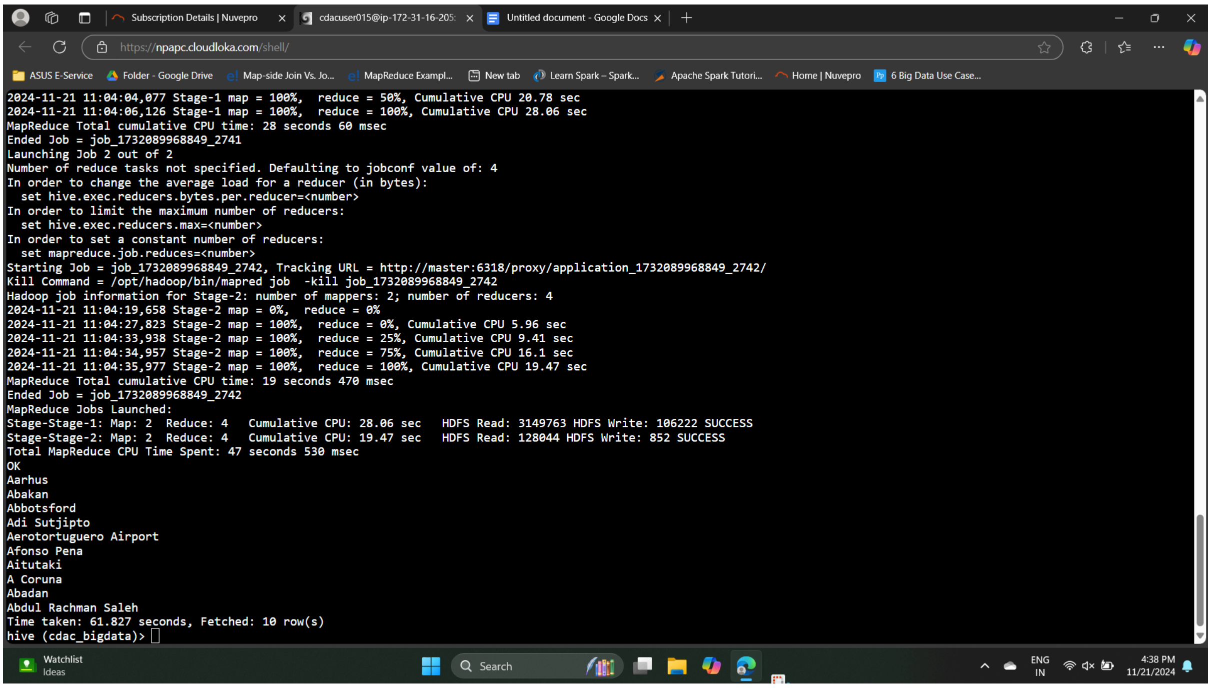
Q1.

1.

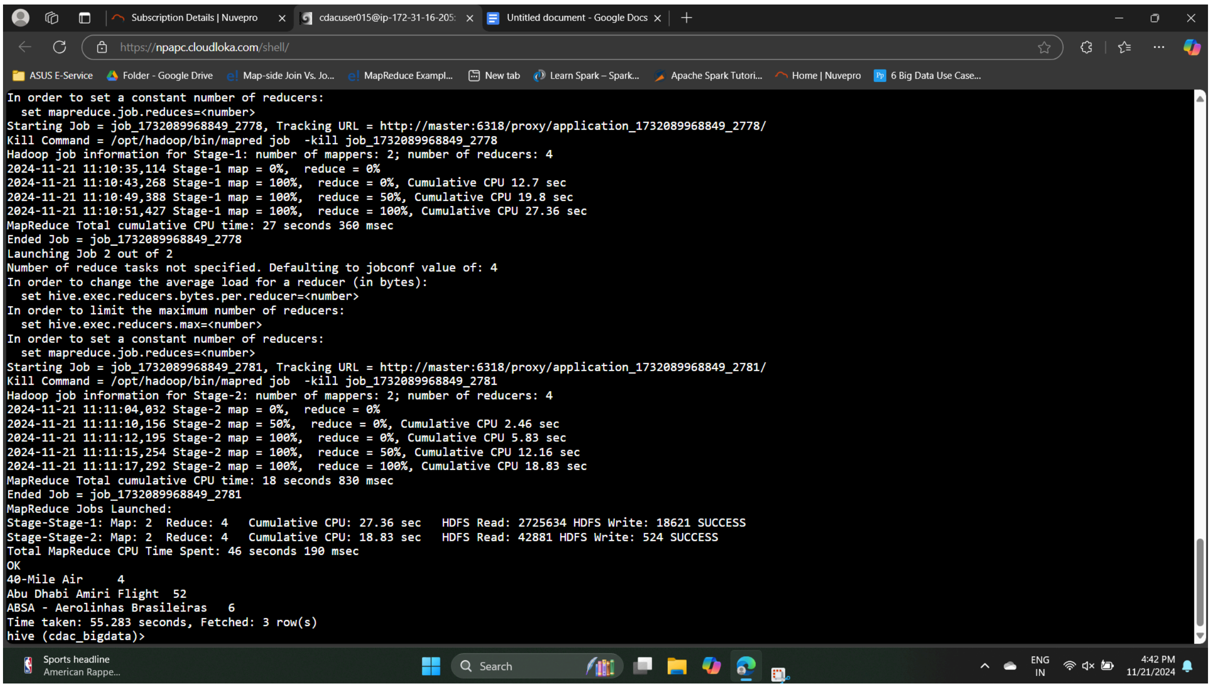
```
  Select a.name from airports a
  join routes r
  On a.airport_id = r.airport_id
Limit 0;
```



```
2024-11-21 11:04:04,077 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 20.78 sec
2024-11-21 11:04:06,126 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.06 sec
MapReduce Total cumulative CPU time: 28 seconds 60 msec
Ended Job = job_1732089968849_2741
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_2742, Tracking URL = http://master:6318/proxy/application_1732089968849_2742/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2742
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 11:04:19,658 Stage-2 map = 0%,  reduce = 0%
2024-11-21 11:04:27,823 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.96 sec
2024-11-21 11:04:33,938 Stage-2 map = 100%,  reduce = 25%, Cumulative CPU 9.41 sec
2024-11-21 11:04:34,957 Stage-2 map = 100%,  reduce = 75%, Cumulative CPU 16.1 sec
2024-11-21 11:04:35,977 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 19.47 sec
MapReduce Total cumulative CPU time: 19 seconds 470 msec
Ended Job = job_1732089968849_2742
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 28.06 sec   HDFS Read: 3149763 HDFS Write: 106222 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 19.47 sec   HDFS Read: 128044 HDFS Write: 852 SUCCESS
Total MapReduce CPU Time Spent: 47 seconds 530 msec
OK
Aarhus
Abakan
Abbotsford
Adi Sutjipto
Aerotortuguero Airport
Afonso Pena
Aitutaki
A Coruna
Abadan
Abdul Rachman Saleh
Time taken: 61.827 seconds, Fetched: 10 row(s)
hive (cdac_bigdata)> 
```

