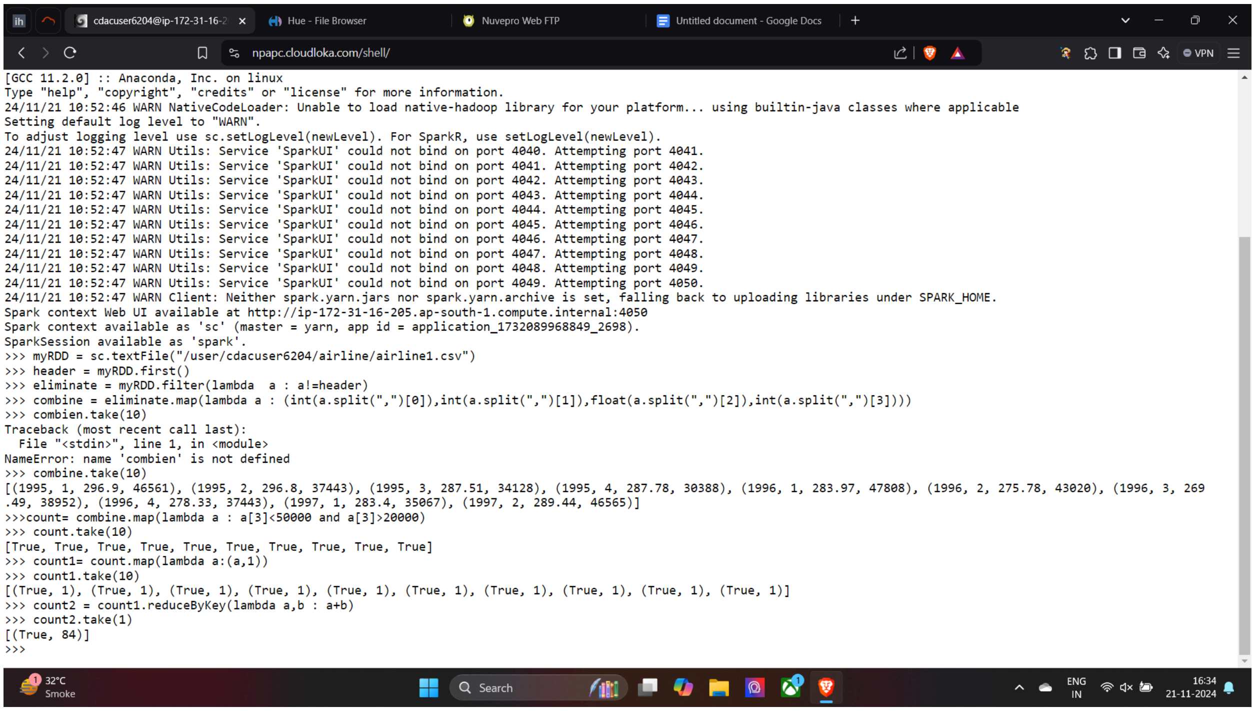
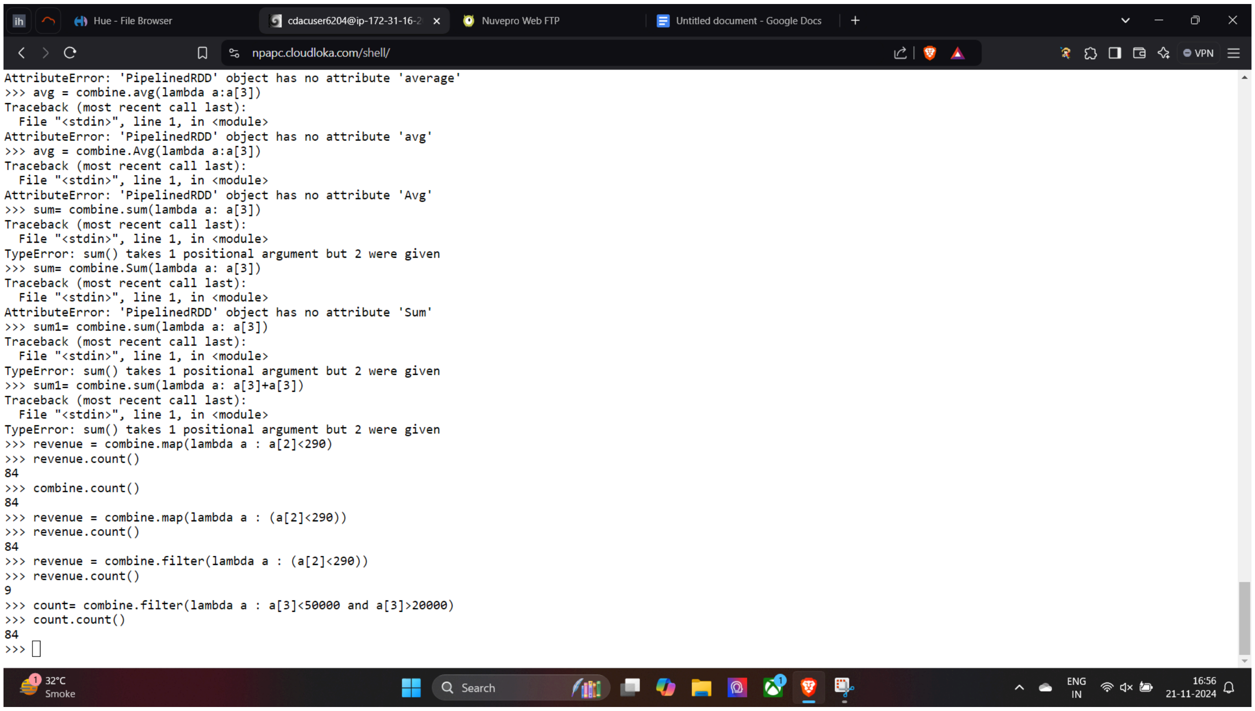
Spark

Q1 1)

