

Name - Sumeet Suryawanshi  
ID - 240840325061

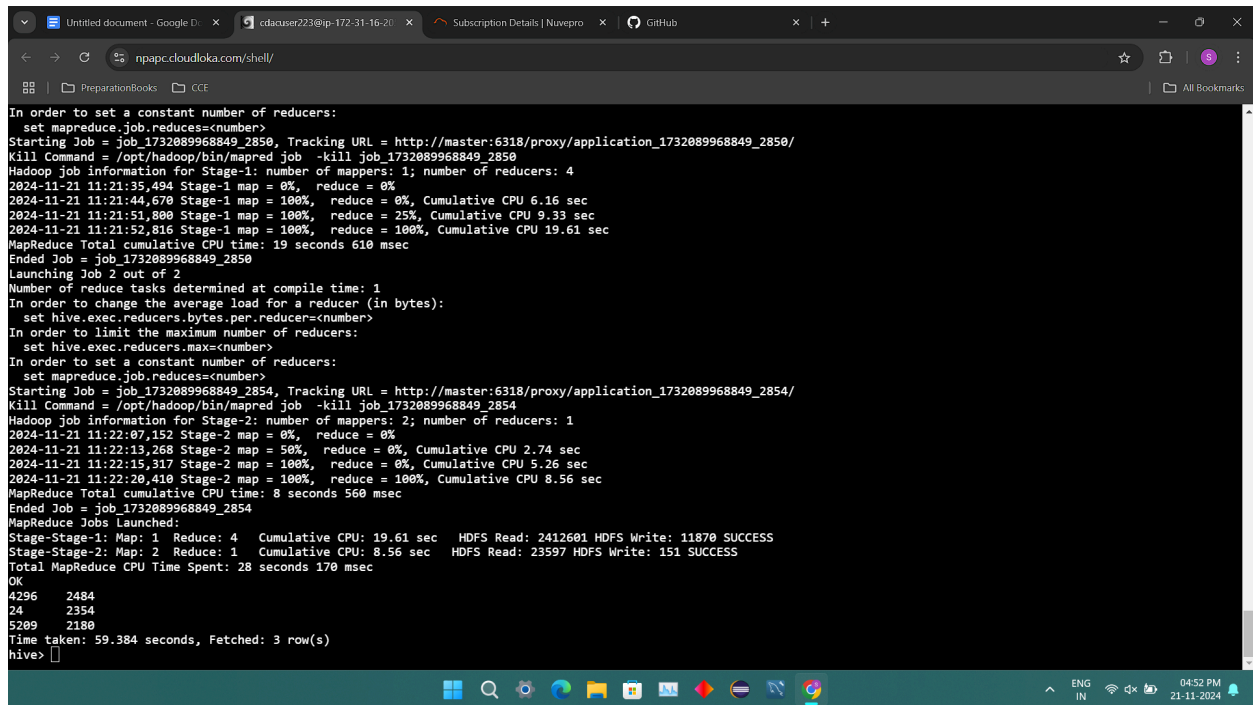
## HIVE

Question 1:

1) Query - `select * from airport a join routes r on a.airport_id = r.src_airport_id where r.src_airport_id is not NULL and r.dest_airport_id is NULL limit 10;`

```
hive> select * from airport a join routes r on a.airport_id = r.src_airport_id where r.src_airport_id is not NULL and r.dest_airport_id is NULL limit 10;
Query ID = cdacuser223_20241121111715_17f38b22-947b-4b6c-be31-ba23909b6bd6
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_2831, Tracking URL = http://master:6318/proxy/application_1732089968849_2831/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2831
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 11:17:27,156 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:17:35,308 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.85 sec
2024-11-21 11:17:41,421 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 23.91 sec
2024-11-21 11:17:43,457 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 27.43 sec
MapReduce Total cumulative CPU time: 27 seconds 430 msec
Ended Job = job_1732089968849_2831
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 27.43 sec HDFS Read: 3179005 HDFS Write: 5724 SUCCESS
Total MapReduce CPU Time Spent: 27 seconds 430 msec
OK
10  Thule Air Base  Thule  Greenland  THU  BGTL  76.531203  -68.703161  251  -4.0  E  America/Thule  GL  921  THU  10
SVR  NULL  BH2
797  Cape Town Intl  Cape Town  South Africa  CPT  FACT  -33.964806  18.601667  151  2.0  U  Africa/Johannesburg  SZ  18946
CPT  797  PBZ  NULL  BEH
910  Ndola  Ndola  Zambia  NLA  FLND  -12.998139  28.664944  4167  2.0  U  Africa/Lusaka  P0  4066  NLA  910  MNS  NULL
0  CN2
1054  Gran Canaria  Gran Canaria  Spain  LPA  GCLP  27.931886  -15.386586  78  0.0  E  Atlantic/Canary  PM  5016  LPA  1054
GLN  NULL  AT4
1094  Nouakchott  Nouakchott  Mauritania  NKC  GQNN  18.097856  -15.947956  7  0.0  N  Africa/Nouakchott  L6  16942
NKC  1094  OUZ  NULL  73G
1118  Makale  Makale  Ethiopia  MQX  HAMK  13.467367  39.533464  7406  3.0  U  Africa/Addis_Ababa  ET  2220  MQX  1118
SZE  NULL  DH8
1353  Provence  Marseille  France  MRS  LFML  43.435555  5.213611  74  1.0  E  Europe/Paris  AH  794  MRS  1353
CFK  NULL  736
1620  Frankfurt  Frankfurt  Germany  FRA  EDDF  50.002500  9.002500  2125  3.0  F  Europe/Frankfurt  1054  FRA  1620  KGO
```

- 2) Query - `select airline_id , count(airline_id) as count from routes where dest airport_id is not NULL and src airport_id is not Null group by airline_id order by count desc limit 3;`



The screenshot shows a terminal window with the following content:

```
npacc.cloudloka.com/shell/
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2850, Tracking URL = http://master:6318/proxy/application_1732089968849_2850/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2850
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 11:21:35,494 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:21:44,670 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.16 sec
2024-11-21 11:21:51,800 Stage-1 map = 100%, reduce = 25%, Cumulative CPU 9.33 sec
2024-11-21 11:21:52,816 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 19.61 sec
MapReduce Total cumulative CPU time: 19 seconds 610 msec
Ended Job = job_1732089968849_2850
Launching Job 2 out of 2
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_2854, Tracking URL = http://master:6318/proxy/application_1732089968849_2854/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2854
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 11:22:07,152 Stage-2 map = 0%, reduce = 0%
2024-11-21 11:22:13,268 Stage-2 map = 50%, reduce = 0%, Cumulative CPU 2.74 sec
2024-11-21 11:22:15,317 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 5.26 sec
2024-11-21 11:22:20,410 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.56 sec
MapReduce Total cumulative CPU time: 8 seconds 560 msec
Ended Job = job_1732089968849_2854
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 4 Cumulative CPU: 19.61 sec HDFS Read: 2412601 HDFS Write: 11870 SUCCESS
Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 8.56 sec HDFS Read: 23597 HDFS Write: 151 SUCCESS
Total MapReduce CPU Time Spent: 28 seconds 170 msec
OK
4296 2484
24 2354
5209 2180
Time taken: 59.384 seconds, Fetched: 3 row(s)
hive>
```

- 3) Query - `select equipment,count(equipment) from routes group by equipment having equipment is not NULL limit 10;`

```
hives> select equipment,count(equipment) from routes group by equipment having equipment is not NULL limit 10;
FAILED: SemanticException Line 0:-1 Expression not in GROUP BY key 'equipment'
hives> select equipment,count(equipment) from routes group by equipment having equipment is not NULL limit 10;
Query ID = cdacuser223_20241121112843_35b9a025-b23a-43c7-b871-5e124b76ece2
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_2889, Tracking URL = http://master:6318/proxy/application_1732089968849_2889/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2889
Hadoop job information for Stage=1: number of mappers: 1; number of reducers: 4
2024-11-21 11:28:54,514 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:29:02,662 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.29 sec
2024-11-21 11:29:09,796 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 16.24 sec
2024-11-21 11:29:10,814 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 19.6 sec
MapReduce Total cumulative CPU time: 19 seconds 600 msec
Ended Job = job_1732089968849_2889
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 4 Cumulative CPU: 19.6 sec HDFS Read: 2412994 HDFS Write: 575 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 600 msec
OK
18
100 319 6
100 319 ER4 BEH 1
73M 733 73C 1
100 284
100 319 320 1
CNA 4
777 1
100 319 2
100 319 ER4 1
Time taken: 29.259 seconds, Fetched: 10 row(s)
hives>
```

