711\_22775, Tracking URL =** [**http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application\_1663041244711\_22775/**](http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1663041244711_22775/)

**Kill Command = /opt/cloudera/parcels/CDH-6.2.1-1.cdh6.2.1.p0.1425774/lib/hadoop/bin/hadoop job -kill job\_1663041244711\_22775**

**Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1**

**2022-12-14 09:39:32,027 Stage-1 map = 0%, reduce = 0%**

**2022-12-14 09:39:56,102 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.5 sec**

**2022-12-14 09:40:15,762 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 12.74 sec**

**MapReduce Total cumulative CPU time: 12 seconds 740 msec**

**Ended Job = job\_1663041244711\_22775**

**Loading data to table exam\_432571.txncat partition (category=null)**

**Time taken to load dynamic partitions: 0.565 seconds**

**Time taken for adding to write entity : 0.004 seconds**

**MapReduce Jobs Launched:**

**Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 12.74 sec HDFS Read: 4429116 HDFS Write: 3355527 HDFS EC Read: 0 SUCCESS**

**Total MapReduce CPU Time Spent: 12 seconds 740 msec**

**OK**

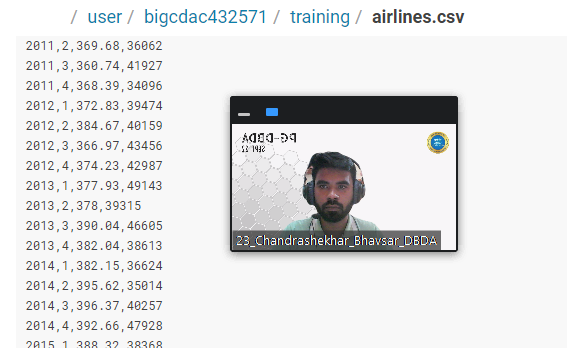
**Time taken: 68.362 seconds**

****

**—----------------------------------------------------------------------------------------------------------------------------**

**PySpark Please find the AIRLINES data set Year Quarter Average revenue per seat Total number of booked seats**

1. **Uploaded file:**

****

1. **Creating schema:**

**>>> from pyspark.sql.types import StructType, IntegerType, StringType, DoubleType, LongType**

**>>> schema1=StructType().add("year",StringType(),True).add("Quarter",StringType(),True).add("avr",DoubleType(),True).add("BookSeats",IntegerType(),Tr**

**ue)**

**>>> df\_with\_schema1=spark.read.format("csv").option("header","True").schema(schema1).load("hdfs://nameservice1/user/bigcdac432571/training/airlines.c**