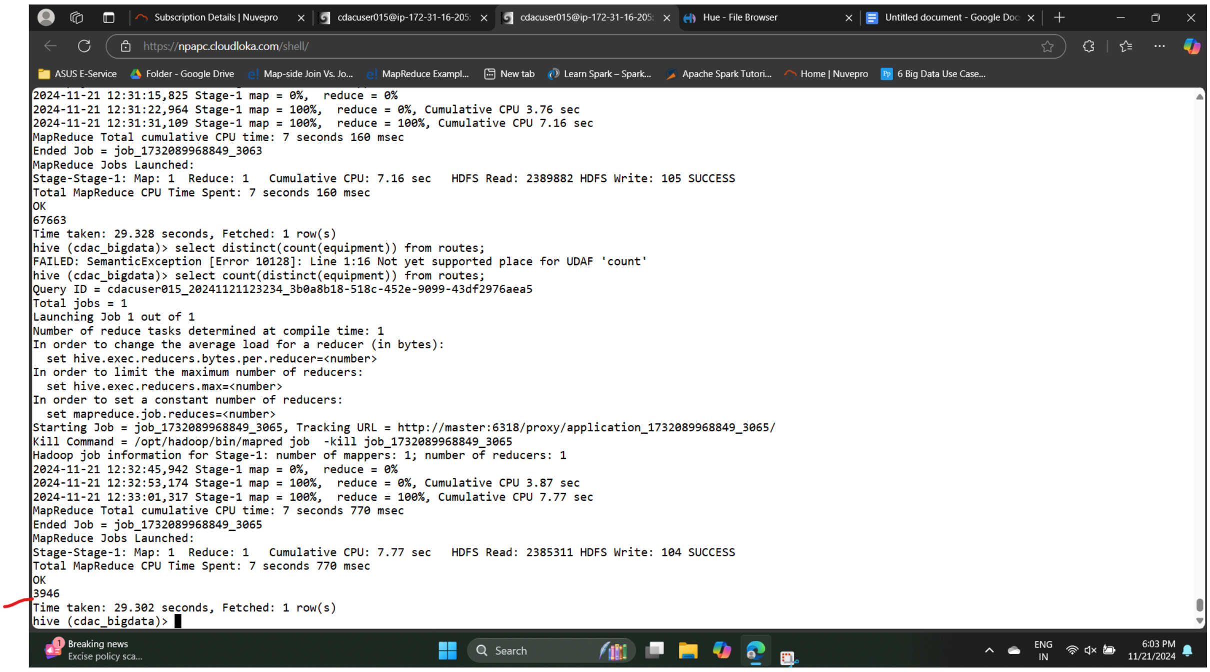
2.

```
select a.name ,count(*) from airlines a join routes r on
r.airline_id = a.airline_id  group by name limit 3;
```



```
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2778, Tracking URL = http://master:6318/proxy/application_1732089968849_2778/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2778
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 11:10:35,114 Stage-1 map = 0%,  reduce = 0%
2024-11-21 11:10:43,268 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 12.7 sec
2024-11-21 11:10:49,388 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 19.8 sec
2024-11-21 11:10:51,427 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.36 sec
MapReduce Total cumulative CPU time: 27 seconds 360 msec
Ended Job = job_1732089968849_2778
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_2781, Tracking URL = http://master:6318/proxy/application_1732089968849_2781/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2781
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 11:11:04,032 Stage-2 map = 0%,  reduce = 0%
2024-11-21 11:11:10,156 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.46 sec
2024-11-21 11:11:12,195 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.83 sec
2024-11-21 11:11:15,254 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 12.16 sec
2024-11-21 11:11:17,292 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 18.83 sec
MapReduce Total cumulative CPU time: 18 seconds 830 msec
Ended Job = job_1732089968849_2781
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 27.36 sec   HDFS Read: 2725634 HDFS Write: 18621 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 18.83 sec   HDFS Read: 42881 HDFS Write: 524 SUCCESS
Total MapReduce CPU Time Spent: 46 seconds 190 msec
OK
40-Mile Air     4
Abu Dhabi Amiri Flight  52
ABSA - Aerolinhas Brasileiras   6
Time taken: 55.283 seconds, Fetched: 3 row(s)
hive (cdac_bigdata)>
```

Q3. select count(distinct(equipment)) from routes;

