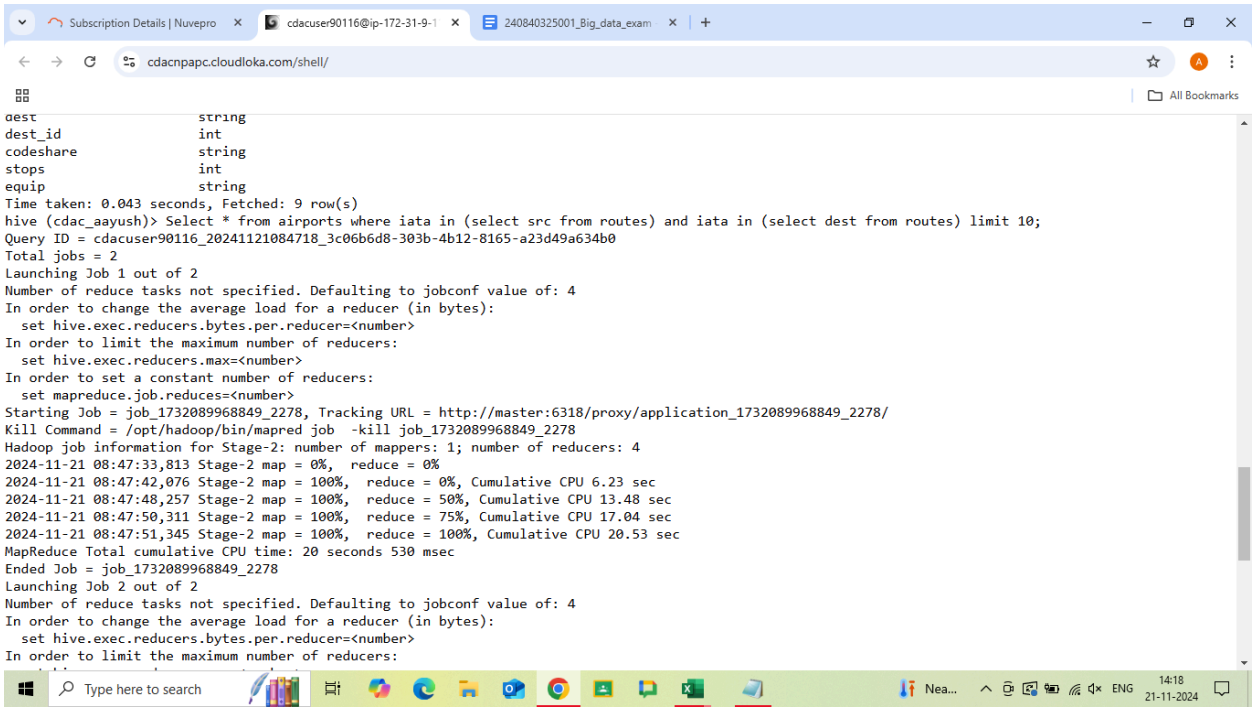
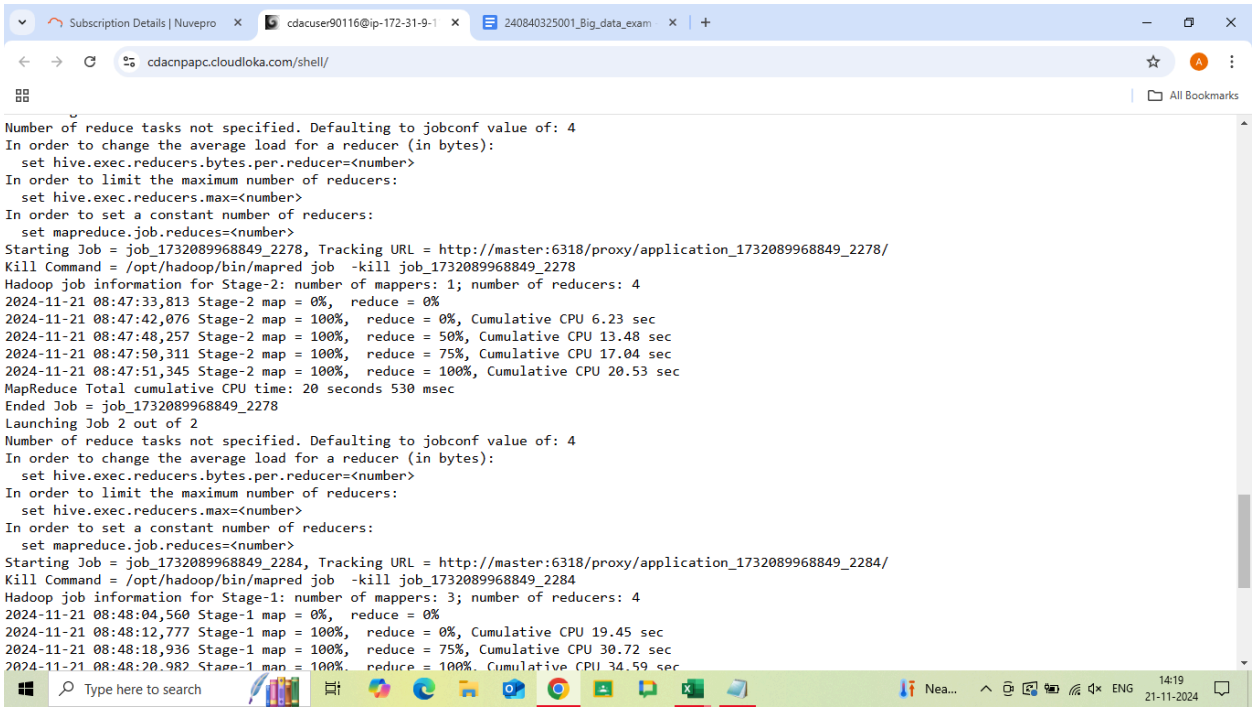
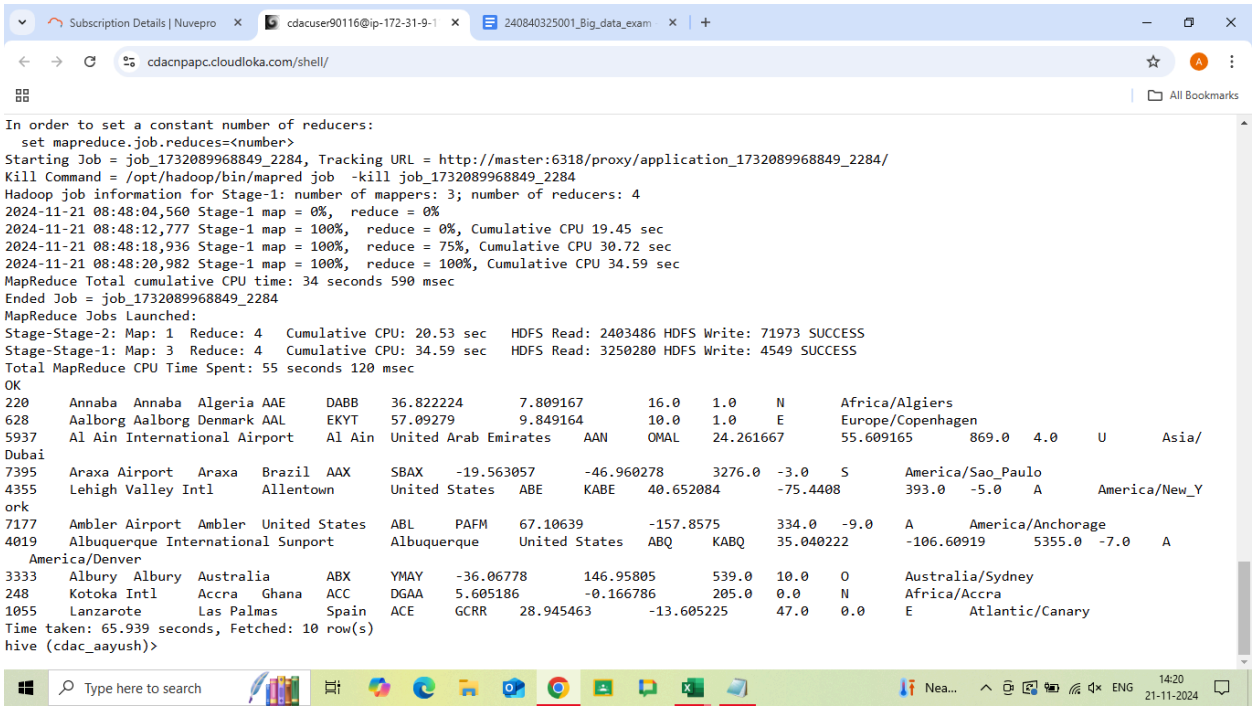
HIVE-

