

BIG DATA MODULE END EXAM

ANIKET KARAD

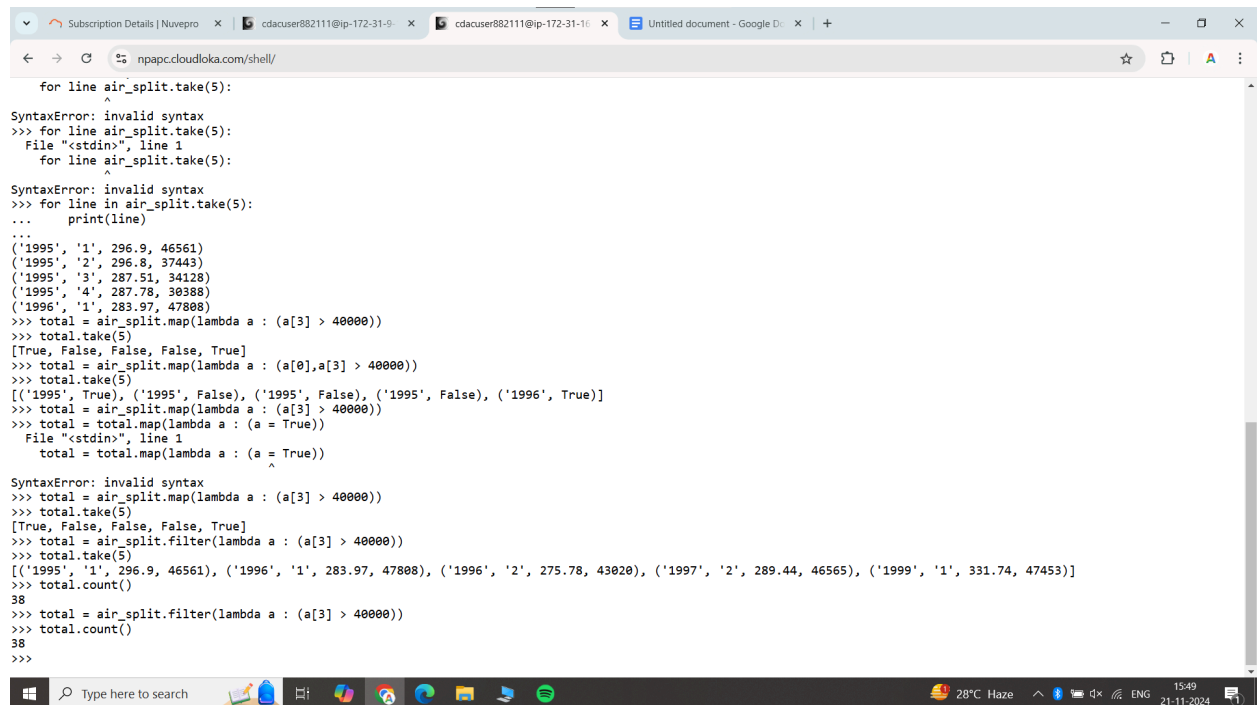
240840325012

SPARK

Q1

Ans

```
total = air_split.filter(lambda a : (a[3] > 40000))
total.count()
```



```
Subscription Details | Nuvepro x cdacuser882111@ip-172-31-9- x cdacuser882111@ip-172-31-16- x Untitled document - Google D- x +
npacc.cloudloka.com/shell/

for line air_split.take(5):
^
SyntaxError: invalid syntax
>>> for line air_split.take(5):
File "<stdin>", line 1
    for line air_split.take(5):
    ^
SyntaxError: invalid syntax
>>> for line in air_split.take(5):
...     print(line)
...
('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)
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total.take(5)
[True, False, False, False, True]
>>> total = air_split.map(lambda a : (a[0], a[3] > 40000))
>>> total.take(5)
[('1995', True), ('1995', False), ('1995', False), ('1995', False), ('1996', True)]
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total = total.map(lambda a : (a = True))
File "<stdin>", line 1
    total = total.map(lambda a : (a = True))
    ^
SyntaxError: invalid syntax
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total.take(5)
[True, False, False, False, True]
>>> total = air_split.filter(lambda a : (a[3] > 40000))
>>> total.take(5)
[('1995', '1', 296.9, 46561), ('1996', '1', 283.97, 47808), ('1996', '2', 275.78, 43020), ('1997', '2', 289.44, 46565), ('1999', '1', 331.74, 47453)]
>>> total.count()
38
>>> total = air_split.filter(lambda a : (a[3] > 40000))
>>> total.count()
38
>>>
```

Q2

1.