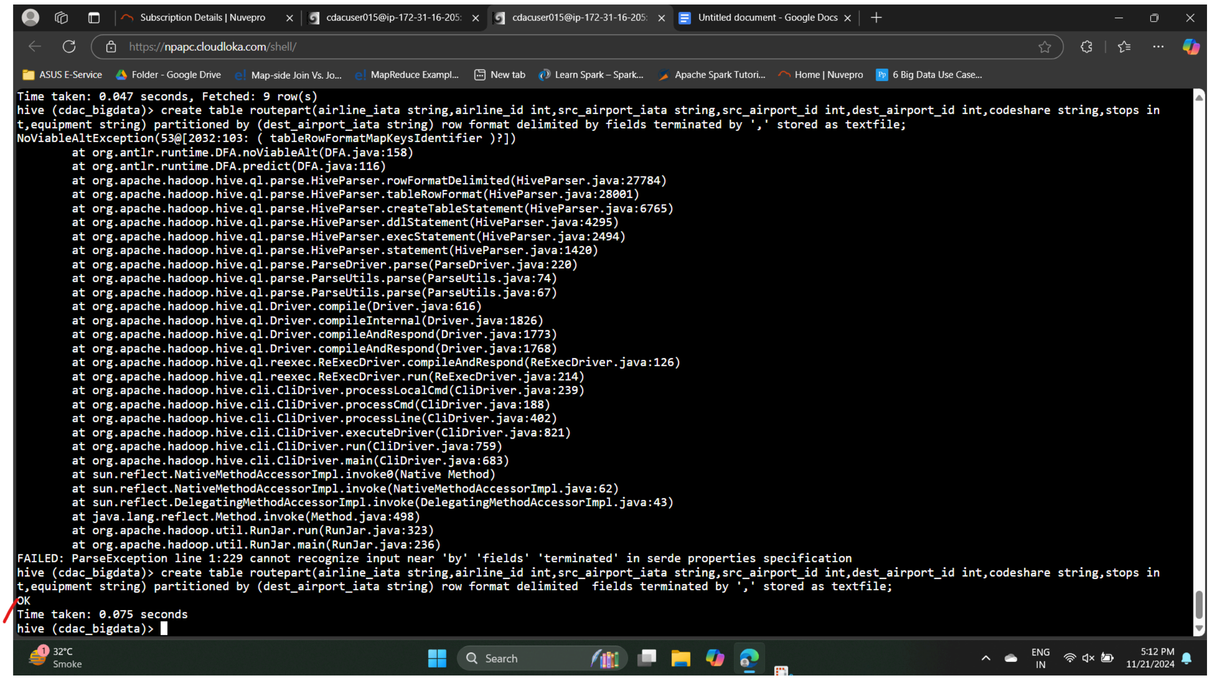


```
2024-11-21 12:31:15,825 Stage-1 map = 0%,  reduce = 0%
2024-11-21 12:31:22,964 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.76 sec
2024-11-21 12:31:31,109 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.16 sec
MapReduce Total cumulative CPU time: 7 seconds 160 msec
Ended Job = job_1732089968849_3063
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.16 sec   HDFS Read: 2389882 HDFS Write: 105 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 160 msec
OK
67663
Time taken: 29.328 seconds, Fetched: 1 row(s)
hive (cdac_bigdata)> select distinct(count(equipment)) from routes;
FAILED: SemanticException [Error 10128]: Line 1:16 Not yet supported place for UDAF 'count'
hive (cdac_bigdata)> select count(distinct(equipment)) from routes;
Query ID = cdacuser015_20241121123234_3b0a8b18-518c-452e-9099-43df2976aea5
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_3065, Tracking URL = http://master:6318/proxy/application_1732089968849_3065/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_3065
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:32:45,942 Stage-1 map = 0%,  reduce = 0%
2024-11-21 12:32:53,174 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.87 sec
2024-11-21 12:33:01,317 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.77 sec
MapReduce Total cumulative CPU time: 7 seconds 770 msec
Ended Job = job_1732089968849_3065
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.77 sec   HDFS Read: 2385311 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 770 msec
OK
3946
Time taken: 29.302 seconds, Fetched: 1 row(s)
hive (cdac_bigdata)>
```

Q2.

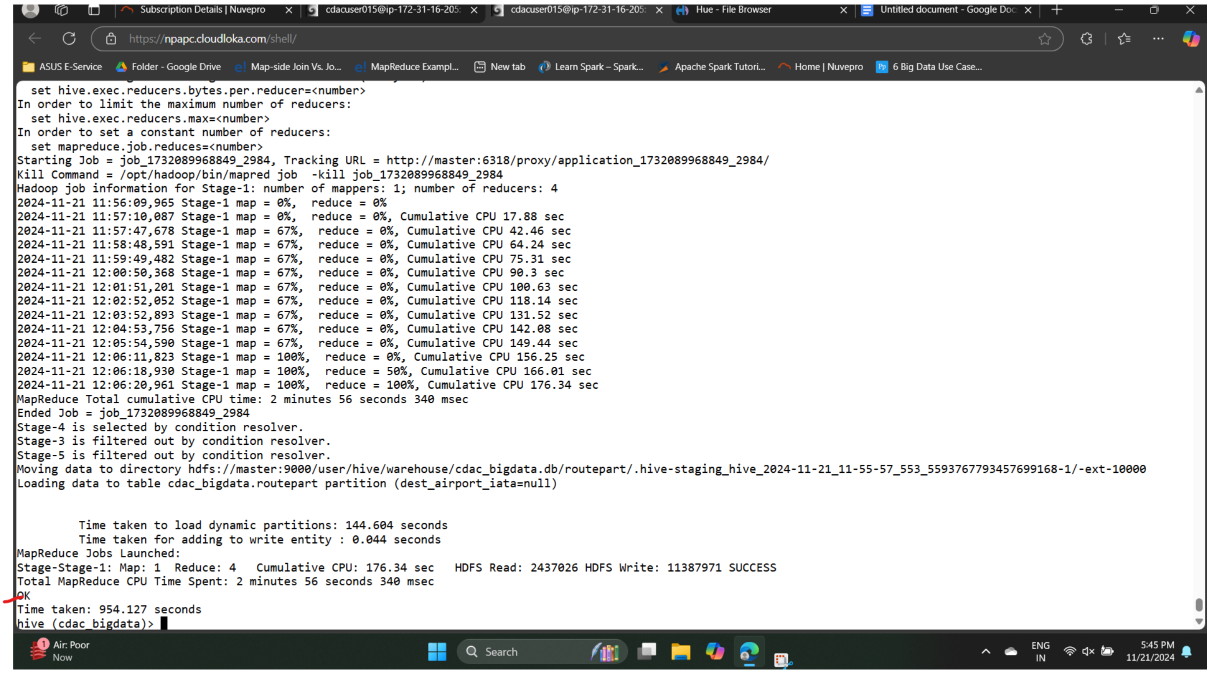
1.