Q1.
Setup hive using hive command-

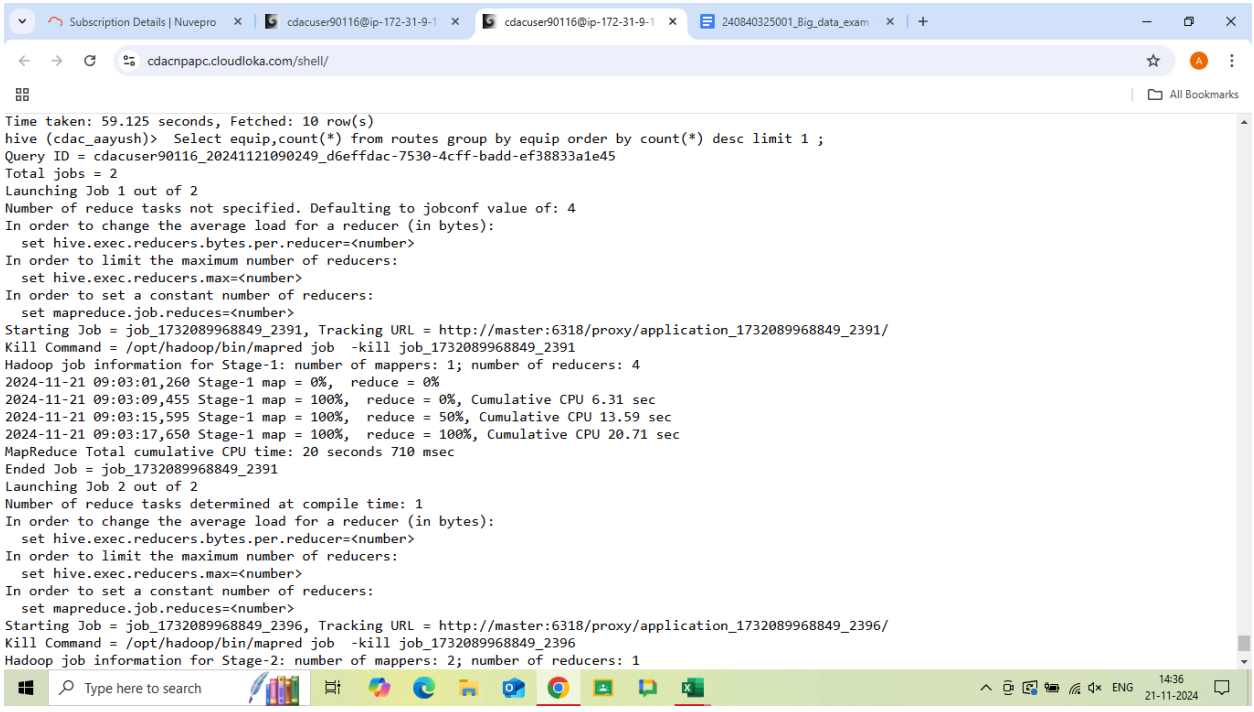
1. Select * from airports where iata in (select src from routes) and iata in (select dest from routes) limit 10;