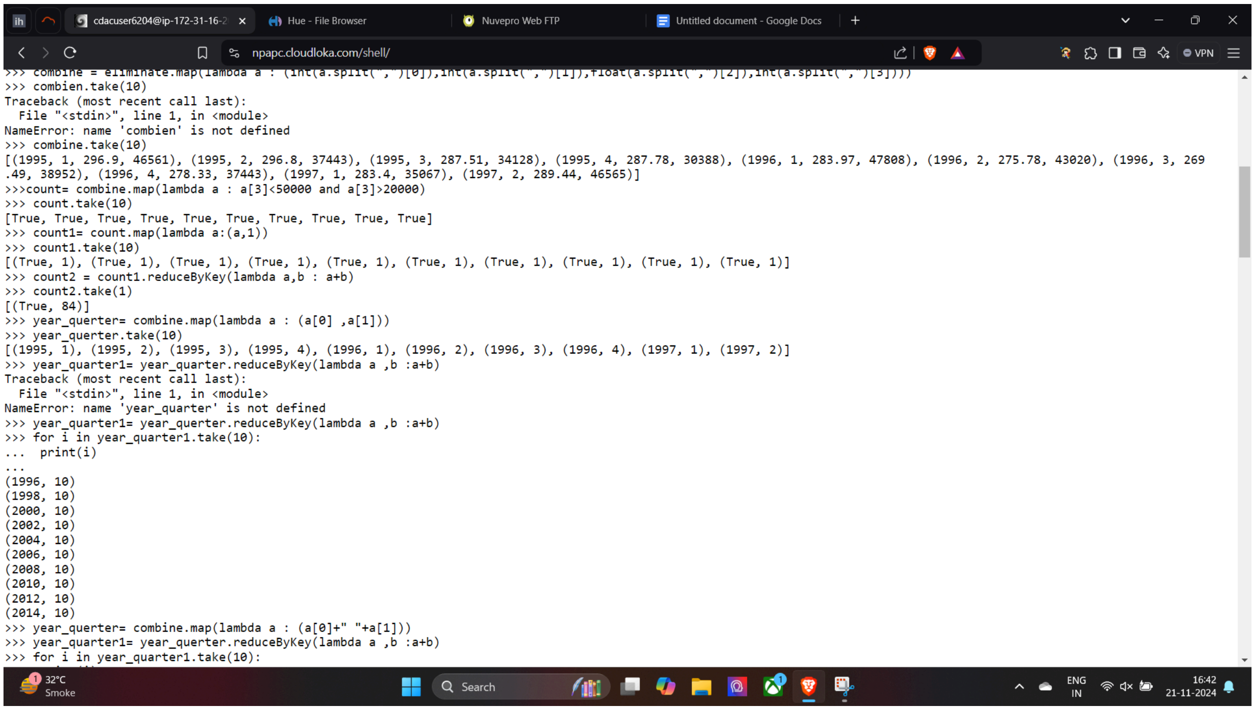

cdacuser6204@ip-172-31-16-2
Hue - File Browser
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:34
21-11-2024

```
[GCC 11.2.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
24/11/21 10:52:46 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4046. Attempting port 4047.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4047. Attempting port 4048.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4048. Attempting port 4049.
24/11/21 10:52:47 WARN Utils: Service 'SparkUI' could not bind on port 4049. Attempting port 4050.
24/11/21 10:52:47 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Spark context Web UI available at http://ip-172-31-16-205.ap-south-1.compute.internal:4050
Spark context available as 'sc' (master = yarn, app id = application_1732089968849_2698).
SparkSession available as 'spark'.
>>> myRDD = sc.textFile("/user/cdacuser6204/airline/airline1.csv")
>>> header = myRDD.first()
>>> eliminate = myRDD.filter(lambda  a : a!=header)
>>> combine = eliminate.map(lambda a : (int(a.split(",")[0]),int(a.split(",")[1]),float(a.split(",")[2]),int(a.split(",")[3])))
>>> combien.take(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'combien' is not defined
>>> combine.take(10)
[(1995, 1, 296.9, 46561), (1995, 2, 296.8, 37443), (1995, 3, 287.51, 34128), (1995, 4, 287.78, 30388), (1996, 1, 283.97, 47808), (1996, 2, 275.78, 43020), (1996, 3, 269
.49, 38952), (1996, 4, 278.33, 37443), (1997, 1, 283.4, 35067), (1997, 2, 289.44, 46565)]
>>>count= combine.map(lambda a : a[3]<50000 and a[3]>20000)
>>> count.take(10)
[True, True, True, True, True, True, True, True, True, True]
>>> count1= count.map(lambda a:(a,1))
>>> count1.take(10)
[(True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1)]
>>> count2 = count1.reduceByKey(lambda a,b : a+b)
>>> count2.take(1)
[(True, 84)]
>>>
```


Hue - File Browser
cdacuser6204@ip-172-31-16-2
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:56
21-11-2024

```
AttributeError: 'PipelinedRDD' object has no attribute 'average'
>>> avg = combine.avg(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'avg'
>>> avg = combine.Avg(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Avg'
>>> sum= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum= combine.Sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Sum'
>>> sum1= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum1= combine.sum(lambda a: a[3]+a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> revenue = combine.map(lambda a : a[2]<290)
>>> revenue.count()
84
>>> combine.count()
84
>>> revenue = combine.map(lambda a : (a[2]<290))
>>> revenue.count()
84
>>> revenue = combine.filter(lambda a : (a[2]<290))
>>> revenue.count()
9
>>> count= combine.filter(lambda a : a[3]<50000 and a[3]>20000)
>>> count.count()
84
>>> 
```

2)

