Hive

Q1.2 Determine the aircraft type(Equipment) that is used on the highest number of routes

```
select distinct ai.name from airlines ai join routes r on
ai.airline_id = r.airline_id where r.equipment = 'CR2';
```

```
hive (cdac_amit)> select distinct ai.name from airlines ai join routes r on ai.airline_id = r.airline_id where r.equipment = 'CR2';
Query ID = cdacuser882114_20241121083929_5f65f904-571f-40ae-958d-36012d8af841
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_2240, Tracking URL = http://master:6318/proxy/application_1732089968849_2240/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2240
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 08:39:39,004 Stage-1 map = 0%,  reduce = 0%
2024-11-21 08:39:47,217 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 5.92 sec
2024-11-21 08:39:48,249 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.98 sec
2024-11-21 08:39:54,399 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 22.63 sec
2024-11-21 08:39:56,444 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 26.15 sec
MapReduce Total cumulative CPU time: 26 seconds 150 msec
Ended Job = job_1732089968849_2240
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_2244, Tracking URL = http://master:6318/proxy/application_1732089968849_2244/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2244
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 08:40:09,072 Stage-2 map = 0%,  reduce = 0%
2024-11-21 08:40:14,285 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.63 sec
2024-11-21 08:40:18,397 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.42 sec
2024-11-21 08:40:21,473 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 11.91 sec
2024-11-21 08:40:23,519 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 18.46 sec
MapReduce Total cumulative CPU time: 18 seconds 460 msec
```

```
  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_2244, Tracking URL = http://master:6318/proxy/application_1732089968849_2244/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2244
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 08:40:09,072 Stage-2 map = 0%,  reduce = 0%
2024-11-21 08:40:14,285 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.63 sec
2024-11-21 08:40:18,397 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.42 sec
2024-11-21 08:40:21,473 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 11.91 sec
2024-11-21 08:40:23,519 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 18.46 sec
MapReduce Total cumulative CPU time: 18 seconds 460 msec
Ended Job = job_1732089968849_2244
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 26.15 sec   HDFS Read: 2727302 HDFS Write: 931 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 18.46 sec   HDFS Read: 22177 HDFS Write: 815 SUCCESS
Total MapReduce CPU Time Spent: 44 seconds 610 msec
OK
BRA-Transportes Aereos
Huaxia
Adria Airways
Air China
LOT Polish Airlines
Aerocondor
Aeroflot Russian Airlines
Ciel Canadien
Isles of Scilly Skybus
Yemenia
Hankook Airline
Iberia Airlines
LTU International
Scandinavian Airlines System
Shandong Airlines
South African Airways
Time taken: 55.353 seconds, Fetched: 16 row(s)
hive (cdac_amit)>
```

3.find the airline that operates the highest number of routes and count of those routes

```
select a.name from airlines a join routes r on a.airline_id =
r.airline_id and r.stops > 0 limit 10;
```

```
hive (cdac_amit)> select a.name from airlines a join routes r on a.airline_id = r.airline_id and r.stops > 0 limit 10;
Query ID = cdacuser882114_20241121085717_6e9801d9-ee85-4683-b630-cec35f1ec64e
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_2345, Tracking URL = http://master:6318/proxy/application_1732089968849_2345/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2345
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 08:57:29,259 Stage-1 map = 0%,  reduce = 0%
2024-11-21 08:57:36,436 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 6.41 sec
2024-11-21 08:57:37,463 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 12.62 sec
2024-11-21 08:57:42,581 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 16.28 sec
2024-11-21 08:57:43,604 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 23.1 sec
2024-11-21 08:57:44,629 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 26.77 sec
MapReduce Total cumulative CPU time: 26 seconds 770 msec
Ended Job = job_1732089968849_2345
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 26.77 sec   HDFS Read: 2730864 HDFS Write: 673 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 770 msec
OK
Canadian North
Scandinavian Airlines System
Cubana de Aviación
Air Canada
Air Canada
AirTran Airways
AirTran Airways
AirTran Airways
Southwest Airlines
Southwest Airlines
Time taken: 31.284 seconds, Fetched: 10 row(s)
hive (cdac_amit)>
```

1.find airports that are lisited as both a source and destination in the routes table.