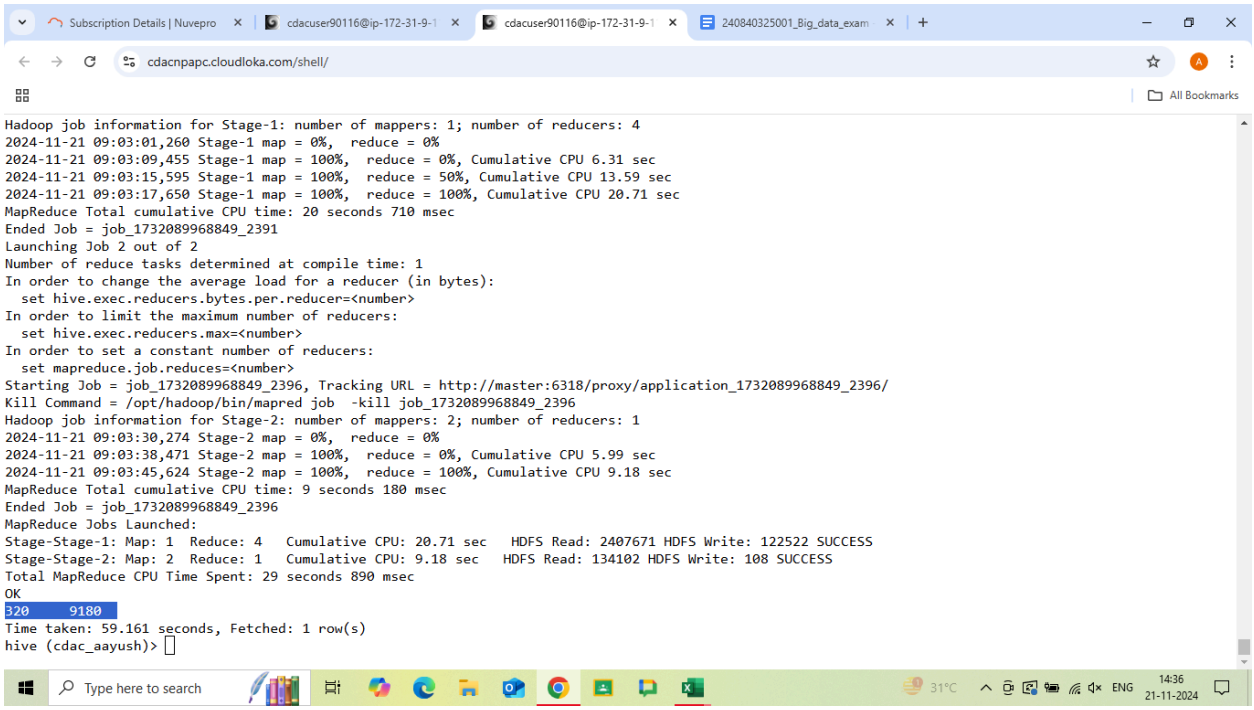

Subscription Details | Nuvepro
cdacuser90116@ip-172-31-9-1
240840325001_Big_data_exam
cdacnpapc.cloudloka.com/shell/
All Bookmarks
dest string
dest_id int
codeshare string
stops int
equip string
Time taken: 0.043 seconds, Fetched: 9 row(s)
hive (cdac_aayush)> Select * from airports where iata in (select src from routes) and iata in (select dest from routes) limit 10;
Query ID = cdacuser90116_20241121084718_3c06b6d8-303b-4b12-8165-a23d49a634b0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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>
Starting Job = job_1732089968849_2278, Tracking URL = http://master:6318/proxy/application_1732089968849_2278/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2278
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 4
2024-11-21 08:47:33,813 Stage-2 map = 0%, reduce = 0%
2024-11-21 08:47:42,076 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 6.23 sec
2024-11-21 08:47:48,257 Stage-2 map = 100%, reduce = 50%, Cumulative CPU 13.48 sec
2024-11-21 08:47:50,311 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 17.04 sec
2024-11-21 08:47:51,345 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 20.53 sec
MapReduce Total cumulative CPU time: 20 seconds 530 msec
Ended Job = job_1732089968849_2278
Launching Job 2 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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:
Type here to search
ENG
14:18
21-11-2024


Subscription Details | Nuvepro
cdacuser90116@ip-172-31-9-1
240840325001_Big_data_exam
cdacnpapc.cloudloka.com/shell/
All Bookmarks
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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>
Starting Job = job_1732089968849_2278, Tracking URL = http://master:6318/proxy/application_1732089968849_2278/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2278
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 4
2024-11-21 08:47:33,813 Stage-2 map = 0%, reduce = 0%
2024-11-21 08:47:42,076 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 6.23 sec
2024-11-21 08:47:48,257 Stage-2 map = 100%, reduce = 50%, Cumulative CPU 13.48 sec
2024-11-21 08:47:50,311 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 17.04 sec
2024-11-21 08:47:51,345 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 20.53 sec
MapReduce Total cumulative CPU time: 20 seconds 530 msec
Ended Job = job_1732089968849_2278
Launching Job 2 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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>
Starting Job = job_1732089968849_2284, Tracking URL = http://master:6318/proxy/application_1732089968849_2284/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2284
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 4
2024-11-21 08:48:04,560 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:48:12,777 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 19.45 sec
2024-11-21 08:48:18,936 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 30.72 sec
Type here to search
ENG
14:19
21-11-2024