**sv")**

**>>> df.with\_schema1.registerTempTable("AIRLINES\_EXAM")**

**Traceback (most recent call last):**

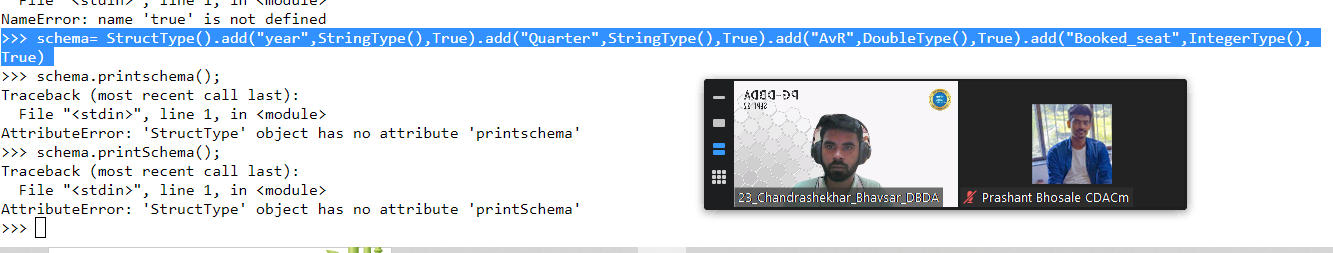
**File "<stdin>", line 1, in <module>**

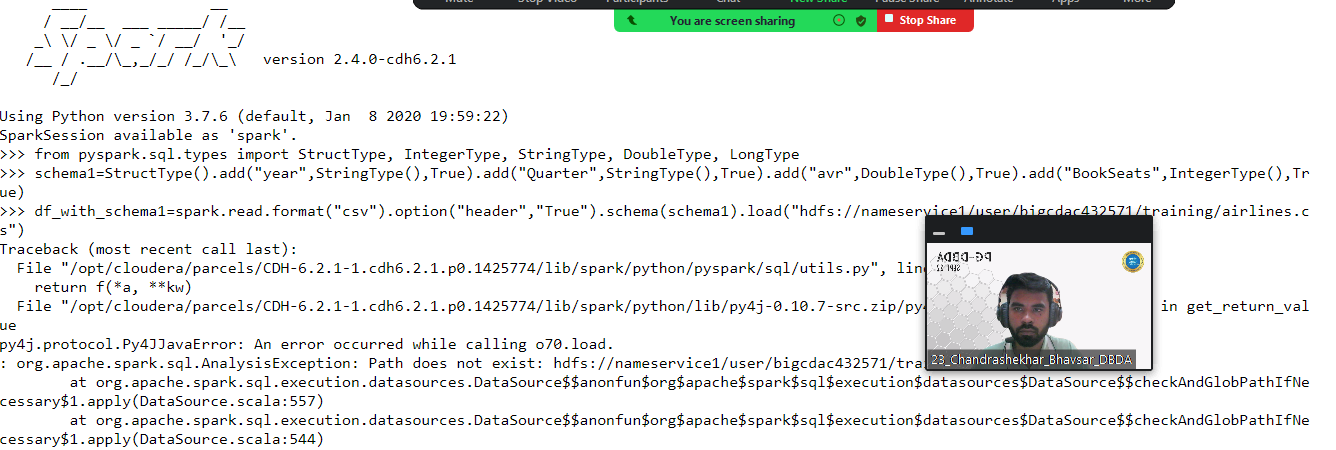
**NameError: name 'df' is not defined**

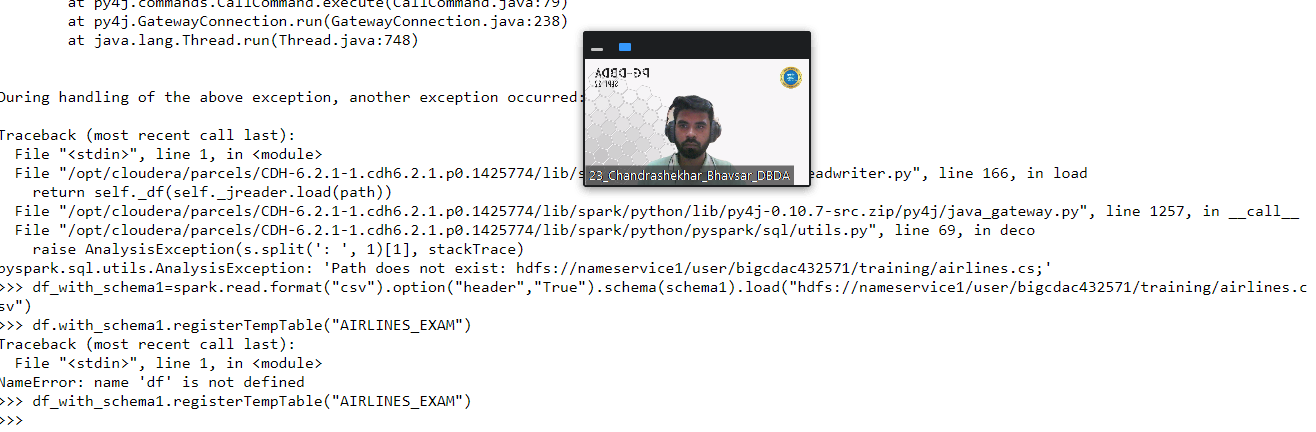
**>>> df\_with\_schema1.registerTempTable("AIRLINES\_EXAM")**

**)**

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

****

****

****

**Select year from AIRLINES\_EXAM where (select BookSeat from Airlines where BookSeats = max(BookSeats))**

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