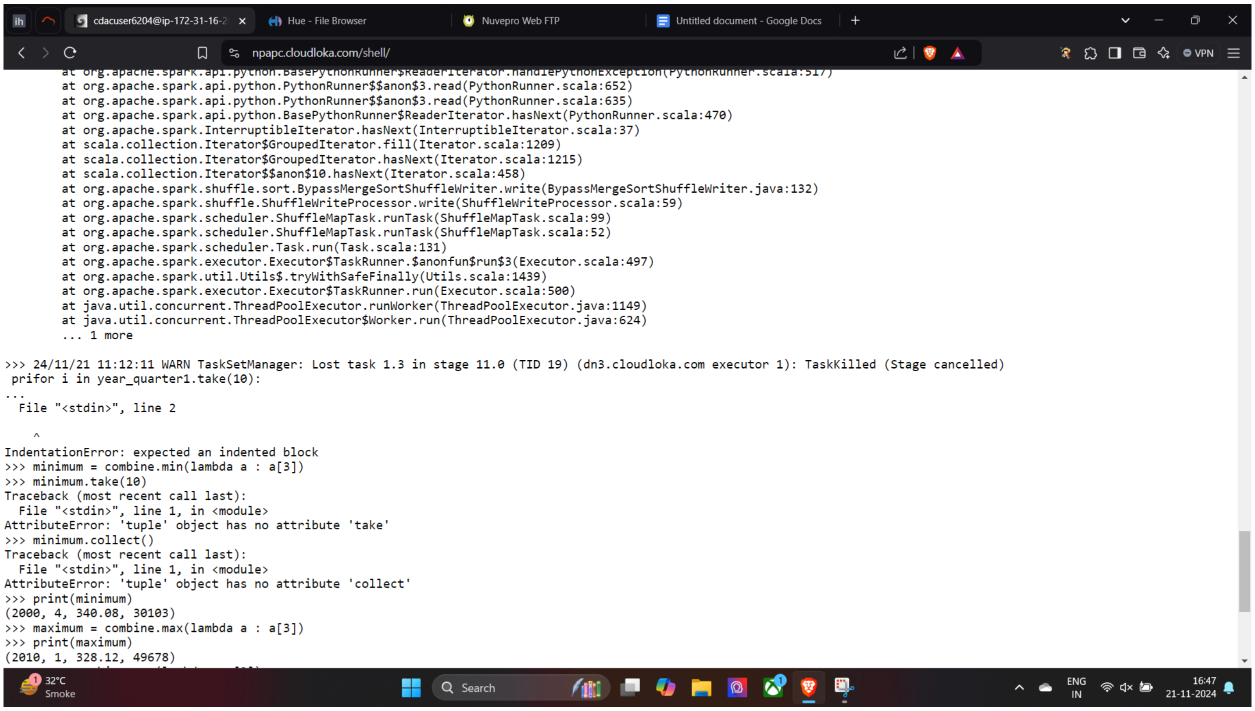


cdacuser6204@ip-172-31-16-2
Hue - File Browser
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:42
21-11-2024

```
>>> combine = eliminate.map(lambda a : (int(a.split(",")[0]),int(a.split(",")[1]),float(a.split(",")[2]),int(a.split(",")[3])))
>>> combien.take(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'combien' is not defined
>>> combine.take(10)
[(1995, 1, 296.9, 46561), (1995, 2, 296.8, 37443), (1995, 3, 287.51, 34128), (1995, 4, 287.78, 30388), (1996, 1, 283.97, 47808), (1996, 2, 275.78, 43020), (1996, 3, 269
.49, 38952), (1996, 4, 278.33, 37443), (1997, 1, 283.4, 35067), (1997, 2, 289.44, 46565)]
>>>count= combine.map(lambda a : a[3]<50000 and a[3]>20000)
>>> count.take(10)
[True, True, True, True, True, True, True, True, True, True]
>>> count1= count.map(lambda a:(a,1))
>>> count1.take(10)
[(True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1), (True, 1)]
>>> count2 = count1.reduceByKey(lambda a,b : a+b)
>>> count2.take(1)
[(True, 84)]
>>> year_querter= combine.map(lambda a : (a[0] ,a[1]))
>>> year_querter.take(10)
[(1995, 1), (1995, 2), (1995, 3), (1995, 4), (1996, 1), (1996, 2), (1996, 3), (1996, 4), (1997, 1), (1997, 2)]
>>> year_quarter1= year_quarter.reduceByKey(lambda a ,b :a+b)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'year_quarter' is not defined
>>> year_quarter1= year_querter.reduceByKey(lambda a ,b :a+b)
>>> for i in year_quarter1.take(10):
...  print(i)
...
(1996, 10)
(1998, 10)
(2000, 10)
(2002, 10)
(2004, 10)
(2006, 10)
(2008, 10)
(2010, 10)
(2012, 10)
(2014, 10)
>>> year_querter= combine.map(lambda a : (a[0]+" "+a[1]))
>>> year_quarter1= year_querter.reduceByKey(lambda a ,b :a+b)
>>> for i in year_quarter1.take(10):
```

Q2

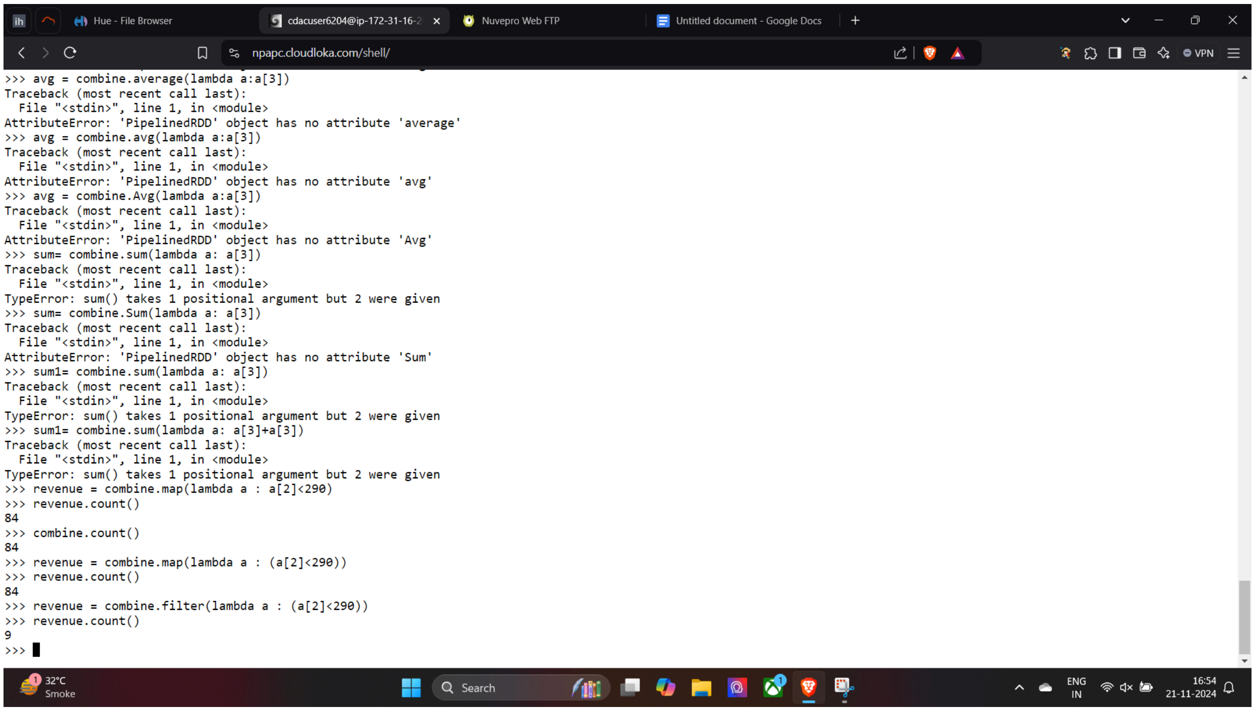
1)



cdacuser6204@ip-172-31-16-2
Hue - File Browser
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:47
21-11-2024

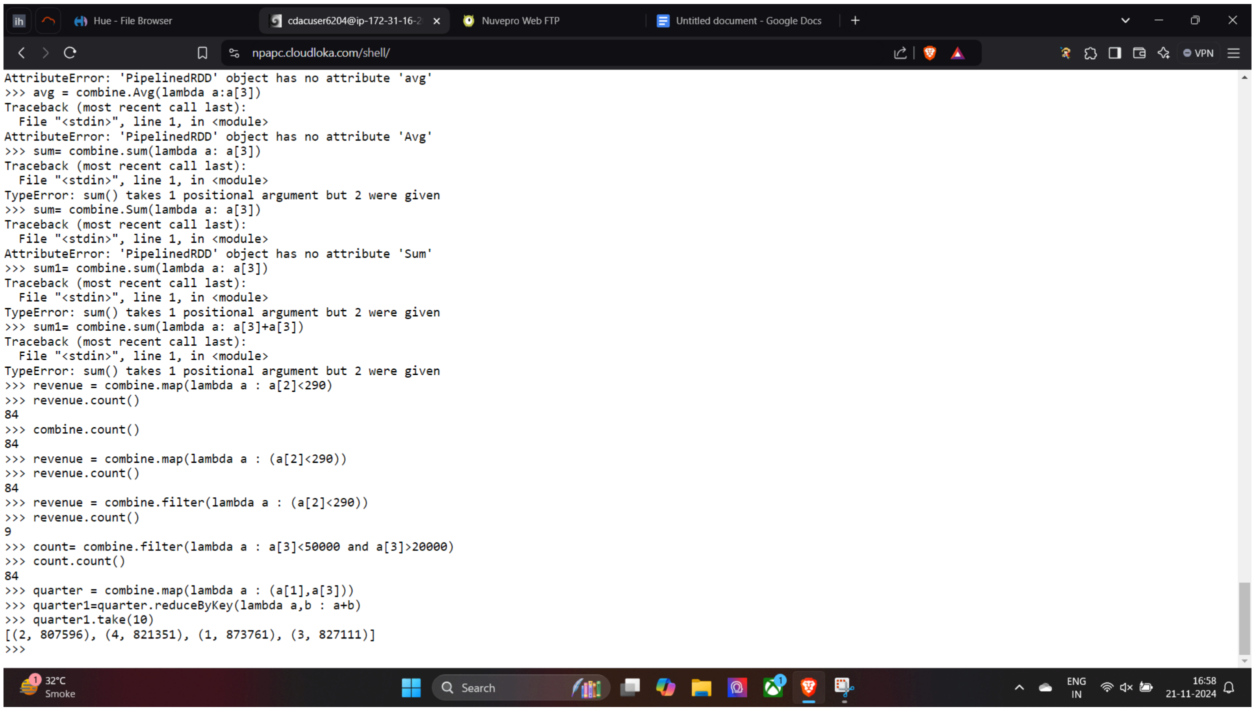
```
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.handlePythonException(PythonRunner.scala:517)
        at org.apache.spark.api.python.PythonRunner$$anon$3.read(PythonRunner.scala:652)
        at org.apache.spark.api.python.PythonRunner$$anon$3.read(PythonRunner.scala:635)
        at org.apache.spark.api.python.BasePythonRunner$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.$anonfun$run$3(Executor.scala:497)
        at org.apache.spark.util.Utils$.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

>>> 24/11/21 11:12:11 WARN TaskSetManager: Lost task 1.3 in stage 11.0 (TID 19) (dn3.cloudloka.com executor 1): TaskKilled (Stage cancelled)
 prifor i in year_quarter1.take(10):
...
  File "<stdin>", line 2

    ^
IndentationError: expected an indented block
>>> minimum = combine.min(lambda a : a[3])
>>> minimum.take(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'tuple' object has no attribute 'take'
>>> minimum.collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'tuple' object has no attribute 'collect'
>>> print(minimum)
(2000, 4, 340.08, 30103)
>>> maximum = combine.max(lambda a : a[3])
>>> print(maximum)
(2010, 1, 328.12, 49678)
```