Ans

```
min = air_split.map(lambda a : a[2]).min()
max = air_split.map(lambda a : a[2]).max()
avg = air_split.map(lambda a : a[2]).mean()
```

```
Subscription Details | Nuvepro x cdacuser882111@ip-172-31-9- x Untitled document - Google D... x +
cdacnpac.cloudloka.com/shell/

>>> airline.count()
85
>>> header = airline.first()
>>> airline_clean = airline.filter(lambda a : a != header)
>>> airline_clean.count()
84
>>> air_split = airline_clean.map(lambda a : (a.split(",")[0],a.split(",")[1],float(a.split(",")[2]),int(a.split(",")[3])))
>>> for line in air_split.take(10):
...     print(line)
...
('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)
>>> min = air_split.map(lambda a : a[2]).min()
>>> min
269.49
>>> min = air_split.map(lambda a : a[2]).min().max().mean()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'float' object has no attribute 'max'
>>> min = air_split.map(lambda a : a[2]).max()
>>> min
396.37
>>> min = air_split.map(lambda a : a[2]).min()
>>> min
269.49
>>> max = air_split.map(lambda a : a[2]).max()
>>> max
396.37
>>> avg = air_split.map(lambda a : a[2]).mean()
>>> avg
329.7475
>>>
```

2.

ANS

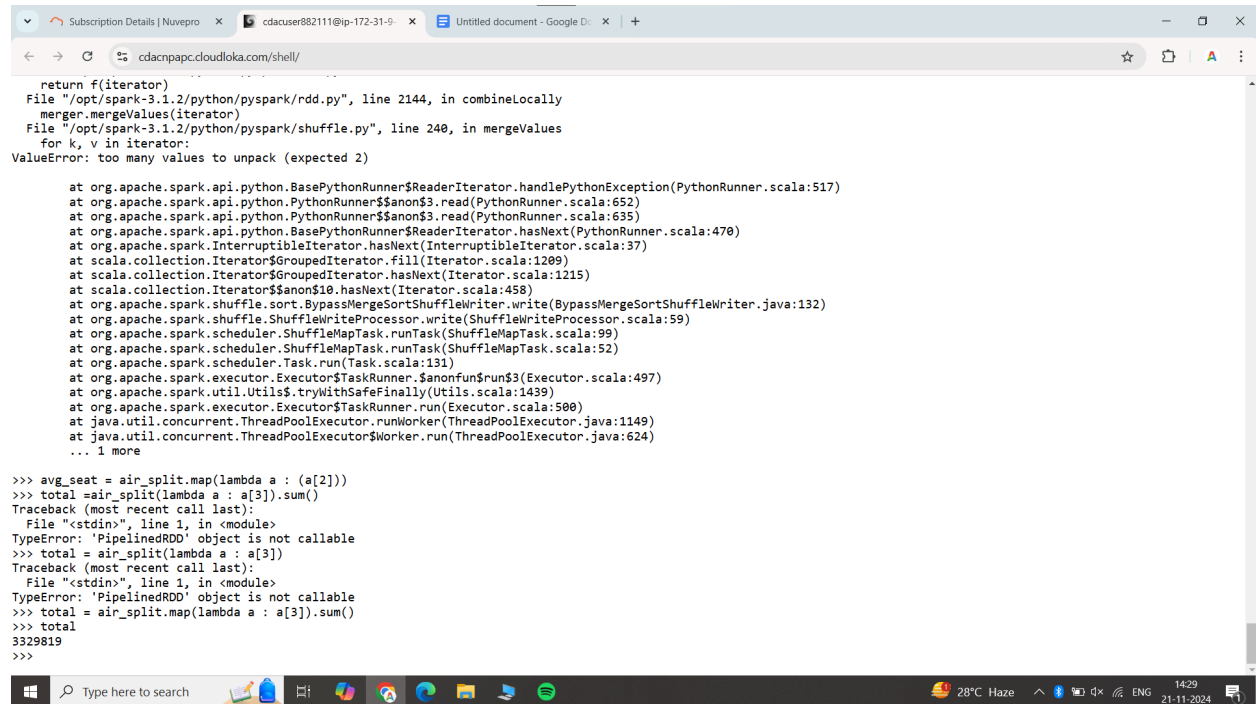
```
total = air_split.filter(lambda a : (a[2] > 290))
total.count()
```

```
Subscription Details | Nuvepro x cdacuser882111@ip-172-31-9- x cdacuser882111@ip-172-31-16-205: ~ - Shell In A Box npapc.cloudloka.com
npapc.cloudloka.com/shell/
File "<stdin>", line 1
for line in air_split.take(5):
^
SyntaxError: invalid syntax
>>> for line in air_split.take(5):
...     print(line)
...
('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)
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total.take(5)
[True, False, False, False, True]
>>> total = air_split.map(lambda a : (a[0],a[3] > 40000))
>>> total.take(5)
[('1995', True), ('1995', False), ('1995', False), ('1995', False), ('1996', True)]
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total = total.map(lambda a : (a = True))
File "<stdin>", line 1
total = total.map(lambda a : (a = True))
^
SyntaxError: invalid syntax
>>> total = air_split.map(lambda a : (a[3] > 40000))
>>> total.take(5)
[True, False, False, False, True]
>>> total = air_split.filter(lambda a : (a[3] > 40000))
>>> total.take(5)
[('1995', '1', 296.9, 46561), ('1996', '1', 283.97, 47808), ('1996', '2', 275.78, 43020), ('1997', '2', 289.44, 46565), ('1999', '1', 331.74, 47453)]
>>> total.count()
38
>>> total = air_split.filter(lambda a : (a[3] > 40000))
>>> total.count()
38
>>> = air_sp = air_split.filter(lambda a : (a[2] > 290))
>>> total = air_split.filter(lambda a : (a[2] > 290))
>>> total.count()
75
>>>
```