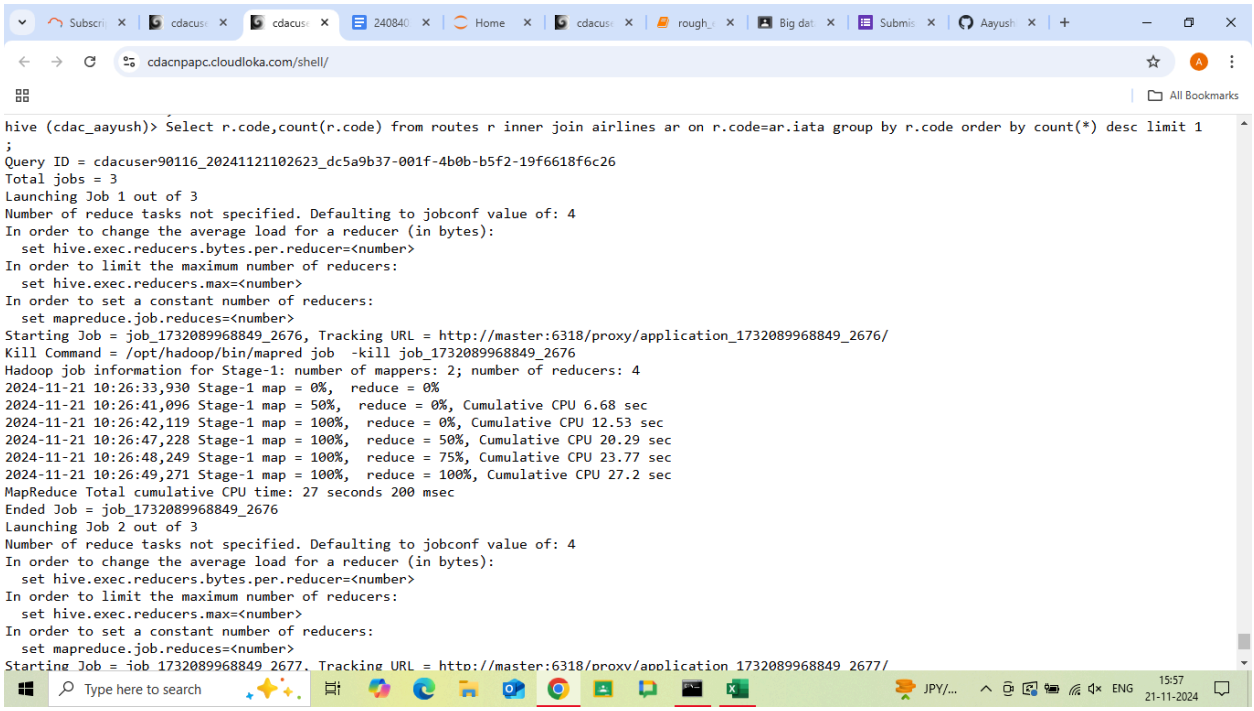


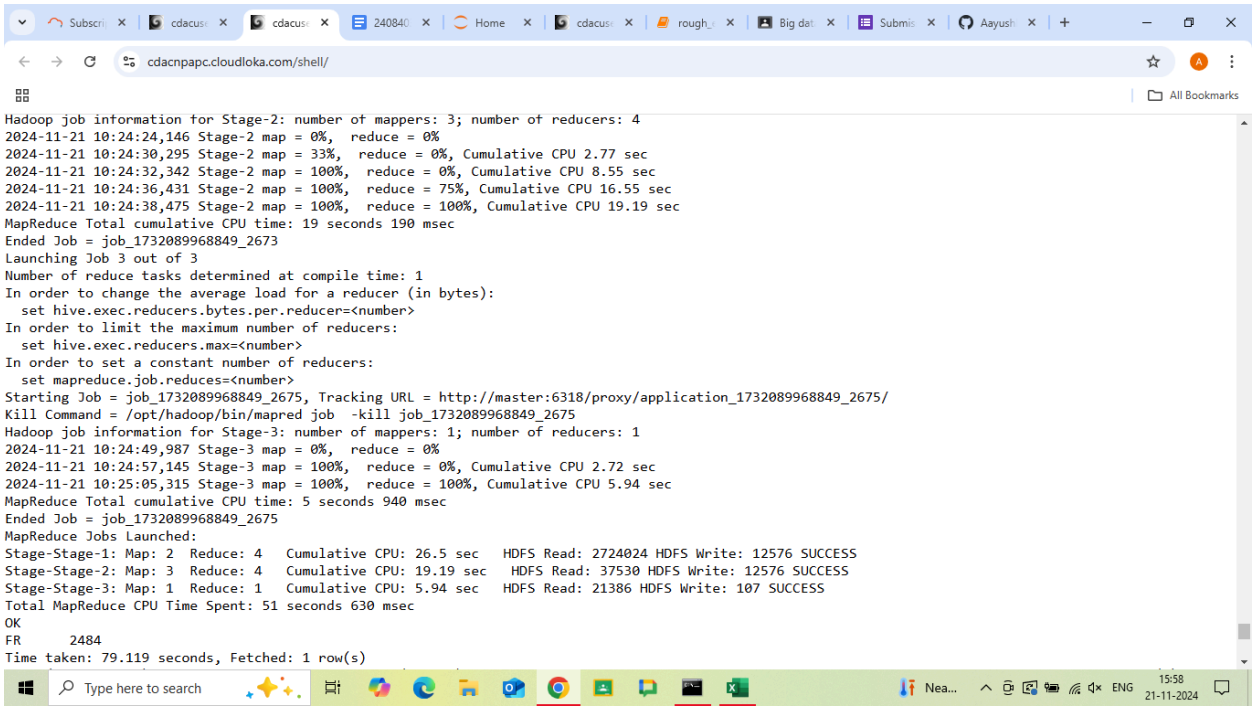
2. Select equip,count(*) from routes group by equip order by count(*) limit 1 ;





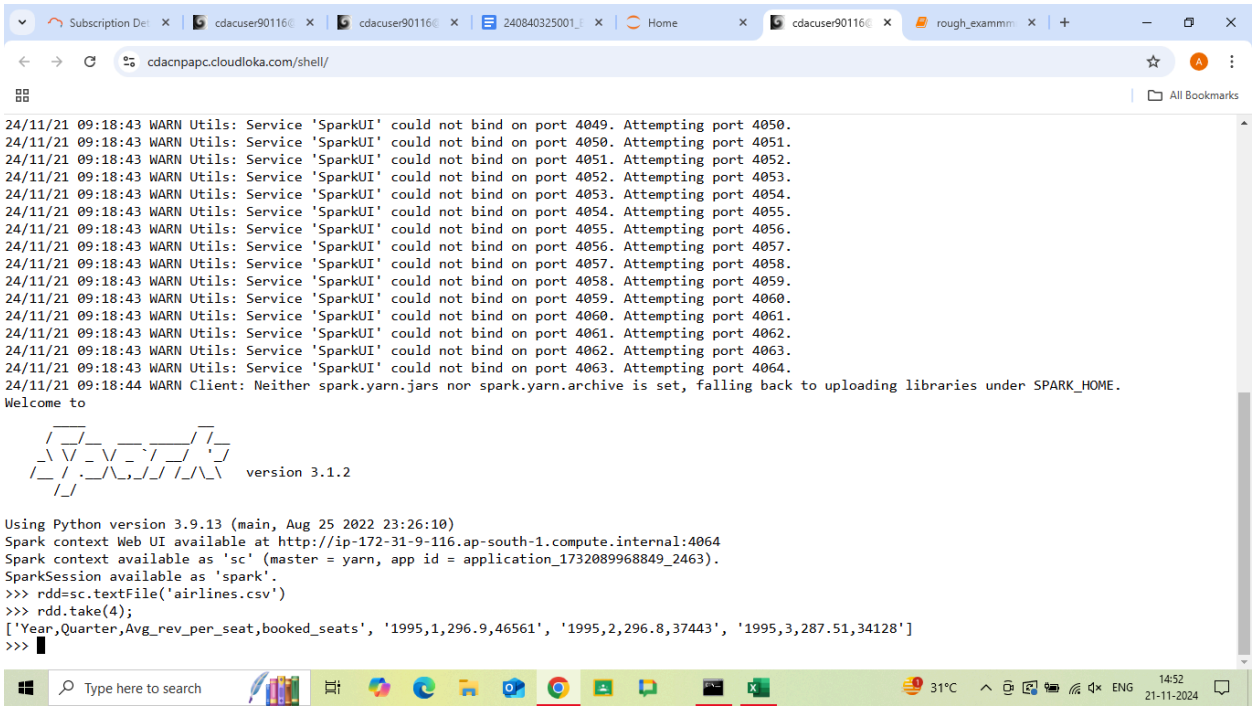
3. Select r.code,count(r.code) from routes r inner join airlines ar on r.code=ar.iata group by r.code order by count(*) desc limit 1;