2)


Hue - File Browser
cdacuser6204@ip-172-31-16-2
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:54
21-11-2024

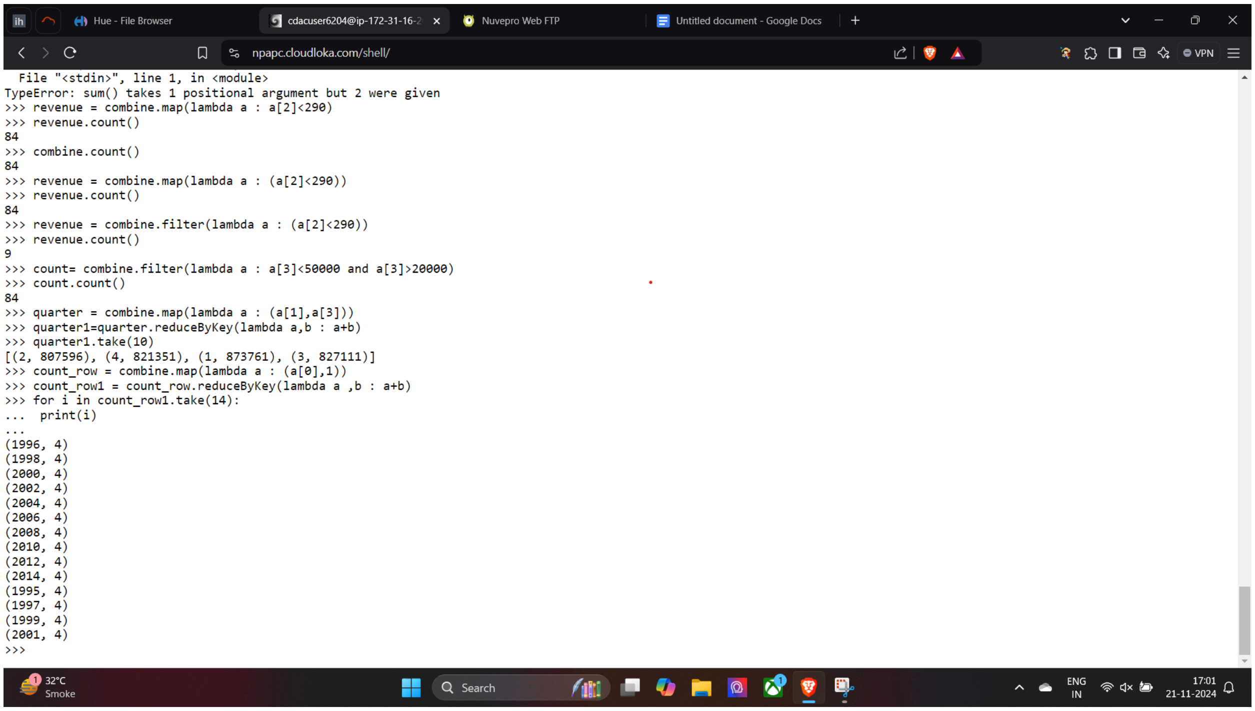
```
>>> avg = combine.average(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'average'
>>> avg = combine.avg(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'avg'
>>> avg = combine.Avg(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Avg'
>>> sum= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum= combine.Sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Sum'
>>> sum1= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum1= combine.sum(lambda a: a[3]+a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> revenue = combine.map(lambda a : a[2]<290)
>>> revenue.count()
84
>>> combine.count()
84
>>> revenue = combine.map(lambda a : (a[2]<290))
>>> revenue.count()
84
>>> revenue = combine.filter(lambda a : (a[2]<290))
>>> revenue.count()
9
>>> █
```

3)


Hue - File Browser
cdacuser6204@ip-172-31-16-2
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
16:58
21-11-2024

```
AttributeError: 'PipelinedRDD' object has no attribute 'avg'
>>> avg = combine.Avg(lambda a:a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Avg'
>>> sum= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum= combine.Sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'Sum'
>>> sum1= combine.sum(lambda a: a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> sum1= combine.sum(lambda a: a[3]+a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> revenue = combine.map(lambda a : a[2]<290)
>>> revenue.count()
84
>>> combine.count()
84
>>> revenue = combine.map(lambda a : (a[2]<290))
>>> revenue.count()
84
>>> revenue = combine.filter(lambda a : (a[2]<290))
>>> revenue.count()
9
>>> count= combine.filter(lambda a : a[3]<50000 and a[3]>20000)
>>> count.count()
84
>>> quarter = combine.map(lambda a : (a[1],a[3]))
>>> quarter1=quarter.reduceByKey(lambda a,b : a+b)
>>> quarter1.take(10)
[(2, 807596), (4, 821351), (1, 873761), (3, 827111)]
>>>
```

4)


Hue - File Browser
cdacuser6204@ip-172-31-16-2
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
17:01
21-11-2024