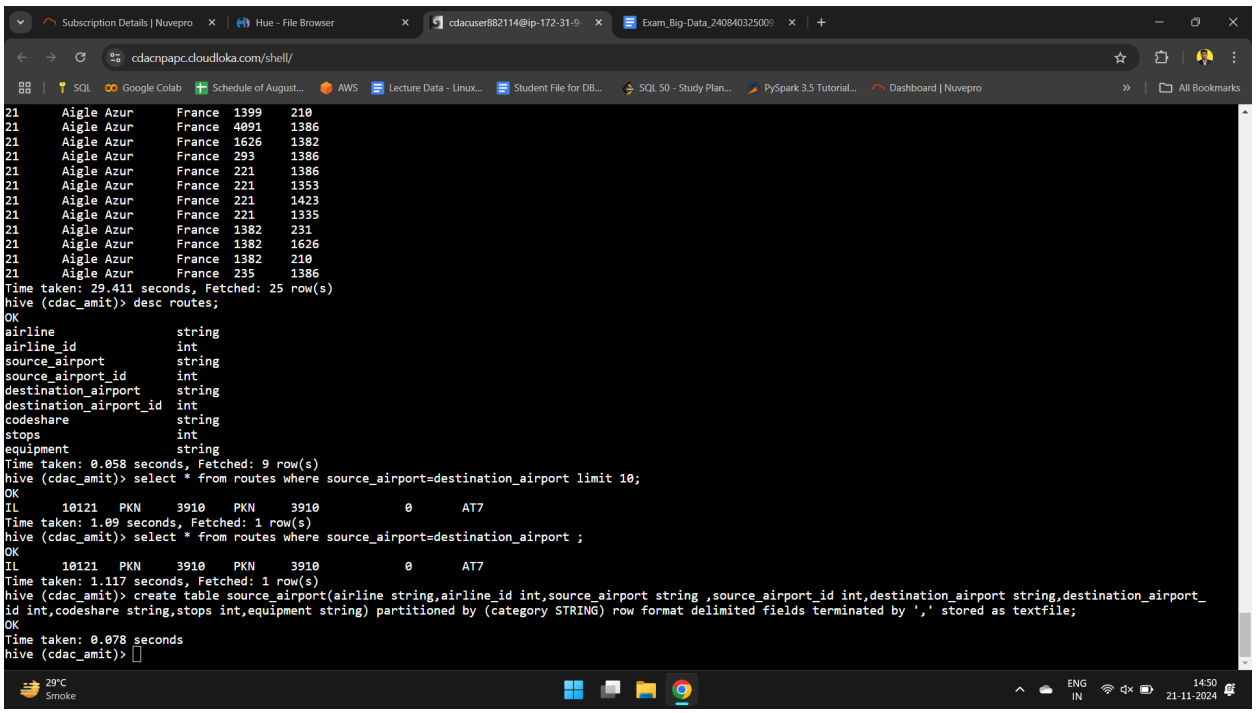
```
select * from routes where source_airport=destination_airport limit
10;
```

```
21      Aigle Azur      France  1423    221
21      Aigle Azur      France  1423    210
21      Aigle Azur      France  1335    6492
21      Aigle Azur      France  1335    231
21      Aigle Azur      France  1335    221
21      Aigle Azur      France  1335    210
21      Aigle Azur      France  1638    1386
21      Aigle Azur      France  1399    231
21      Aigle Azur      France  1399    210
21      Aigle Azur      France  4091    1386
21      Aigle Azur      France  1626    1382
21      Aigle Azur      France  293     1386
21      Aigle Azur      France  221     1386
21      Aigle Azur      France  221     1353
21      Aigle Azur      France  221     1423
21      Aigle Azur      France  221     1335
21      Aigle Azur      France  1382    231
21      Aigle Azur      France  1382    1626
21      Aigle Azur      France  1382    210
21      Aigle Azur      France  235     1386
Time taken: 29.411 seconds, Fetched: 25 row(s)
hive (cdac_amit)> desc routes;
OK
airline                 string
airline_id              int
source_airport          string
source_airport_id       int
destination_airport     string
destination_airport_id  int
codeshare               string
stops                   int
equipment               string
Time taken: 0.058 seconds, Fetched: 9 row(s)
hive (cdac_amit)> select * from routes where source_airport=destination_airport limit 10;
OK
IL      10121   PKN     3910    PKN     3910            0       AT7
Time taken: 1.09 seconds, Fetched: 1 row(s)
hive (cdac_amit)>
```