cdacnpapc.cloudloka.com/shell/
All Bookmarks
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 4
2024-11-21 10:24:24,146 Stage-2 map = 0%, reduce = 0%
2024-11-21 10:24:30,295 Stage-2 map = 33%, reduce = 0%, Cumulative CPU 2.77 sec
2024-11-21 10:24:32,342 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 8.55 sec
2024-11-21 10:24:36,431 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 16.55 sec
2024-11-21 10:24:38,475 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 19.19 sec
MapReduce Total cumulative CPU time: 19 seconds 190 msec
Ended Job = job_1732089968849_2673
Launching Job 3 out of 3
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>
Starting Job = job_1732089968849_2675, Tracking URL = http://master:6318/proxy/application_1732089968849_2675/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2675
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2024-11-21 10:24:49,987 Stage-3 map = 0%, reduce = 0%
2024-11-21 10:24:57,145 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 2.72 sec
2024-11-21 10:25:05,315 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 5.94 sec
MapReduce Total cumulative CPU time: 5 seconds 940 msec
Ended Job = job_1732089968849_2675
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 26.5 sec HDFS Read: 2724024 HDFS Write: 12576 SUCCESS
Stage-Stage-2: Map: 3 Reduce: 4 Cumulative CPU: 19.19 sec HDFS Read: 37530 HDFS Write: 12576 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 5.94 sec HDFS Read: 21386 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 630 msec
OK
FR 2484
Time taken: 79.119 seconds, Fetched: 1 row(s)
Type here to search
ENG
15:58
21-11-2024

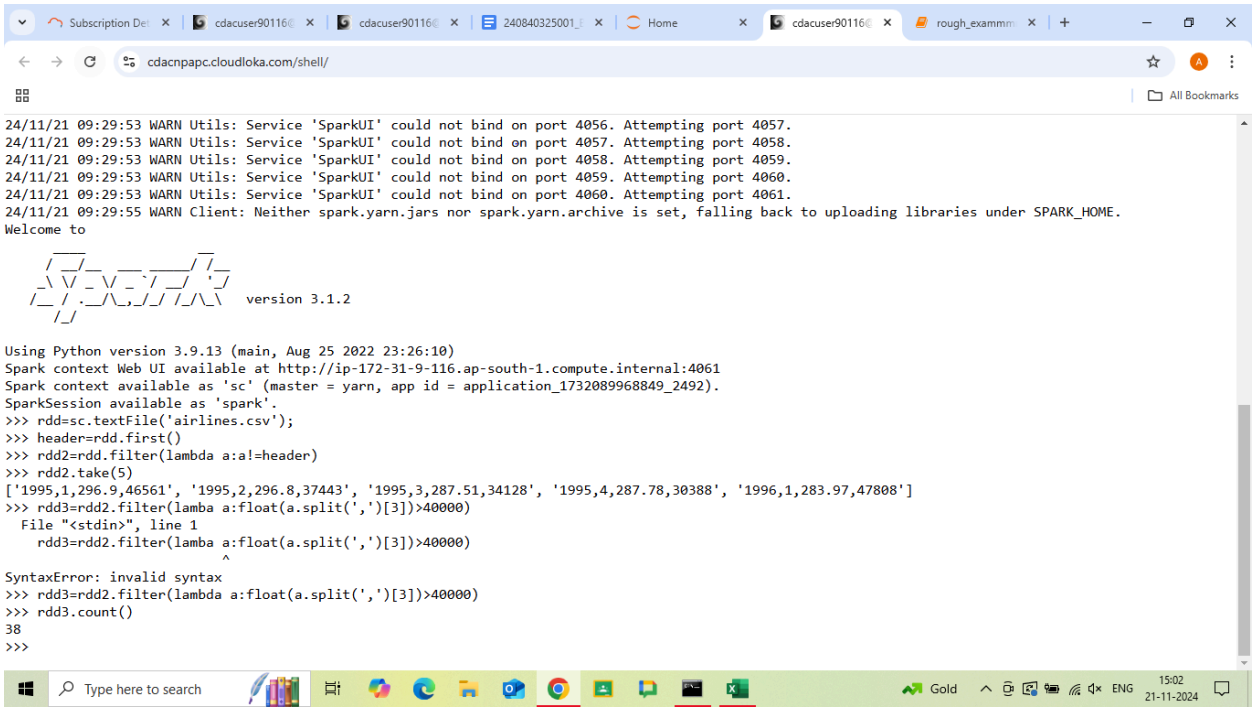
SPARK-

Q1.



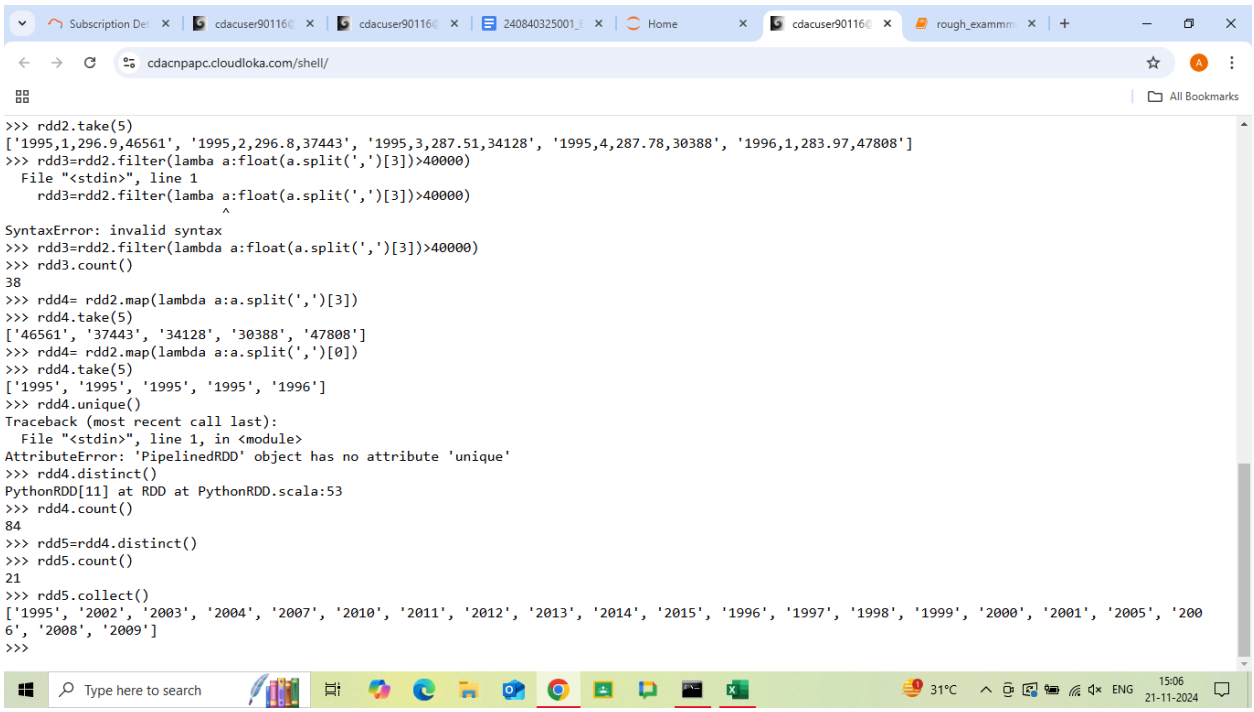
1.

```
header=rdd.first()
rdd2=rdd.filter(lambda a:a!=header)
rdd2.take(5)
rdd3=rdd2.filter(lamba a:float(a.split(',')[3])>40000)
rdd3.count()
```