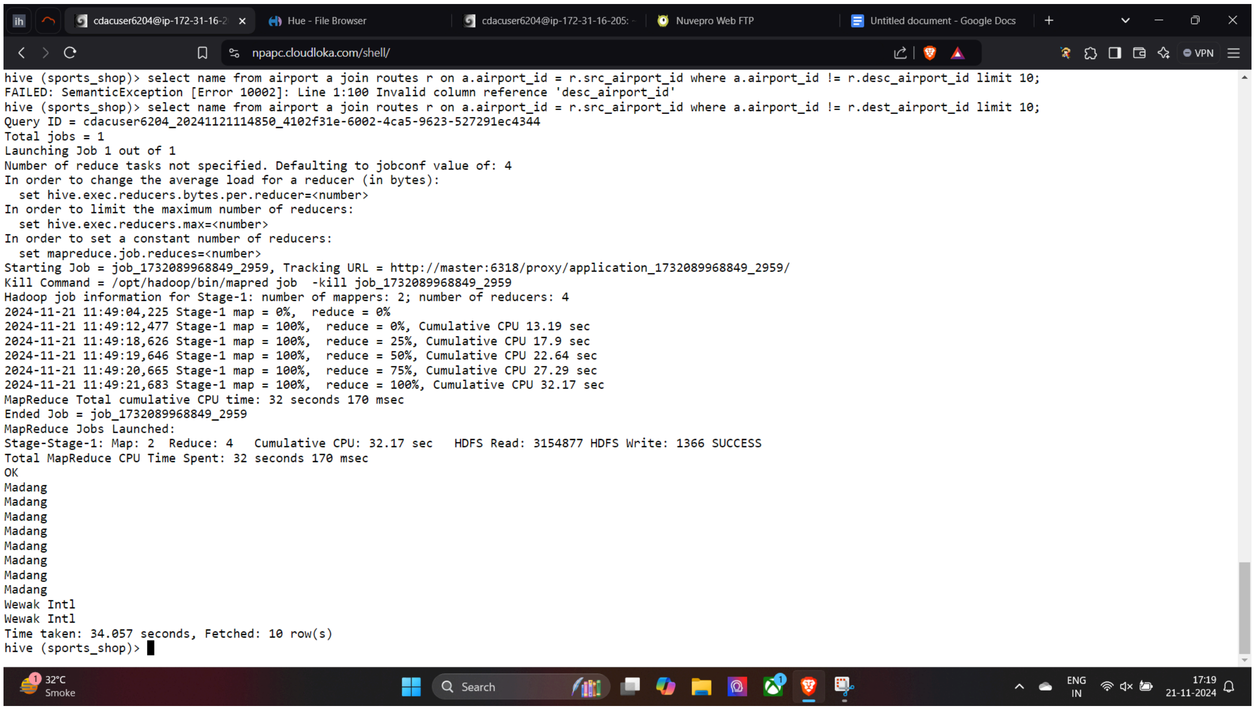
```
  File "<stdin>", line 1, in <module>
TypeError: sum() takes 1 positional argument but 2 were given
>>> revenue = combine.map(lambda a : a[2]<290)
>>> revenue.count()
84
>>> combine.count()
84
>>> revenue = combine.map(lambda a : (a[2]<290))
>>> revenue.count()
84
>>> revenue = combine.filter(lambda a : (a[2]<290))
>>> revenue.count()
9
>>> count= combine.filter(lambda a : a[3]<50000 and a[3]>20000)
>>> count.count()
84
>>> quarter = combine.map(lambda a : (a[1],a[3]))
>>> quarter1=quarter.reduceByKey(lambda a,b : a+b)
>>> quarter1.take(10)
[(2, 807596), (4, 821351), (1, 873761), (3, 827111)]
>>> count_row = combine.map(lambda a : (a[0],1))
>>> count_row1 = count_row.reduceByKey(lambda a ,b : a+b)
>>> for i in count_row1.take(14):
...  print(i)
...
(1996, 4)
(1998, 4)
(2000, 4)
(2002, 4)
(2004, 4)
(2006, 4)
(2008, 4)
(2010, 4)
(2012, 4)
(2014, 4)
(1995, 4)
(1997, 4)
(1999, 4)
(2001, 4)
>>>
```

5)

Hive

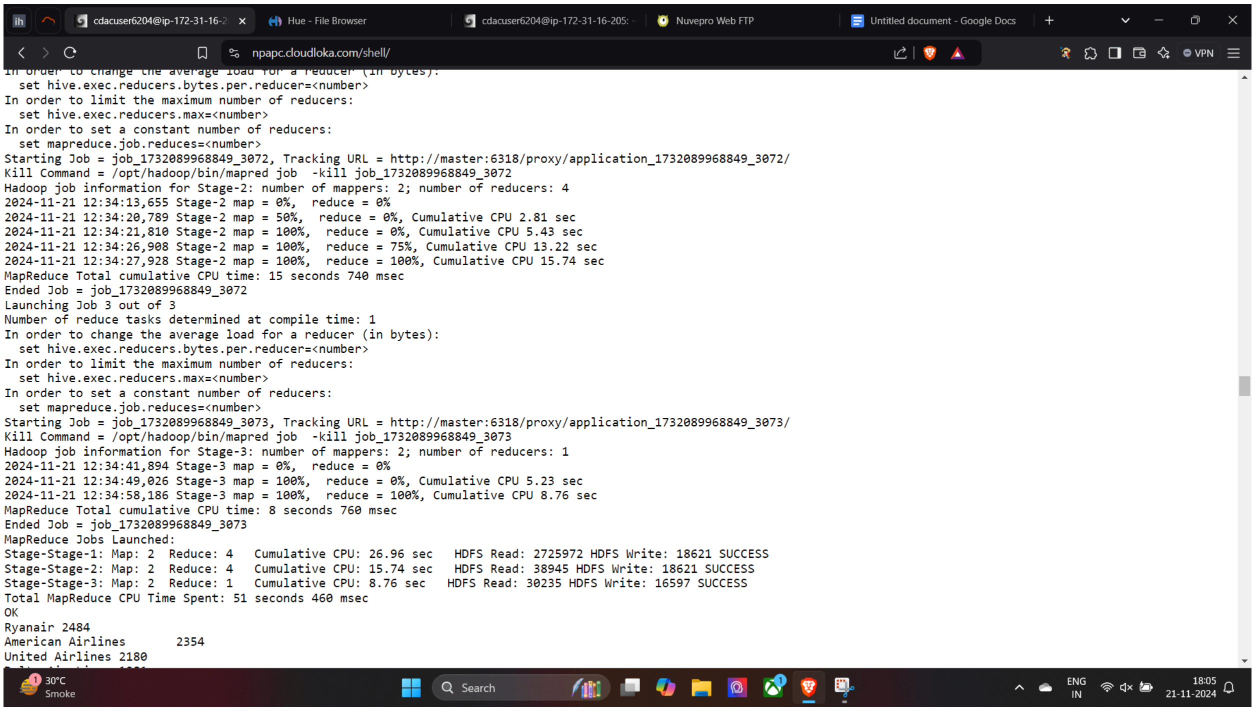
Q1) 1)

```
hive (sports_shop)> select name from airport a join routes r on
a.airport_id = r.src_airport_id where a.airport_id !=
r.dest_airport_id limit 10;
```


cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205: -
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
32°C
Smoke
Search
ENG
IN
17:19
21-11-2024