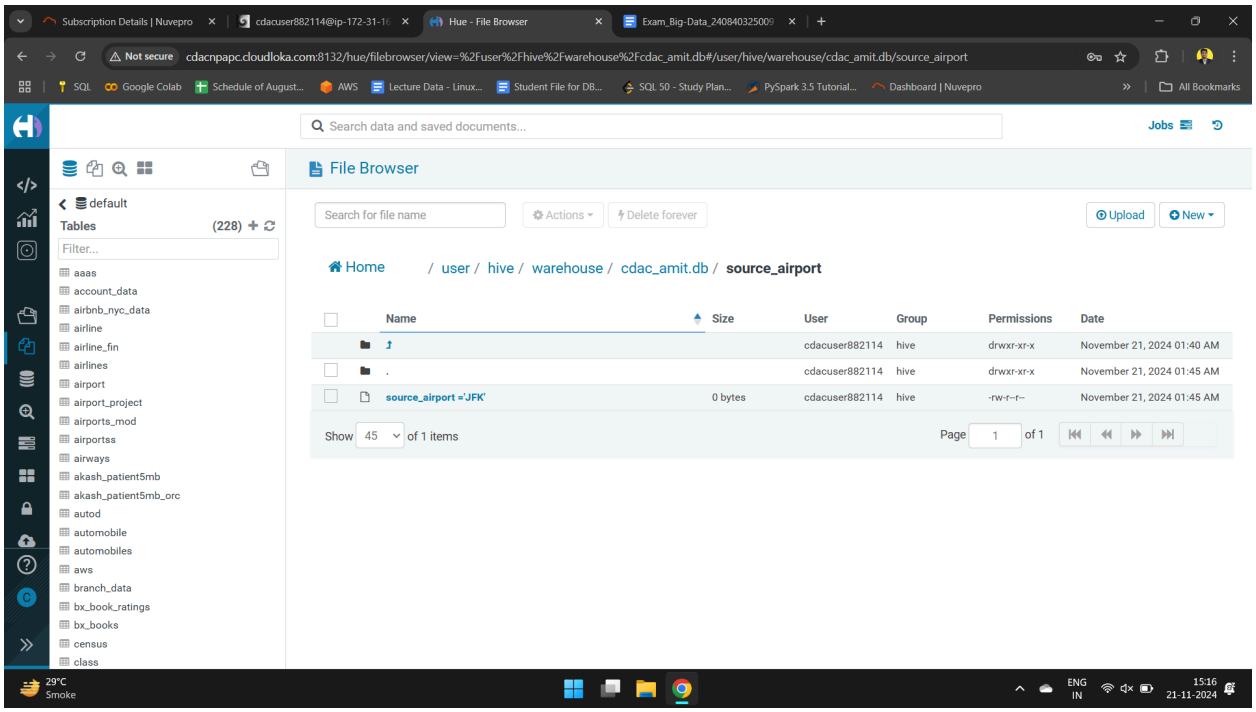
Question 2.

1.create a partitioned table to store the routes data by source_airport .write the SQL query to create this table and insert data into it.

```
create table source_airport(airline string,airline_id
int,source_airport string ,source_airport_id int,destination_airport
string,destination_airport_
id int,codeshare string,stops int,equipment string) partitioned by
(category STRING) row format delimited fields terminated by ',' stored
as textfile;
```