## SPARK

```
npapc.cloudloka.com/shell/

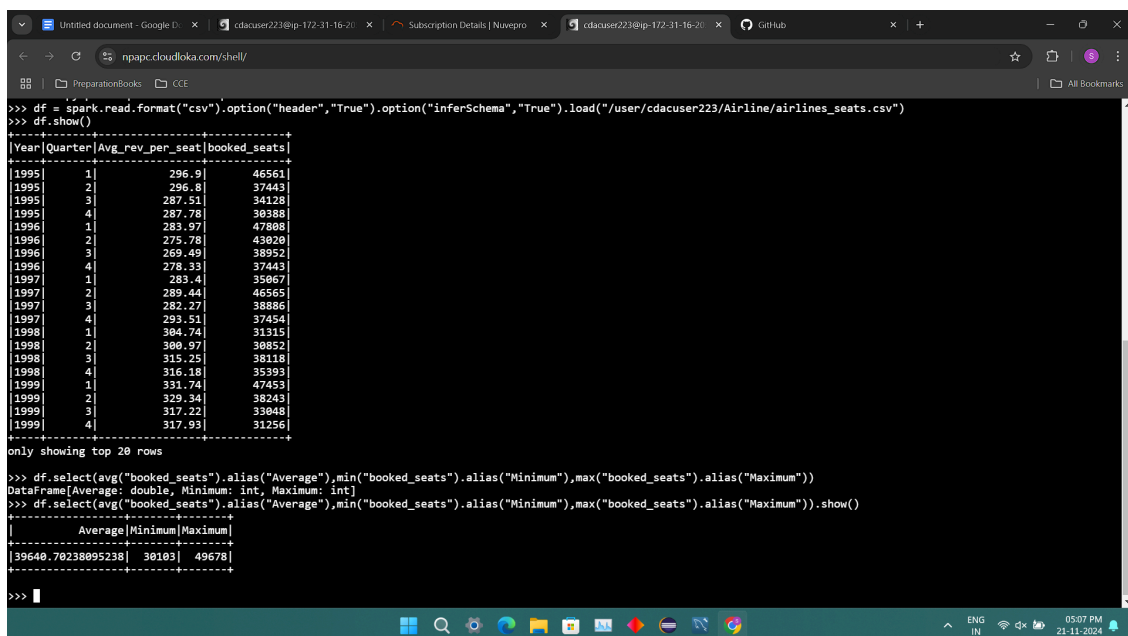
version 3.1.2

Using Python version 3.9.13 (main, Aug 25 2022 23:26:10)
Spark context Web UI available at http://ip-172-31-16-205.ap-south-1.compute.internal:4041
Spark context available as 'sc' (master = yarn, app id = application_1732089968849_2880).
SparkSession available as 'spark'.
>>> from pyspark.sql.functions import *
>>> df = spark.read.format("csv").option("header","True").option("inferSchema","True").load("/user/cdacuser223/Airline/airlines_seats.csv")
>>> df.show()
+---+-----+-----+-----+
|Year|Quarter|Avg_rev_per_seat|booked_seats|
+---+-----+-----+-----+
|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|
|1997|3|282.27|38886|
|1997|4|293.51|37454|
|1998|1|304.74|31315|
|1998|2|300.97|30852|
|1998|3|315.25|38118|
|1998|4|316.18|35393|
|1999|1|331.74|47453|
|1999|2|329.34|38243|
|1999|3|317.22|33048|
|1999|4|317.93|31256|
+---+-----+-----+-----+
only showing top 20 rows
>>>
```

## Question 2 : using DataFrame

### 1) Query -

```
df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum")).show()
```



The screenshot shows a Jupyter Notebook interface with a browser window at the top displaying the URL `npapcloudlaka.com/shell/`. The notebook contains the following code and output:

```
>>> df = spark.read.format("csv").option("header","True").option("inferSchema","True").load("/user/cdacuser223/Airline/airlines_seats.csv")
>>> df.show()
```