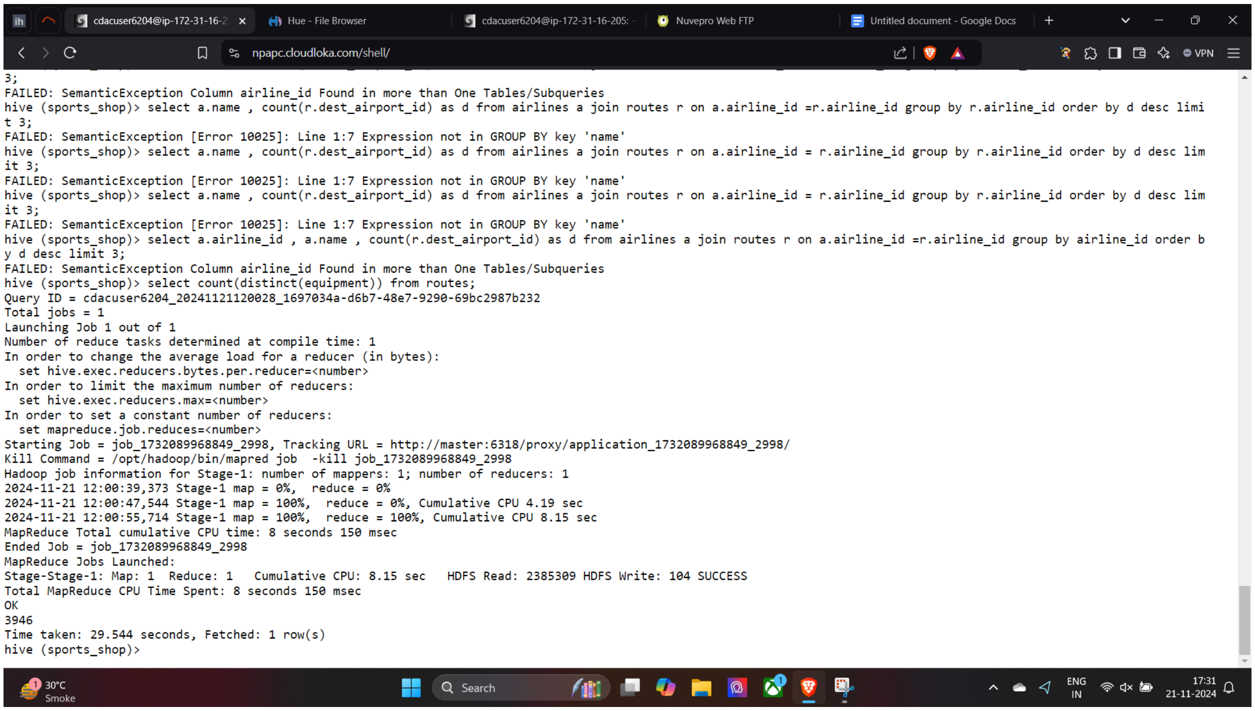
```
hive (sports_shop)> select name from airport a join routes r on a.airport_id = r.src_airport_id where a.airport_id != r.desc_airport_id limit 10;
FAILED: SemanticException [Error 10002]: Line 1:100 Invalid column reference 'desc_airport_id'
hive (sports_shop)> select name from airport a join routes r on a.airport_id = r.src_airport_id where a.airport_id != r.dest_airport_id limit 10;
Query ID = cdacuser6204_20241121114850_4102f31e-6002-4ca5-9623-527291ec4344
Total jobs = 1
Launching Job 1 out of 1
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_2959, Tracking URL = http://master:6318/proxy/application_1732089968849_2959/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2959
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 11:49:04,225 Stage-1 map = 0%,  reduce = 0%
2024-11-21 11:49:12,477 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.19 sec
2024-11-21 11:49:18,626 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 17.9 sec
2024-11-21 11:49:19,646 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 22.64 sec
2024-11-21 11:49:20,665 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 27.29 sec
2024-11-21 11:49:21,683 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 32.17 sec
MapReduce Total cumulative CPU time: 32 seconds 170 msec
Ended Job = job_1732089968849_2959
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 32.17 sec   HDFS Read: 3154877 HDFS Write: 1366 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 170 msec
OK
Madang
Madang
Madang
Madang
Madang
Madang
Madang
Madang
Wewak Intl
Wewak Intl
Time taken: 34.057 seconds, Fetched: 10 row(s)
hive (sports_shop)>
```

2) **hive (sports_shop)> select a.name , count(r.airline_id) as s from airlines a join routes r on a.airline_id=r.airline_id group by a.name order by s desc limit 3;**


cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205: -
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
30°C
Smoke
Search
ENG
IN
18:05
21-11-2024

```
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_3072, Tracking URL = http://master:6318/proxy/application_1732089968849_3072/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_3072
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 12:34:13,655 Stage-2 map = 0%,  reduce = 0%
2024-11-21 12:34:20,789 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.81 sec
2024-11-21 12:34:21,810 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.43 sec
2024-11-21 12:34:26,908 Stage-2 map = 100%,  reduce = 75%, Cumulative CPU 13.22 sec
2024-11-21 12:34:27,928 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 15.74 sec
MapReduce Total cumulative CPU time: 15 seconds 740 msec
Ended Job = job_1732089968849_3072
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_3073, Tracking URL = http://master:6318/proxy/application_1732089968849_3073/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_3073
Hadoop job information for Stage-3: number of mappers: 2; number of reducers: 1
2024-11-21 12:34:41,894 Stage-3 map = 0%,  reduce = 0%
2024-11-21 12:34:49,026 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.23 sec
2024-11-21 12:34:58,186 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 8.76 sec
MapReduce Total cumulative CPU time: 8 seconds 760 msec
Ended Job = job_1732089968849_3073
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 26.96 sec   HDFS Read: 2725972 HDFS Write: 18621 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 15.74 sec   HDFS Read: 38945 HDFS Write: 18621 SUCCESS
Stage-Stage-3: Map: 2  Reduce: 1   Cumulative CPU: 8.76 sec   HDFS Read: 30235 HDFS Write: 16597 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 460 msec
OK
Ryanair 2484
American Airlines       2354
United Airlines 2180
```

3) **hive (sports_shop)> select count(distinct(equipment)) from routes;**


cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205: ~
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
30°C
Smoke
Search
ENG
IN
17:31
21-11-2024