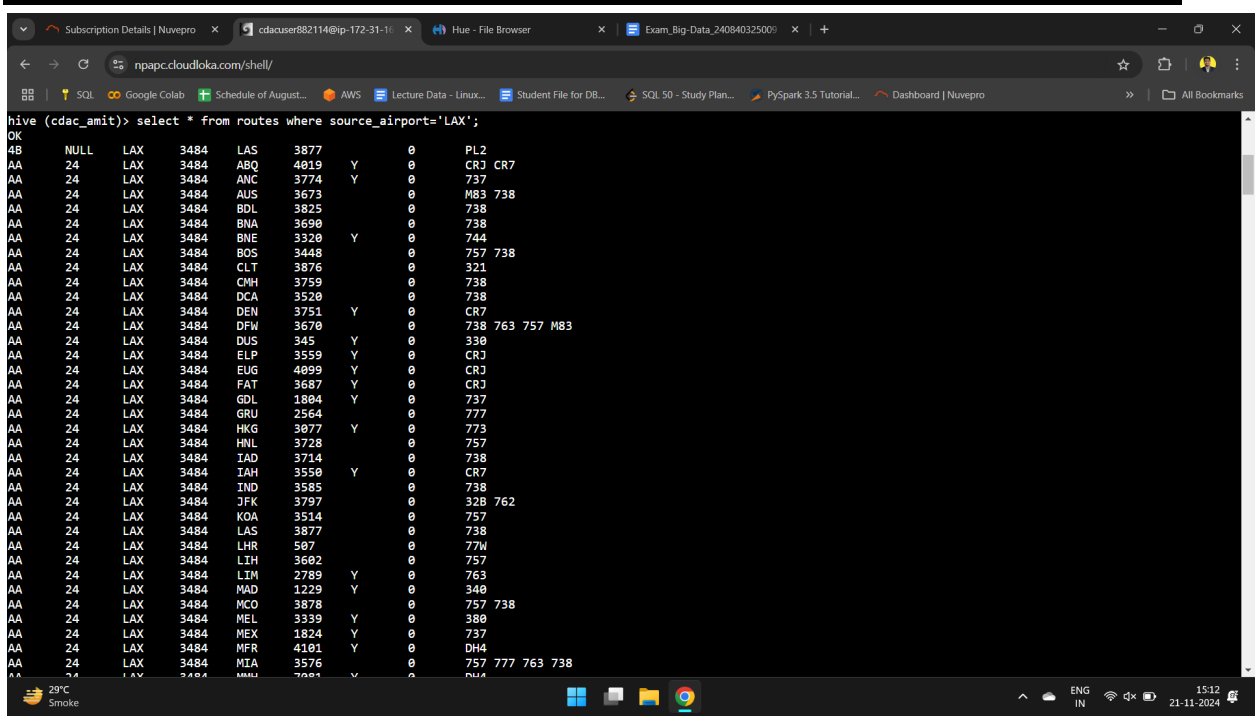


2.


Subscription Details | Nuvepro
cdacuser882114@ip-172-31-16
Hue - File Browser
Exam_Big-Data_240840325009
Not secure
cdacnpapc.cloudloka.com:8132/hue/filebrowser/view=%2Fuser%2Fhive%2Fwarehouse%2Fcdac_amit.db#/user/hive/warehouse/cdac_amit.db/source_airport
SQL
Google Colab
Schedule of August...
AWS
Lecture Data - Linux...
Student File for DB...
SQL 50 - Study Plan...
PySpark 3.5 Tutorial...
Dashboard | Nuvepro
All Bookmarks
Search data and saved documents...
Jobs
File Browser
default
Tables (228)
Filter...
aaas
account_data
airbnb_nyc_data
airline
airline_fin
airlines
airport
airport_project
airports_mod
airportss
airways
akash_patient5mb
akash_patient5mb_orc
autod
automobile
automobiles
aws
branch_data
bx_book_ratings
bx_books
census
class
Search for file name
Actions
Delete forever
Upload
New
Home / user / hive / warehouse / cdac_amit.db / source_airport
Name
Size
User
Group
Permissions
Date
cdacuser882114 hive drwxr-xr-x November 21, 2024 01:40 AM
cdacuser882114 hive drwxr-xr-x November 21, 2024 01:45 AM
source_airport ='JFK'
0 bytes
cdacuser882114 hive -rw-r--r-- November 21, 2024 01:45 AM
Show 45 of 1 items
Page 1 of 1
29°C
Smoke
ENG
IN
15:16
21-11-2024

3.Write a query to retrieve all routes from source_airports=”LAX”;

```
select * from routes where source_airport='LAX';
OK
4B      NULL    LAX     3484    LAS     3877            0       PL2
AA      24      LAX     3484    ABQ     4019    Y       0       CRJ
CR7
AA      24      LAX     3484    ANC     3774    Y       0       737
AA      24      LAX     3484    AUS     3673            0       M83
738
AA      24      LAX     3484    BDL     3825            0       738
AA      24      LAX     3484    BNA     3690            0       738
AA      24      LAX     3484    BNE     3320    Y       0       744
AA      24      LAX     3484    BOS     3448            0       757
738
AA      24      LAX     3484    CLT     3876            0       321
AA      24      LAX     3484    LHR     507             0       77W
AA      24      LAX     3484    LIH     3602            0       757
AA      24      LAX     3484    LIM     2789    Y       0       763
```