create table routepart(airline_iata string,airline_id int,src_airport_iata string,src_airport_id int,dest_airport_id int,codeshare string,stops int,equipment string) partitioned by (dest_airport_iata string) row format delimited fields terminated by ',' stored as textfile;



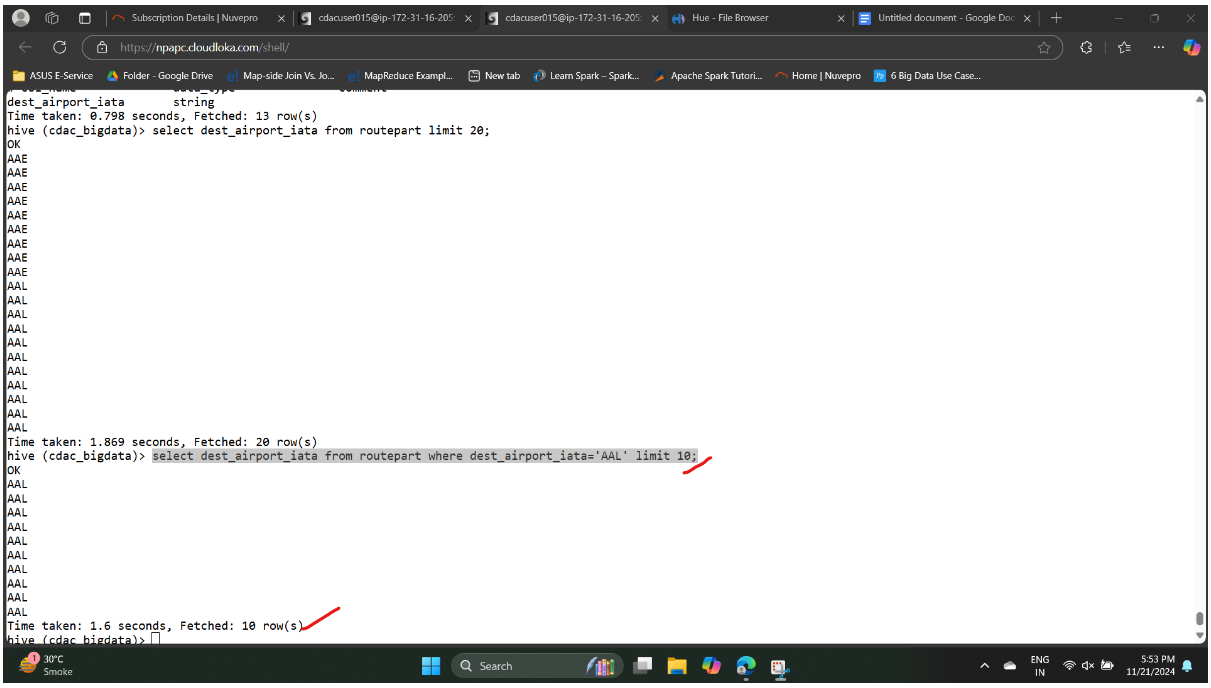
```
Time taken: 0.047 seconds, Fetched: 9 row(s)
hive (cdac_bigdata)> create table routepart(airline_iata string,airline_id int,src_airport_iata string,src_airport_id int,dest_airport_id int,codeshare string,stops in
t,equipment string) partitioned by (dest_airport_iata string) row format delimited by fields terminated by ',' stored as textfile;
NoViableAltException(53@[2032:103: ( tableRowFormatMapKeysIdentifier )?])
        at org.antlr.runtime.DFA.noViableAlt(DFA.java:158)
        at org.antlr.runtime.DFA.predict(DFA.java:116)
        at org.apache.hadoop.hive.ql.parse.HiveParser.rowFormatDelimited(HiveParser.java:27784)
        at org.apache.hadoop.hive.ql.parse.HiveParser.tableRowFormat(HiveParser.java:28001)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:6765)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:4295)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2494)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1420)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:220)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:229 cannot recognize input near 'by' 'fields' 'terminated' in serde properties specification
hive (cdac_bigdata)> create table routepart(airline_iata string,airline_id int,src_airport_iata string,src_airport_id int,dest_airport_id int,codeshare string,stops in
t,equipment string) partitioned by (dest_airport_iata string) row format delimited  fields terminated by ',' stored as textfile;
OK
Time taken: 0.075 seconds
hive (cdac_bigdata)>
```

```
2.insert overwrite table routepart partition(dest_airport_iata)
select
airline_iata,airline_id,src_airport_iata,src_airport_id,dest_airpo
rt_id,codes
hare,stops,equipment,dest_airport_iata from routes;
```