2.

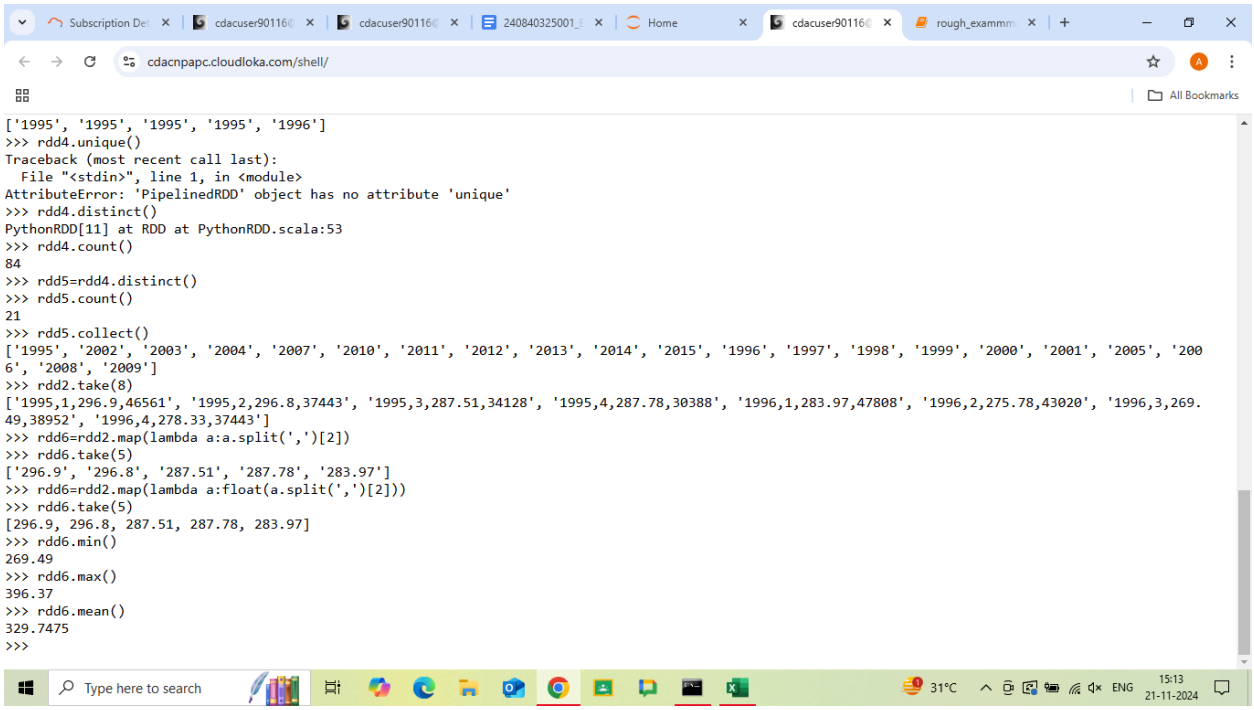
```
rdd4= rdd2.map(lambda a:a.split(',')[0])
rdd4.take(5)
rdd5=rdd4.distinct()
rdd5.count()
rdd5.collect()
```



Q2.

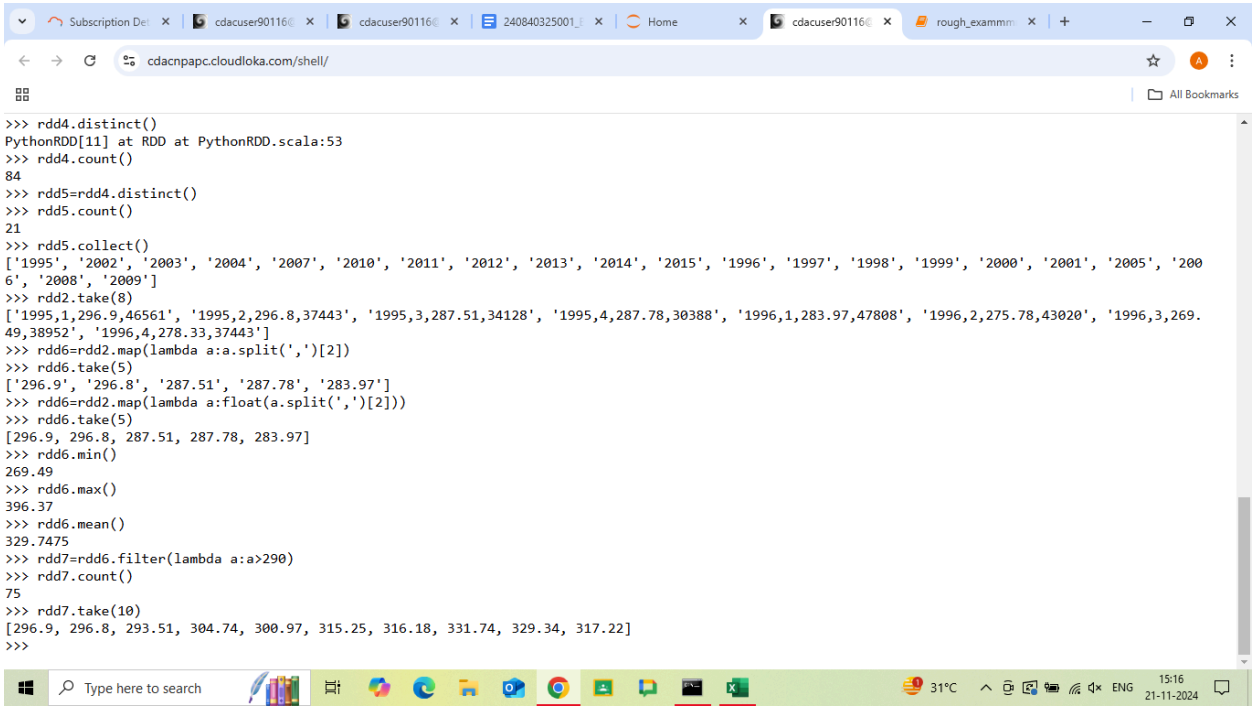
1.

```
rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
rdd6.take(5)
rdd6.min()
rdd6.max()
rdd6.mean()
```