SPARK

Q1.1

```
        at org.apache.spark.InterruptibleIterator.to(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toBuffer(TraversableOnce.scala:307)
        at scala.collection.TraversableOnce.toBuffer$(TraversableOnce.scala:307)
        at org.apache.spark.InterruptibleIterator.toBuffer(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toArray(TraversableOnce.scala:294)
        at scala.collection.TraversableOnce.toArray$(TraversableOnce.scala:288)
        at org.apache.spark.InterruptibleIterator.toArray(InterruptibleIterator.scala:28)
        at org.apache.spark.rdd.RDD.$anonfun$collect$2(RDD.scala:1030)
        at org.apache.spark.SparkContext.$anonfun$runJob$5(SparkContext.scala:2236)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:131)
        at org.apache.spark.executor.Executor$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

>>> airlines=sc.textFile("/user/cdacuser882114/airlines.csv")
>>> airlines.min().max()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'str' object has no attribute 'max'
>>> airlines.min()
'1995,1,296.9,46561'
>>> airlines.average()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'RDD' object has no attribute 'average'
>>> airlines.max()
'Year,Quarter,Avg_rev_per_seat,booked_seats'
>>> airlines[3].max()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'RDD' object is not subscriptable
>>> airlines.count()
85
>>>
```

Q2.

1.

```
        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.foreach(Iterator.scala:941)
        at scala.collection.Iterator.foreach$(Iterator.scala:941)
        at org.apache.spark.InterruptibleIterator.foreach(InterruptibleIterator.scala:28)
        at scala.collection.generic.Growable.$plus$plus$eq(Growable.scala:62)
        at scala.collection.generic.Growable.$plus$plus$eq$(Growable.scala:53)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:105)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:49)
        at scala.collection.TraversableOnce.to(TraversableOnce.scala:315)
        at scala.collection.TraversableOnce.to$(TraversableOnce.scala:313)
        at org.apache.spark.InterruptibleIterator.to(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toBuffer(TraversableOnce.scala:307)
        at scala.collection.TraversableOnce.toBuffer$(TraversableOnce.scala:307)
        at org.apache.spark.InterruptibleIterator.toBuffer(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toArray(TraversableOnce.scala:294)
        at scala.collection.TraversableOnce.toArray$(TraversableOnce.scala:288)
        at org.apache.spark.InterruptibleIterator.toArray(InterruptibleIterator.scala:28)
        at org.apache.spark.rdd.RDD.$anonfun$collect$2(RDD.scala:1030)
        at org.apache.spark.SparkContext.$anonfun$runJob$5(SparkContext.scala:2236)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:131)
        at org.apache.spark.executor.Executor$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

>>> airlines=sc.textFile("/user/cdacuser882114/airlines.csv")
>>> airlines.min().max()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'str' object has no attribute 'max'
>>> airlines.min()
'1995,1,296.9,46561'
>>>
```

2.count

```
        at org.apache.spark.InterruptibleIterator.to(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toBuffer(TraversableOnce.scala:307)
        at scala.collection.TraversableOnce.toBuffer$(TraversableOnce.scala:307)
        at org.apache.spark.InterruptibleIterator.toBuffer(InterruptibleIterator.scala:28)
        at scala.collection.TraversableOnce.toArray(TraversableOnce.scala:294)
        at scala.collection.TraversableOnce.toArray$(TraversableOnce.scala:288)
        at org.apache.spark.InterruptibleIterator.toArray(InterruptibleIterator.scala:28)
        at org.apache.spark.rdd.RDD.$anonfun$collect$2(RDD.scala:1030)
        at org.apache.spark.SparkContext.$anonfun$runJob$5(SparkContext.scala:2236)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:131)
        at org.apache.spark.executor.Executor$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

>>> airlines=sc.textFile("/user/cdacuser882114/airlines.csv")
>>> airlines.min().max()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'str' object has no attribute 'max'
>>> airlines.min()
'1995,1,296.9,46561'
>>> airlines.average()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'RDD' object has no attribute 'average'
>>> airlines.max()
'Year,Quarter,Avg_rev_per_seat,booked_seats'
>>> airlines[3].max()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'RDD' object is not subscriptable
>>> airlines.count()
85
>>>
```