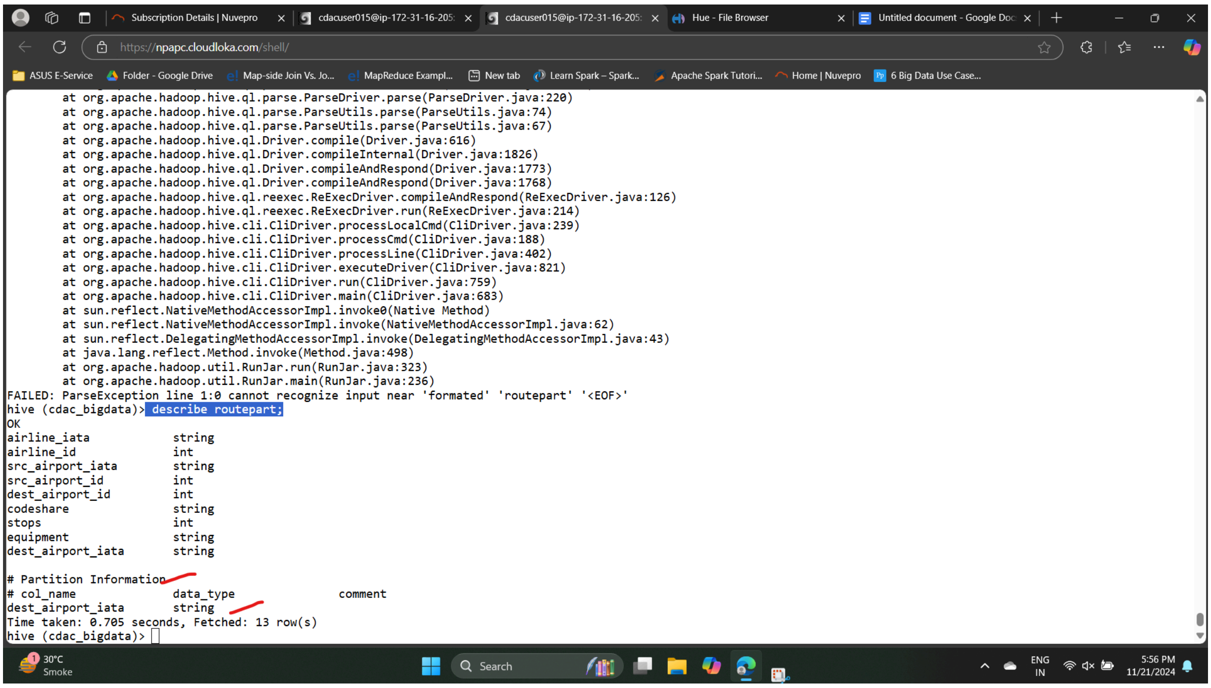
```
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2984, Tracking URL = http://master:6318/proxy/application_1732089968849_2984/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2984
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 11:56:09,965 Stage-1 map = 0%,  reduce = 0%
2024-11-21 11:57:10,087 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 17.88 sec
2024-11-21 11:57:47,678 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 42.46 sec
2024-11-21 11:58:48,591 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 64.24 sec
2024-11-21 11:59:49,482 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 75.31 sec
2024-11-21 12:00:50,368 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 90.3 sec
2024-11-21 12:01:51,201 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 100.63 sec
2024-11-21 12:02:52,052 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 118.14 sec
2024-11-21 12:03:52,893 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 131.52 sec
2024-11-21 12:04:53,756 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 142.08 sec
2024-11-21 12:05:54,590 Stage-1 map = 67%,  reduce = 0%, Cumulative CPU 149.44 sec
2024-11-21 12:06:11,823 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 156.25 sec
2024-11-21 12:06:18,930 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 166.01 sec
2024-11-21 12:06:20,961 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 176.34 sec
MapReduce Total cumulative CPU time: 2 minutes 56 seconds 340 msec
Ended Job = job_1732089968849_2984
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/cdac_bigdata.db/routepart/.hive-staging_hive_2024-11-21_11-55-57_553_5593767793457699168-1/-ext-10000
Loading data to table cdac_bigdata.routepart partition (dest_airport_iata=null)


         Time taken to load dynamic partitions: 144.604 seconds
         Time taken for adding to write entity : 0.044 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 176.34 sec   HDFS Read: 2437026 HDFS Write: 11387971 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 56 seconds 340 msec
OK
Time taken: 954.127 seconds
hive (cdac_bigdata)>
```

3.select dest_airport_iata from routepart where dest_airport_iata='AAL' limit 10;



```
# col_name              data_type               comment
dest_airport_iata       string
Time taken: 0.798 seconds, Fetched: 13 row(s)
hive (cdac_bigdata)> select dest_airport_iata from routepart limit 20;
OK
AAE
AAE
AAE
AAE
AAE
AAE
AAE
AAE
AAE
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
Time taken: 1.869 seconds, Fetched: 20 row(s)
hive (cdac_bigdata)> select dest_airport_iata from routepart where dest_airport_iata='AAL' limit 10;
OK
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
AAL
Time taken: 1.6 seconds, Fetched: 10 row(s)
hive (cdac_bigdata)> 
```

4. describe routepart;