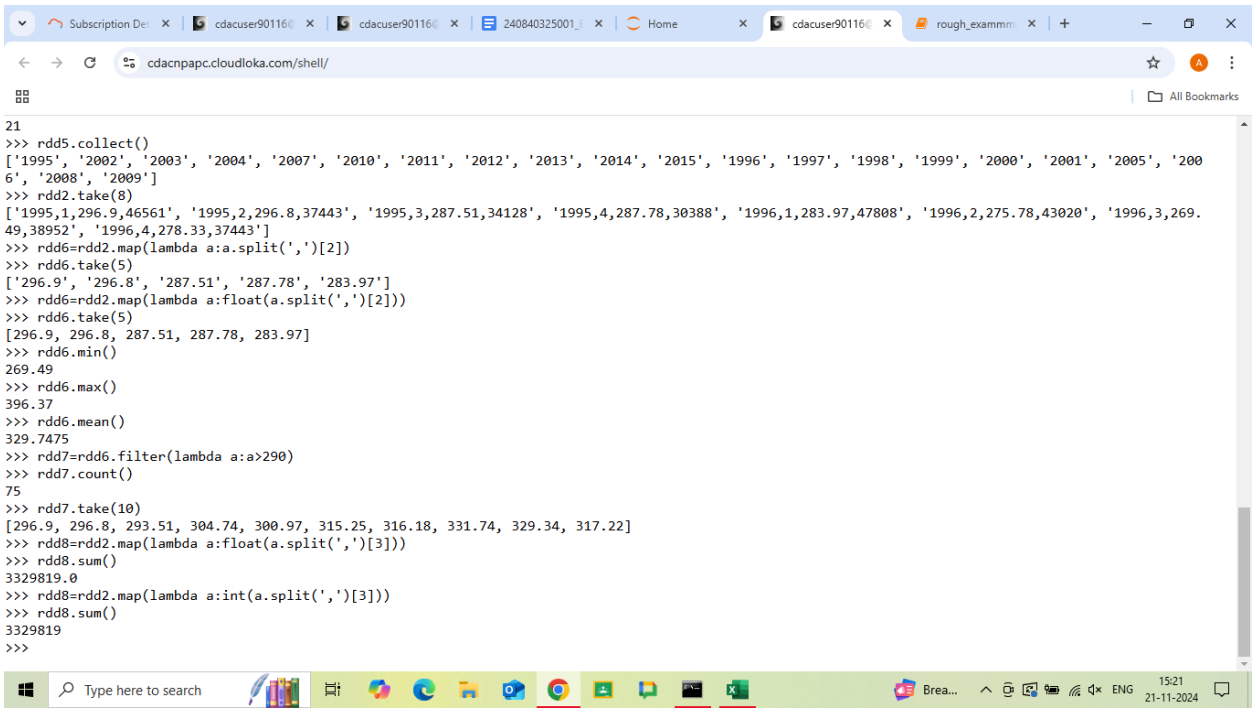


2.

rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
rdd7=rdd6.filter(lambda a:a>290)
rdd7.count()

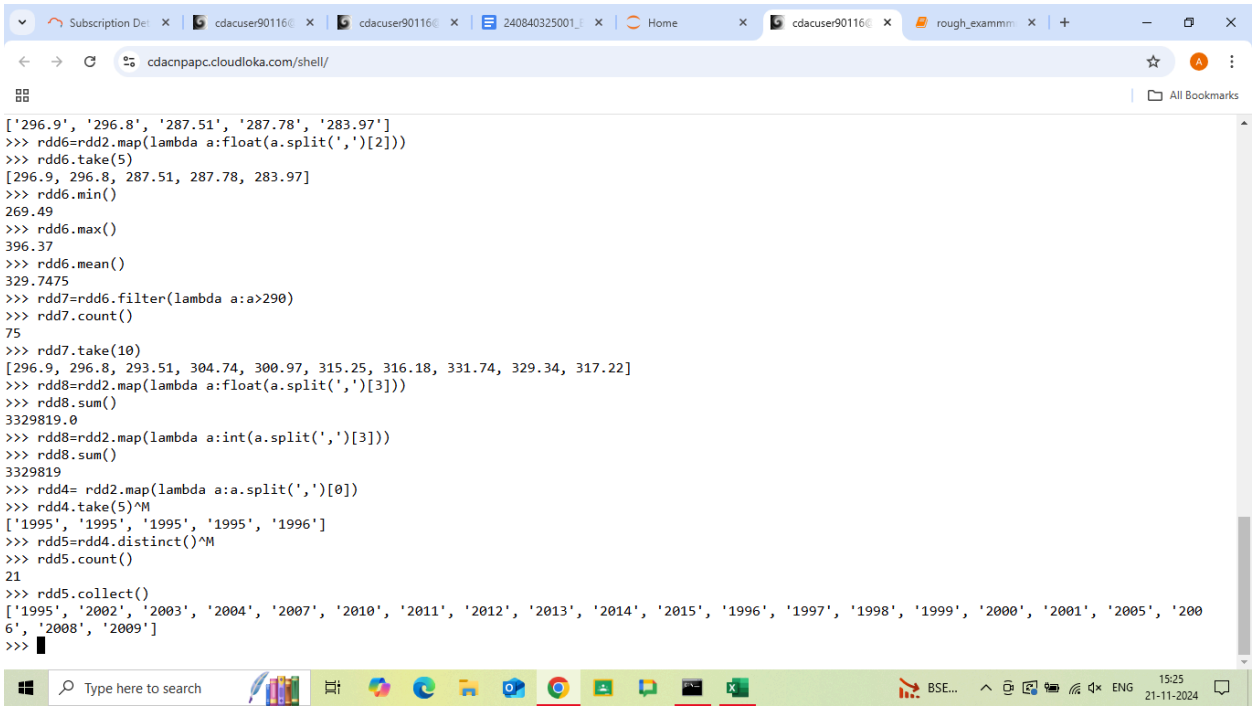


3. rdd8=rdd2.map(lambda a:int(a.split(',')[3]))
   rdd8.sum()