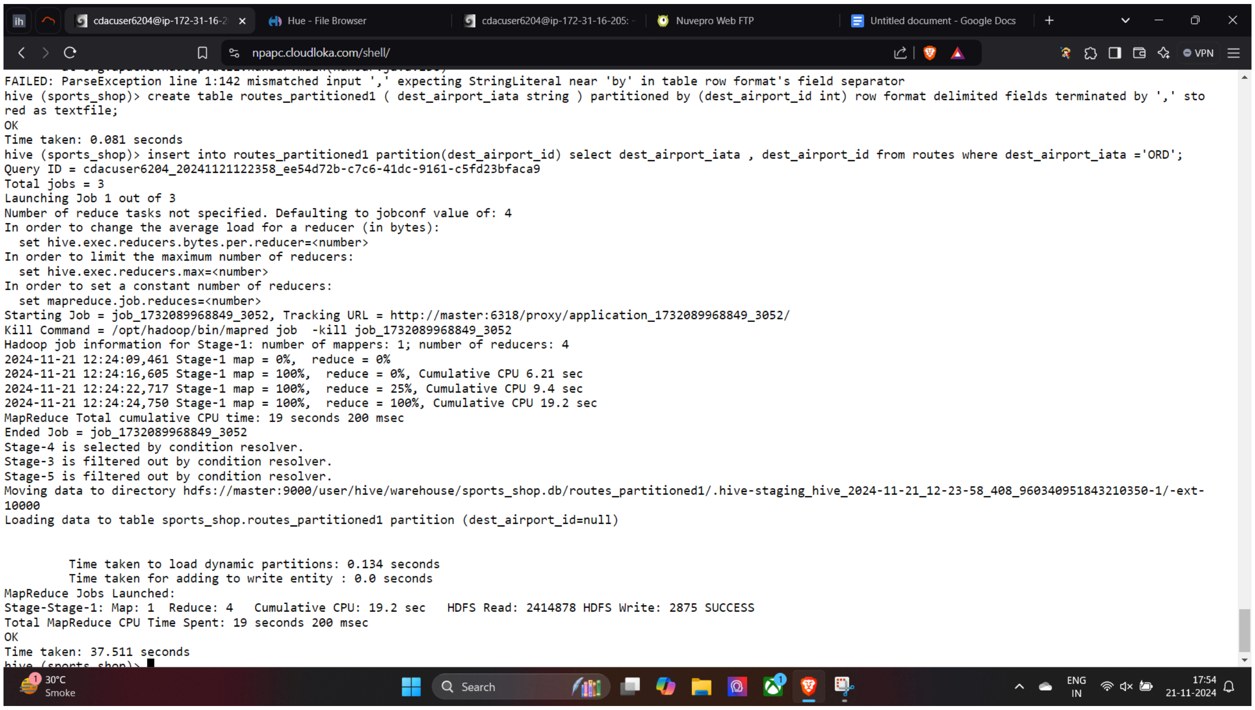
```
3;
FAILED: SemanticException Column airline_id Found in more than One Tables/Subqueries
hive (sports_shop)> select a.name , count(r.dest_airport_id) as d from airlines a join routes r on a.airline_id =r.airline_id group by r.airline_id order by d desc limi
t 3;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'name'
hive (sports_shop)> select a.name , count(r.dest_airport_id) as d from airlines a join routes r on a.airline_id = r.airline_id group by r.airline_id order by d desc lim
it 3;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'name'
hive (sports_shop)> select a.name , count(r.dest_airport_id) as d from airlines a join routes r on a.airline_id = r.airline_id group by r.airline_id order by d desc lim
it 3;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'name'
hive (sports_shop)> select a.airline_id , a.name , count(r.dest_airport_id) as d from airlines a join routes r on a.airline_id =r.airline_id group by airline_id order b
y d desc limit 3;
FAILED: SemanticException Column airline_id Found in more than One Tables/Subqueries
hive (sports_shop)> select count(distinct(equipment)) from routes;
Query ID = cdacuser6204_20241121120028_1697034a-d6b7-48e7-9290-69bc2987b232
Total jobs = 1
Launching Job 1 out of 1
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_2998, Tracking URL = http://master:6318/proxy/application_1732089968849_2998/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2998
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:00:39,373 Stage-1 map = 0%,  reduce = 0%
2024-11-21 12:00:47,544 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.19 sec
2024-11-21 12:00:55,714 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.15 sec
MapReduce Total cumulative CPU time: 8 seconds 150 msec
Ended Job = job_1732089968849_2998
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.15 sec   HDFS Read: 2385309 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 150 msec
OK
3946
Time taken: 29.544 seconds, Fetched: 1 row(s)
hive (sports_shop)>
```

Q2)

```
   hive (sports_shop)> create table routes_partioned(airline_iata
string , airline_id int , src_airport_iata string , src_airport_id int
, dest_airport_iata string , codes
     hare string , stops int , equipment string) partitioned by (
     dest_airport_id int) row format delimited fields terminated by
     ',' stored as textfile;
     OK
```

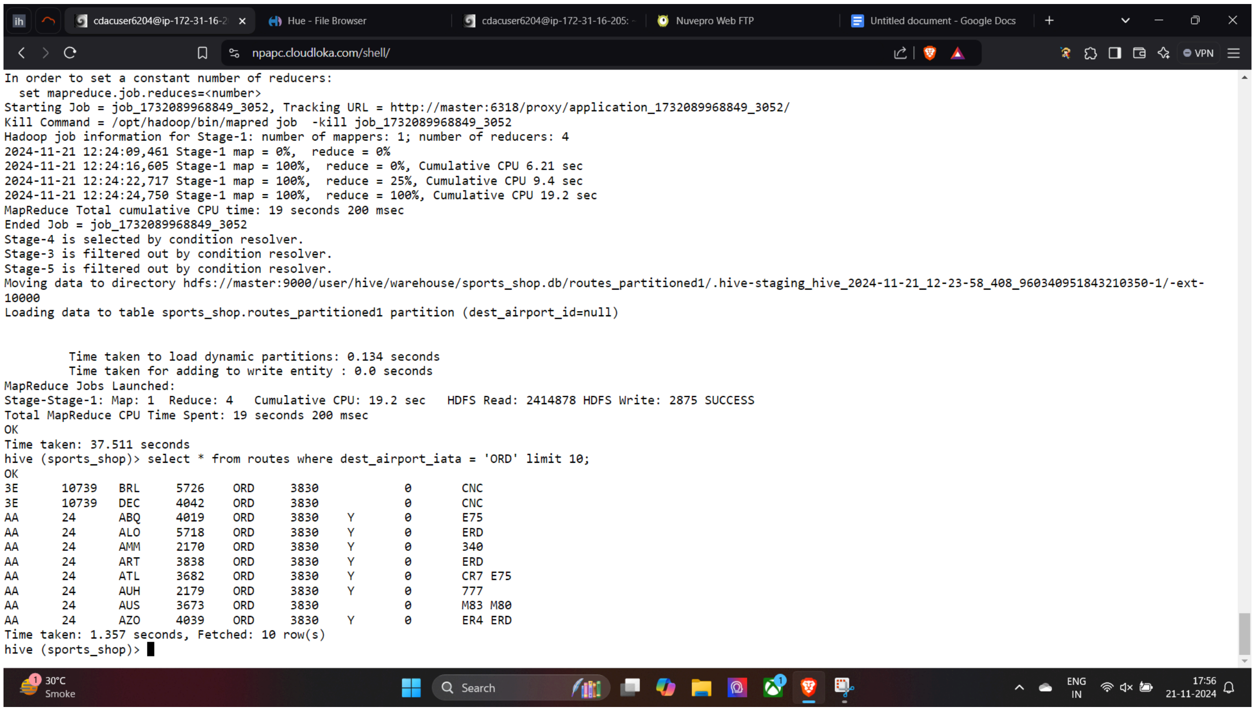
```
1)Another  - create table routes_partitioned1 ( dest_airport_iata
string ) partitioned by (dest_airport_id int) row format delimited
fields terminated by ‘,’ stored as textfile
```

```
2) hive (sports_shop)> insert into routes_partitioned1
partition(dest_airport_id) select dest_airport_iata , dest_airport_id
from routes where dest_airport_iata ='ORD';
```