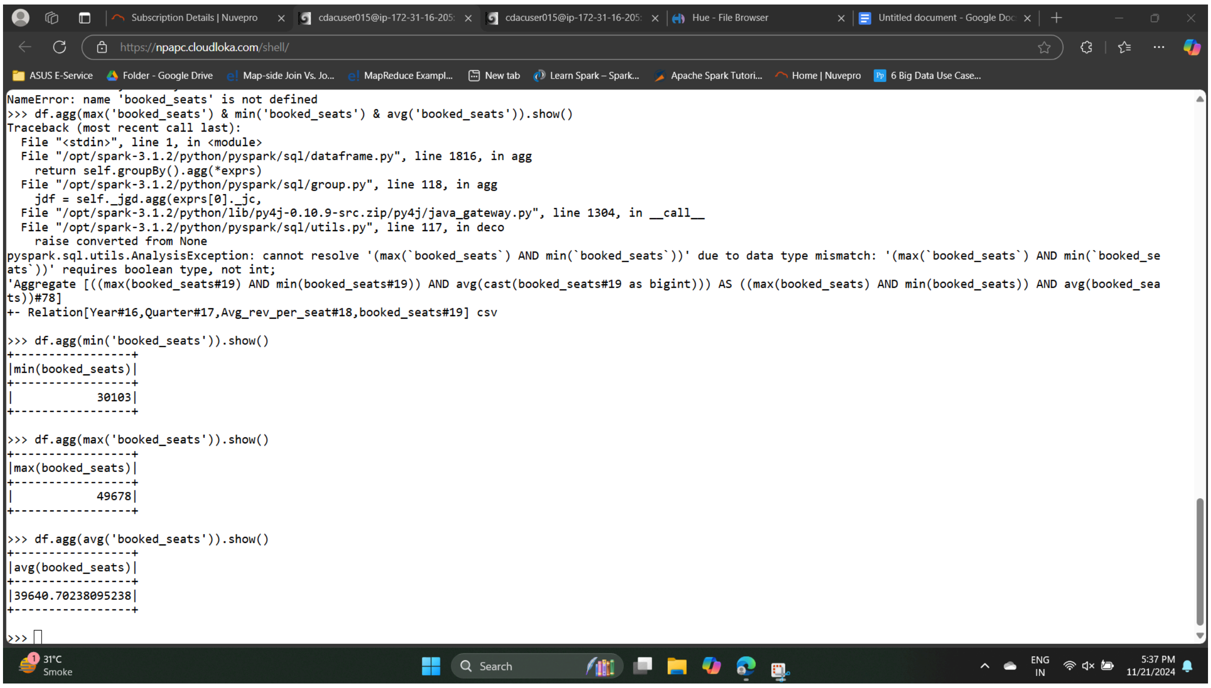


```
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:220)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:74)
        at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:67)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:616)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1826)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1773)
        at org.apache.hadoop.hive.ql.Driver.compileAndRespond(Driver.java:1768)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.compileAndRespond(ReExecDriver.java:126)
        at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:214)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:239)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:188)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:402)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:0 cannot recognize input near 'formated' 'routepart' '<EOF>'
hive (cdac_bigdata)> describe routepart;
OK
airline_iata            string
airline_id              int
src_airport_iata        string
src_airport_id          int
dest_airport_id         int
codeshare               string
stops                   int
equipment               string
dest_airport_iata       string

# Partition Information
# col_name              data_type               comment
dest_airport_iata       string
Time taken: 0.705 seconds, Fetched: 13 row(s)
hive (cdac_bigdata)> 
```

Spark

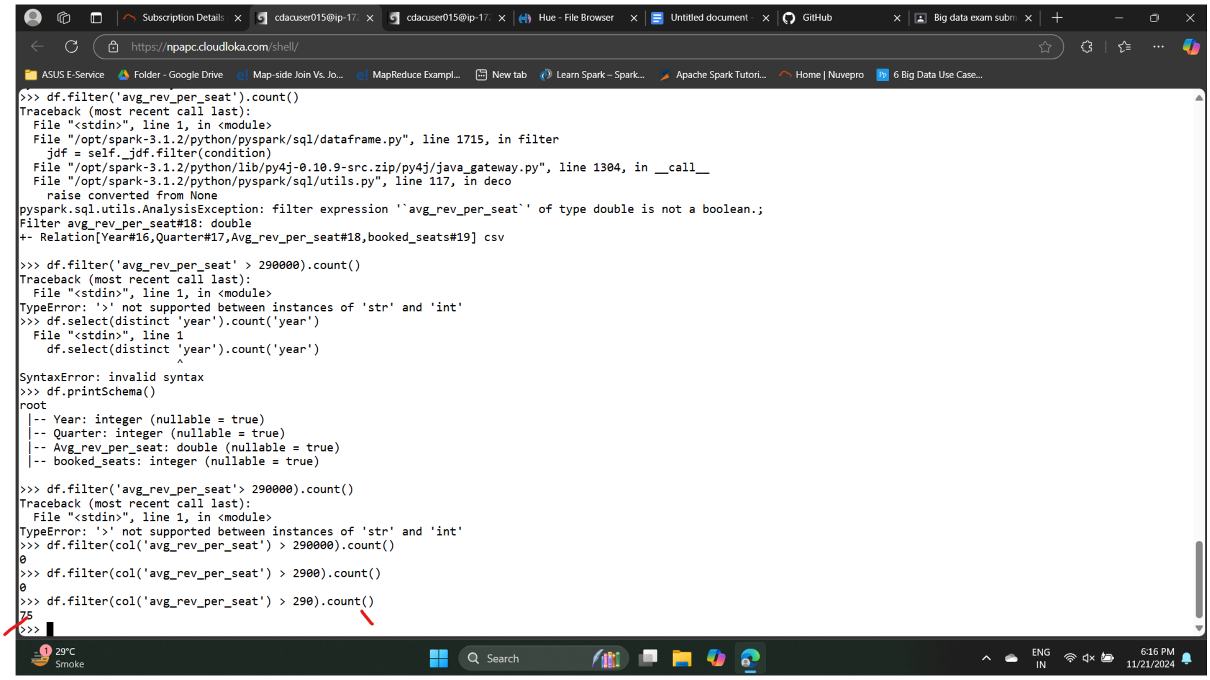
Q2.