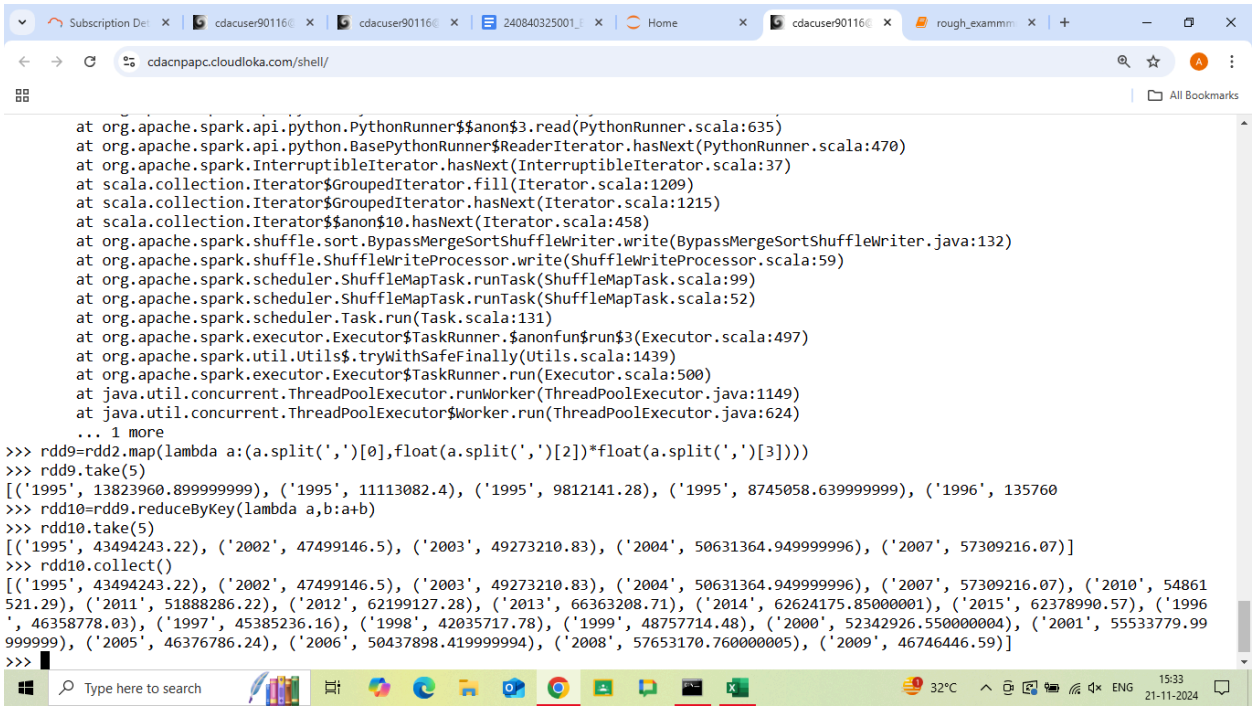
4.

   rdd4= rdd2.map(lambda a:a.split(',')[0])
   rdd4.take(5)
   rdd5=rdd4.distinct()
   rdd5.count()
   rdd5.collect()



5.

```
rdd9=rdd2.map(lambda a:(a.split(',')[0],float(a.split(',')[2])*float(a.split(',')[3])))
rdd9.take(5)
rdd10=rdd9.reduceByKey(lambda a,b:a+b)
rdd10.take(5)
rdd10.collect()
```


cdacnpapc.cloudloka.com/shell/
at org.apache.spark.api.python.PythonRunner$$anon$3.read(PythonRunner.scala:635)
at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.hasNext(PythonRunner.scala:470)
at org.apache.spark.InterruptibleIterator.hasNext(InterruptibleIterator.scala:37)
at scala.collection.Iterator$GroupedIterator.fill(Iterator.scala:1209)
at scala.collection.Iterator$GroupedIterator.hasNext(Iterator.scala:1215)
at scala.collection.Iterator$$anon$10.hasNext(Iterator.scala:458)
at org.apache.spark.shuffle.sort.BypassMergeSortShuffleWriter.write(BypassMergeSortShuffleWriter.java:132)
at org.apache.spark.shuffle.ShuffleWriteProcessor.write(ShuffleWriteProcessor.scala:59)
at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:99)
at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:52)
at org.apache.spark.scheduler.Task.run(Task.scala:131)
at org.apache.spark.executor.Executor$TaskRunner.$anonfunrun3(Executor.scala:497)
at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1439)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:500)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
... 1 more
>>> rdd9=rdd2.map(lambda a:(a.split(',')[0],float(a.split(',')[2])*float(a.split(',')[3])))
>>> rdd9.take(5)
[('1995', 13823960.899999999), ('1995', 11113082.4), ('1995', 9812141.28), ('1995', 8745058.639999999), ('1996', 135760
>>> rdd10=rdd9.reduceByKey(lambda a,b:a+b)
>>> rdd10.take(5)
[('1995', 43494243.22), ('2002', 47499146.5), ('2003', 49273210.83), ('2004', 50631364.949999996), ('2007', 57309216.07)]
>>> rdd10.collect()
[('1995', 43494243.22), ('2002', 47499146.5), ('2003', 49273210.83), ('2004', 50631364.949999996), ('2007', 57309216.07), ('2010', 54861
521.29), ('2011', 51888286.22), ('2012', 62199127.28), ('2013', 66363208.71), ('2014', 62624175.85000001), ('2015', 62378990.57), ('1996
', 46358778.03), ('1997', 45385236.16), ('1998', 42035717.78), ('1999', 48757714.48), ('2000', 52342926.550000004), ('2001', 55533779.99
999999), ('2005', 46376786.24), ('2006', 50437898.419999994), ('2008', 57653170.760000005), ('2009', 46746446.59)]
>>>
Type here to search
32°C
ENG
15:33
21-11-2024