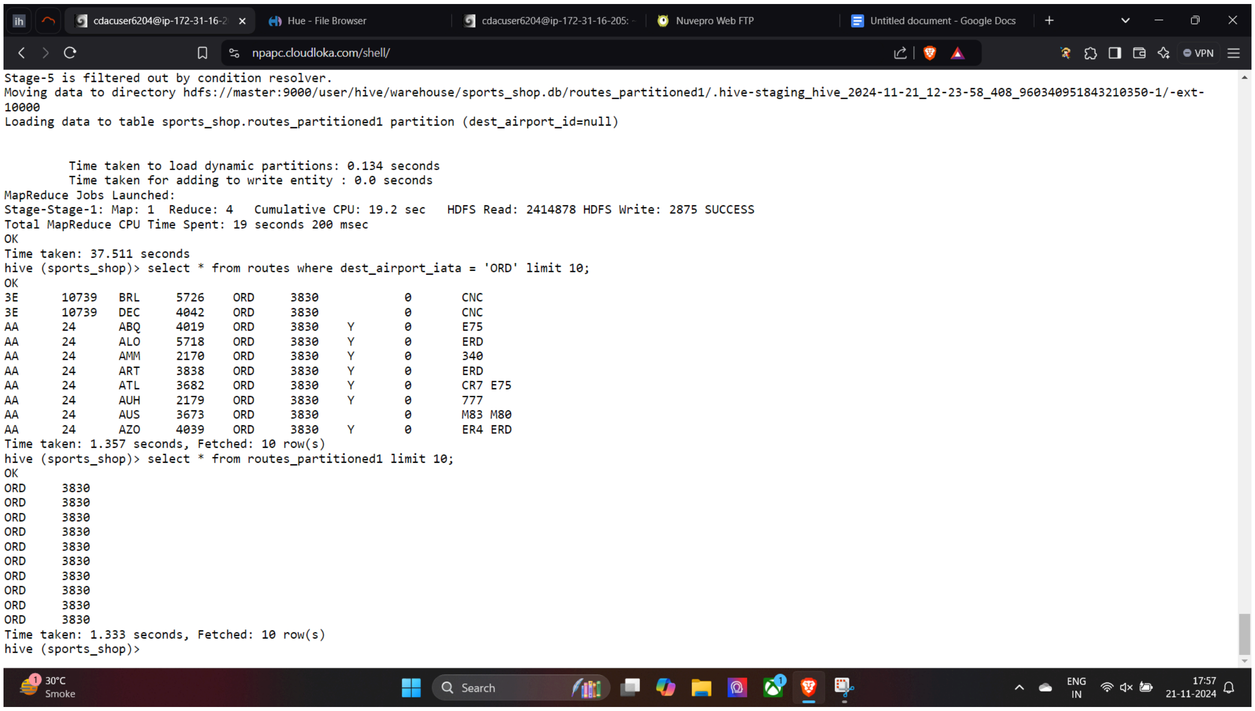

cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205: ~
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
30°C
Smoke
Search
ENG
IN
17:54
21-11-2024

```
FAILED: ParseException line 1:142 mismatched input ',' expecting StringLiteral near 'by' in table row format's field separator
hive (sports_shop)> create table routes_partitioned1 ( dest_airport_iata string ) partitioned by (dest_airport_id int) row format delimited fields terminated by ',' sto
red as textfile;
OK
Time taken: 0.081 seconds
hive (sports_shop)> insert into routes_partitioned1 partition(dest_airport_id) select dest_airport_iata , dest_airport_id from routes where dest_airport_iata ='ORD';
Query ID = cdacuser6204_20241121122358_ee54d72b-c7c6-41dc-9161-c5fd23bfaca9
Total jobs = 3
Launching Job 1 out of 3
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_3052, Tracking URL = http://master:6318/proxy/application_1732089968849_3052/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_3052
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 12:24:09,461 Stage-1 map = 0%,  reduce = 0%
2024-11-21 12:24:16,605 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.21 sec
2024-11-21 12:24:22,717 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 9.4 sec
2024-11-21 12:24:24,750 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.2 sec
MapReduce Total cumulative CPU time: 19 seconds 200 msec
Ended Job = job_1732089968849_3052
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://master:9000/user/hive/warehouse/sports_shop.db/routes_partitioned1/.hive-staging_hive_2024-11-21_12-23-58_408_9603409518432103501-1/-ext-
10000
Loading data to table sports_shop.routes_partitioned1 partition (dest_airport_id=null)


         Time taken to load dynamic partitions: 0.134 seconds
         Time taken for adding to write entity : 0.0 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 19.2 sec   HDFS Read: 2414878 HDFS Write: 2875 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 200 msec
OK
Time taken: 37.511 seconds
hive (sports_shop)>
```

3)


cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205:
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
30°C
Smoke
Search
ENG
IN
17:56
21-11-2024

```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_3052, Tracking URL = http://master:6318/proxy/application_1732089968849_3052/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_3052
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 12:24:09,461 Stage-1 map = 0%,  reduce = 0%
2024-11-21 12:24:16,605 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.21 sec
2024-11-21 12:24:22,717 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 9.4 sec
2024-11-21 12:24:24,750 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.2 sec
MapReduce Total cumulative CPU time: 19 seconds 200 msec
Ended Job = job_1732089968849_3052
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://master:9000/user/hive/warehouse/sports_shop.db/routes_partitioned1/.hive-staging_hive_2024-11-21_12-23-58_408_960340951843210350-1/-ext-
10000
Loading data to table sports_shop.routes_partitioned1 partition (dest_airport_id=null)


         Time taken to load dynamic partitions: 0.134 seconds
         Time taken for adding to write entity : 0.0 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 19.2 sec   HDFS Read: 2414878 HDFS Write: 2875 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 200 msec
OK
Time taken: 37.511 seconds
hive (sports_shop)> select * from routes where dest_airport_iata = 'ORD' limit 10;
OK
3E      10739   BRL     5726    ORD     3830            0       CNC
3E      10739   DEC     4042    ORD     3830            0       CNC
AA      24      ABQ     4019    ORD     3830    Y       0       E75
AA      24      ALO     5718    ORD     3830    Y       0       ERD
AA      24      AMM     2170    ORD     3830    Y       0       340
AA      24      ART     3838    ORD     3830    Y       0       ERD
AA      24      ATL     3682    ORD     3830    Y       0       CR7 E75
AA      24      AUH     2179    ORD     3830    Y       0       777
AA      24      AUS     3673    ORD     3830            0       M83 M80
AA      24      AZO     4039    ORD     3830    Y       0       ER4 ERD
Time taken: 1.357 seconds, Fetched: 10 row(s)
hive (sports_shop)>
```


cdacuser6204@ip-172-31-16-2
Hue - File Browser
cdacuser6204@ip-172-31-16-205:
Nuvepro Web FTP
Untitled document - Google Docs
npapc.cloudloka.com/shell/
VPN
30°C
Smoke
Search
ENG
IN
17:57
21-11-2024

```
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://master:9000/user/hive/warehouse/sports_shop.db/routes_partitioned1/.hive-staging_hive_2024-11-21_12-23-58_408_960340951843210350-1/-ext-
10000
Loading data to table sports_shop.routes_partitioned1 partition (dest_airport_id=null)


         Time taken to load dynamic partitions: 0.134 seconds
         Time taken for adding to write entity : 0.0 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 19.2 sec   HDFS Read: 2414878 HDFS Write: 2875 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 200 msec
OK
Time taken: 37.511 seconds
hive (sports_shop)> select * from routes where dest_airport_iata = 'ORD' limit 10;
OK
3E      10739   BRL     5726    ORD     3830            0       CNC
3E      10739   DEC     4042    ORD     3830            0       CNC
AA      24      ABQ     4019    ORD     3830    Y       0       E75
AA      24      ALO     5718    ORD     3830    Y       0       ERD
AA      24      AMM     2170    ORD     3830    Y       0       340
AA      24      ART     3838    ORD     3830    Y       0       ERD
AA      24      ATL     3682    ORD     3830    Y       0       CR7 E75
AA      24      AUH     2179    ORD     3830    Y       0       777
AA      24      AUS     3673    ORD     3830            0       M83 M80
AA      24      AZO     4039    ORD     3830    Y       0       ER4 ERD
Time taken: 1.357 seconds, Fetched: 10 row(s)
```

```
hive (sports_shop)> select * from routes_partitioned1 limit 10;
OK
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
ORD     3830
Time taken: 1.333 seconds, Fetched: 10 row(s)
hive (sports_shop)>
```