1.df.agg(max('booked_seats')).show()



```
NameError: name 'booked_seats' is not defined
>>> df.agg(max('booked_seats') & min('booked_seats') & avg('booked_seats')).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1816, in agg
    return self.groupBy().agg(*exprs)
  File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 118, in agg
    jdf = self._jgd.agg(exprs[0]._jc,
  File "/opt/spark-3.1.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark-3.1.2/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve '(max(`booked_seats`) AND min(`booked_seats`))' due to data type mismatch: '(max(`booked_se
ats`))' requires boolean type, not int;
'Aggregate [((max(booked_seats#19) AND min(booked_seats#19)) AND avg(cast(booked_seats#19 as bigint))) AS ((max(booked_seats) AND min(booked_seats)) AND avg(booked_sea
ts))#78]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.agg(min('booked_seats')).show()
+-----------------+
|min(booked_seats)|
+-----------------+
|            30103|
+-----------------+

>>> df.agg(max('booked_seats')).show()
+-----------------+
|max(booked_seats)|
+-----------------+
|            49678|
+-----------------+

>>> df.agg(avg('booked_seats')).show()
+-----------------+
|avg(booked_seats)|
+-----------------+
|39640.70238095238|
+-----------------+

>>> 
```

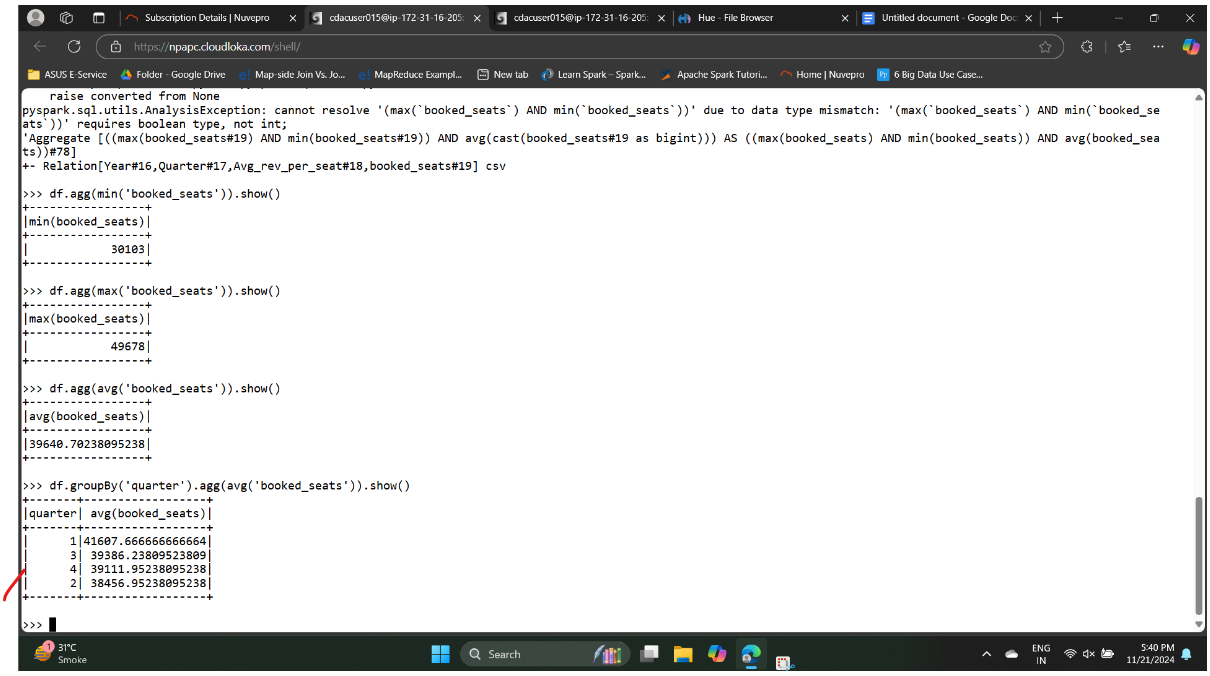
2. df.filter(col('avg_rev_per_seat') > 290).count()



```
>>> df.filter('avg_rev_per_seat').count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1715, in filter
    jdf = self._jdf.filter(condition)
  File "/opt/spark-3.1.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark-3.1.2/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: filter expression '`avg_rev_per_seat`' of type double is not a boolean.;
Filter avg_rev_per_seat#18: double
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.filter('avg_rev_per_seat' > 290000).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: '>' not supported between instances of 'str' and 'int'
>>> df.select(distinct 'year').count('year')
  File "<stdin>", line 1
    df.select(distinct 'year').count('year')
                       ^
SyntaxError: invalid syntax
>>> df.printSchema()
root
 |-- Year: integer (nullable = true)
 |-- Quarter: integer (nullable = true)
 |-- Avg_rev_per_seat: double (nullable = true)
 |-- booked_seats: integer (nullable = true)

>>> df.filter('avg_rev_per_seat'> 290000).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: '>' not supported between instances of 'str' and 'int'
>>> df.filter(col('avg_rev_per_seat') > 290000).count()
0
>>> df.filter(col('avg_rev_per_seat') > 2900).count()
0
>>> df.filter(col('avg_rev_per_seat') > 290).count()
75
>>> 
```

.