3.

Ans

```
total = air_split.map(lambda a : a[3]).sum()
(This for total no. of seats booked combining all quarters in
dataset)
```



```

return f(iterator)
File "/opt/spark-3.1.2/python/pyspark/rdd.py", line 2144, in combineLocally
  merger.mergeValues(iterator)
File "/opt/spark-3.1.2/python/pyspark/shuffle.py", line 240, in mergeValues
  for k, v in iterator:
ValueError: too many values to unpack (expected 2)

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

>>> avg_seat = air_split.map(lambda a : (a[2]))
>>> total = air_split(lambda a : a[3]).sum()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'PipelinedRDD' object is not callable
>>> total = air_split(lambda a : a[3])
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'PipelinedRDD' object is not callable
>>> total = air_split.map(lambda a : a[3]).sum()
>>> total
3329819
>>>

```

```
Total_book = air_split.map(lambda a : (a[0],a[3]))
Total_seat = total_book.reduceByKey(lambda a,b : a+b)
total _seat.take(5)
```

```
Subscription Details | Nuvepro x cdacuser882111@ip-172-31-9- x Untitled document - Google D... x +
cdacnpac.cloudloka.com/shell/

return func(split, prev_func(split, iterator))
File "/opt/spark-3.1.2/python/pyspark/rdd.py", line 2916, in pipeline_func
return func(split, prev_func(split, iterator))
File "/opt/spark-3.1.2/python/pyspark/rdd.py", line 418, in func
return f(iterator)
File "/opt/spark-3.1.2/python/pyspark/rdd.py", line 2144, in combineLocally
merger.mergeValues(iterator)
File "/opt/spark-3.1.2/python/pyspark/shuffle.py", line 240, in mergeValues
for k, v in iterator:
TypeError: cannot unpack non-iterable int object

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

>>> total_seat
PythonRDD[32] at RDD at PythonRDD.scala:53
>>> total_book = air_split.map(lambda a : (a[0],a[3]))
>>> total_book.take(5)
[('1995', 46561), ('1995', 37443), ('1995', 34128), ('1995', 30388), ('1996', 47808)]
>>> total_seat = total_book.reduceByKey(lambda a,b : a+b)
>>> total_seat.take(5)
[('1995', 148520), ('2002', 152195), ('2003', 156153), ('2004', 164800), ('2007', 176299)]
>>>
```

5.

Ans

```
total_rev = air_split.map(lambda a : (a[0],a[2]*a[3]))
total_rev_red = total_rev.reduceByKey(lambda a,b : a+b)
total_rev_red.take(5)
```

```
Subscription Details | Nuvepro x cdacuser882111@ip-172-31-9- x Untitled document - Google D... x +
cdacnpapc.cloudloka.com/shell/

...
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

>>> total_rev = air_split.map(lambda a : (a[0],a[2]*a[3]))
>>> total_rev.take(5)
[('1995', 13823960.899999999), ('1995', 11113082.4), ('1995', 9812141.28), ('1995', 8745058.639999999), ('1996', 13576037.760000002)]
>>> total_rev_red = total_rev.reduceByKey(lambda a,b : a+b)
>>> total_rev_red.take(5)
[('1995', 43494243.22), ('2002', 47499146.5), ('2003', 49273210.83), ('2004', 50631364.949999996), ('2007', 57309216.07)]
>>> for line in total_rev_red.take(10):
...     print(line)
...
('1995', 43494243.22)
('2002', 47499146.5)
('2003', 49273210.83)
('2004', 50631364.949999996)
('2007', 57309216.07)
('2010', 54861521.29)
('2011', 51888286.22)
('2012', 62199127.28)
('2013', 66363208.71)
('2014', 62624175.85000001)
>>>
```