Year	Quarter	Avg_rev_per_seat	booked_seats
1995	1	296.9	46561
1995	2	296.8	37443
1995	3	287.51	34128
1995	4	287.78	30388
1996	1	283.97	47888
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
1997	3	282.27	38886
1997	4	293.51	37454
1998	1	304.74	31315
1998	2	300.97	30852
1998	3	315.25	38118
1998	4	316.18	35393
1999	1	331.74	47453
1999	2	329.34	38243
1999	3	317.22	33048
1999	4	317.93	31256

only showing top 20 rows

```
>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum"))
DataFrame[Average: double, Minimum: int, Maximum: int]
>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum")).show()
```

Average	Minimum	Maximum
39640.70238095238	30103	49678

The bottom of the screenshot shows a Windows taskbar with the time 05:07 PM on 23-11-2024.

2) Query - `df.filter(col("Avg_rev_per_seat")<290).count()`

```
npapc.cloudloka.com/shell/

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
1997| 3| 282.27| 38886
1997| 4| 293.51| 37454
1998| 1| 304.74| 31315
1998| 2| 300.97| 30852
1998| 3| 315.25| 38118
1998| 4| 316.18| 35393
1999| 1| 331.74| 47453
1999| 2| 329.34| 38243
1999| 3| 317.22| 33048
1999| 4| 317.93| 31256
-----+-----+
only showing top 20 rows

>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum"))
DataFrame[Average: double, Minimum: int, Maximum: int]
>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum")).show()
+-----+-----+-----+
| Average|Minimum|Maximum|
+-----+-----+-----+
|39640.70238095238| 30103| 49678|
+-----+-----+-----+

>>> df.filter(col("Avg_rev_per_seat")<290).count()Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'ccol' is not defined
>>>
>>> df.filter(col("Avg_rev_per_seat")<290).count()
9
>>>
```

3) Query -

```
df.groupBy("Quarter").agg(avg("Avg_rev_per_seat").alias("Avg_per_Quarter")).show(10)
```

```
npapc.cloudloka.com/shell/

>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum"))
DataFrame[Average: double, Minimum: int, Maximum: int]
>>> df.select(avg("booked_seats").alias("Average"),min("booked_seats").alias("Minimum"),max("booked_seats").alias("Maximum")).show()
+-----+-----+-----+
| Average|Minimum|Maximum|
+-----+-----+-----+
|39640.70238095238| 30103| 49678|
+-----+-----+-----+

>>> df.filter(col("Avg_rev_per_seat")<290).count()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'ccol' is not defined
>>>
>>> df.filter(col("Avg_rev_per_seat")<290).count()
9
>>> df.groupBy("Quarter").select(avg("Avg_rev_per_seat")).limit(10).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'select'
>>> df.select(avg("Avg_rev_per_seat")).groupBy("Quarter").limit(10).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'limit'
>>> df.select(avg("Avg_rev_per_seat")).groupBy("Quarter").show(10)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'show'
>>> df.groupBy("Quarter").agg(avg("Avg_rev_per_seat").alias("Avg_per_Quarter")).show(10)
+-----+-----+
|Quarter| Avg_per_Quarter|
+-----+-----+
| 1|330.61238095238093|
| 3| 327.5557142857143|
| 4|328.48285714285714|
| 2|332.33904761904756|
+-----+-----+

>>> 
```

#### 4) Query -

```
df.groupBy("Year").agg(count("Year").alias("Year_count")).show()
```

```
npapc.cloudloka.com/shell/

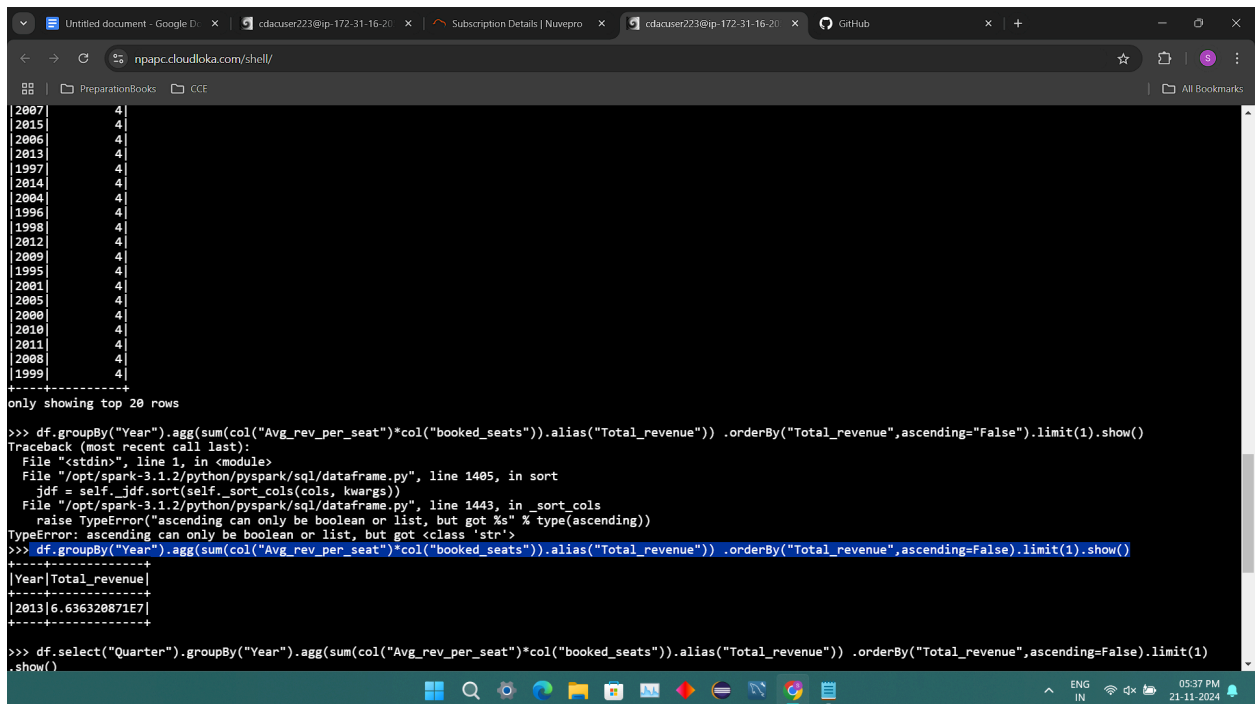
+-----+-----+
| 1|330.61238095238093|
| 3| 327.5557142857143|
| 4|328.48285714285714|
| 2|332.33904761904756|
+-----+-----+

>>> df.groupBy("Year").agg(count("Year").alias("Year_count")).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'ddf' is not defined
>>> df.groupBy("Year").agg(count("Year").alias("Year_count")).show()
+-----+-----+
|Year|Year_count|
+-----+-----+
|2003| 4|
|2007| 4|
|2015| 4|
|2006| 4|
|2013| 4|
|1997| 4|
|2014| 4|
|2004| 4|
|1996| 4|
|1998| 4|
|2012| 4|
|2009| 4|
|1995| 4|
|2001| 4|
|2005| 4|
|2000| 4|
|2010| 4|
|2011| 4|
|2008| 4|
|1999| 4|
+-----+-----+
only showing top 20 rows

>>> 
```

## 5) Query -

```
df.groupBy("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")).alias("Total_revenue"))  
.orderBy("Total_revenue",ascending=False).limit(1).show()
```

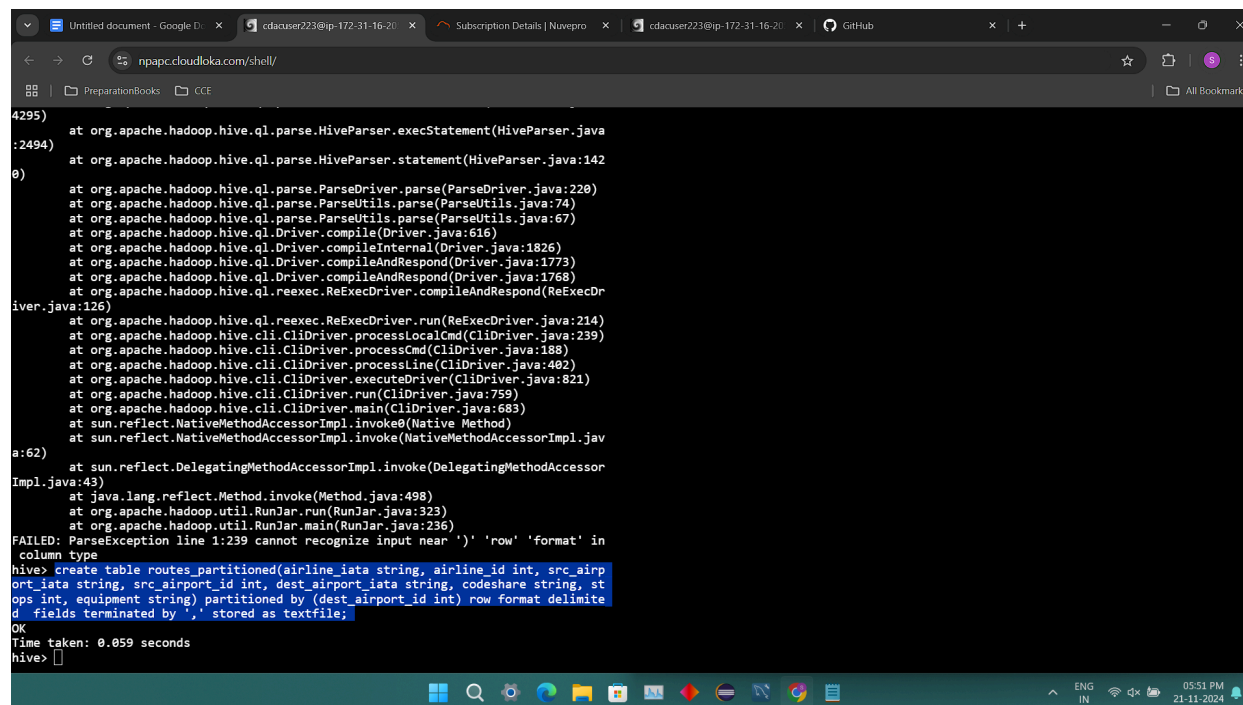


```
npapc.cloudloka.com/shell/
|
| PreparationBooks | CCE
|
|2007| 4|
|2015| 4|
|2006| 4|
|2013| 4|
|1997| 4|
|2014| 4|
|2004| 4|
|1996| 4|
|1998| 4|
|2012| 4|
|2009| 4|
|1995| 4|
|2001| 4|
|2005| 4|
|2000| 4|
|2010| 4|
|2011| 4|
|2008| 4|
|1999| 4|
+-----+
only showing top 20 rows
>>> df.groupBy("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")).alias("Total_revenue")) .orderBy("Total_revenue",ascending=False).limit(1).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1405, in sort
    jdf = self._jdf.sort(self._sort_cols(cols, kwargs))
  File "/opt/spark-3.1.2/python/pyspark/sql/dataframe.py", line 1443, in _sort_cols
    raise TypeError("ascending can only be boolean or list, but got %s" % type(ascending))
TypeError: ascending can only be boolean or list, but got <class 'str'>
>>> df.groupBy("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")).alias("Total_revenue")) .orderBy("Total_revenue",ascending=False).limit(1).show()
+-----+
|Year|Total_revenue|
+-----+
|2013|6.636320871E7|
+-----+
>>> df.select("Quarter").groupBy("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")).alias("Total_revenue")) .orderBy("Total_revenue",ascending=False).limit(1).show()
```

## HIVE

Question 2 :

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



```
4295) at org.apache.hadoop.hive.q1.parse.HiveParser.execStatement(HiveParser.java
:2494) at org.apache.hadoop.hive.q1.parse.HiveParser.statement(HiveParser.java:142
0) at org.apache.hadoop.hive.q1.parse.ParseDriver.parse(ParseDriver.java:220)
at org.apache.hadoop.hive.q1.parse.ParseUtils.parse(ParseUtils.java:74)
at org.apache.hadoop.hive.q1.parse.ParseUtils.parse(ParseUtils.java:67)
at org.apache.hadoop.hive.q1.Driver.compile(Driver.java:616)
at org.apache.hadoop.hive.q1.Driver.compileInternal(Driver.java:1826)
at org.apache.hadoop.hive.q1.Driver.compileAndRespond(Driver.java:1773)
at org.apache.hadoop.hive.q1.Driver.compileAndRespond(Driver.java:1768)
at org.apache.hadoop.hive.q1.rexec.ReExecDriver.compileAndRespond(ReExecDr
iver.java:126)
at org.apache.hadoop.hive.q1.rexec.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.jav
a:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessor