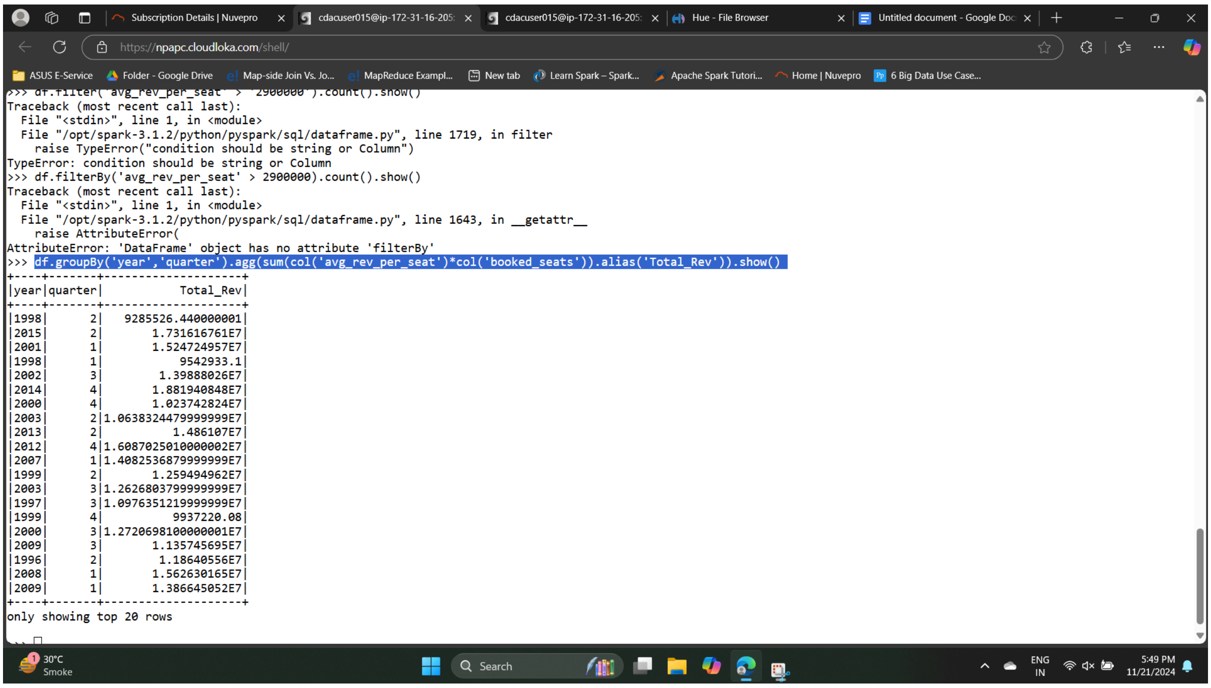
3. df.groupBy('quarter').agg(avg('booked_seats')).show()



```
    raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve '(max(`booked_seats`) AND min(`booked_seats`))' due to data type mismatch: '(max(`booked_seats`) AND min(`booked_se
ats`))' requires boolean type, not int;
'Aggregate [((max(booked_seats#19) AND min(booked_seats#19)) AND avg(cast(booked_seats#19 as bigint))) AS ((max(booked_seats) AND min(booked_seats)) AND avg(booked_sea
ts))#78]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.agg(min('booked_seats')).show()
+-----------------+
|min(booked_seats)|
+-----------------+
|            30103|
+-----------------+

>>> df.agg(max('booked_seats')).show()
+-----------------+
|max(booked_seats)|
+-----------------+
|            49678|
+-----------------+

>>> df.agg(avg('booked_seats')).show()
+-----------------+
|avg(booked_seats)|
+-----------------+
|39640.70238095238|
+-----------------+

>>> df.groupBy('quarter').agg(avg('booked_seats')).show()
+-------+-----------------+
|quarter|avg(booked_seats)|
+-------+-----------------+
|      1|41607.666666666664|
|      3| 39386.23809523809|
|      4| 39111.95238095238|
|      2| 38456.95238095238|
+-------+-----------------+

>>> 
```

4. df.select(distinct 'year').count().show()

5. df.groupBy('year','quarter').agg(sum(col('avg_rev_per_seat')*col('booked_seats')).alias('Total_Rev')).show()



```
>>> df.filter('avg_rev_per_seat' > '2900000').count().show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1719, in filter
    raise TypeError("condition should be string or Column")
TypeError: condition should be string or Column
>>> df.filterBy('avg_rev_per_seat' > 2900000).count().show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1643, in __getattr__
    raise AttributeError(
AttributeError: 'DataFrame' object has no attribute 'filterBy'
>>> df.groupBy('year','quarter').agg(sum(col('avg_rev_per_seat')*col('booked_seats')).alias('Total_Rev')).show()
+----+-------+-------------------+
|year|quarter|          Total_Rev|
+----+-------+-------------------+
|1998|      2|   9285526.440000001|
|2015|      2|      1.731616761E7|
|2001|      1|      1.524724957E7|
|1998|      1|          9542933.1|
|2002|      3|       1.39888026E7|
|2014|      4|      1.881940848E7|
|2000|      4|      1.023742824E7|
|2003|      2|1.0638324479999999E7|
|2013|      2|         1.486107E7|
|2012|      4|1.6087025010000002E7|
|2007|      1|1.4082536879999999E7|
|1999|      2|      1.259494962E7|
|2003|      3|1.2626803799999999E7|
|1997|      3|1.0976351219999999E7|
|1999|      4|         9937220.08|
|2000|      3|1.2720698100000001E7|
|2009|      3|      1.135745695E7|
|1996|      2|       1.18640556E7|
|2008|      1|      1.562630165E7|
|2009|      1|      1.386645052E7|
+----+-------+-------------------+
only showing top 20 rows

>>> 
```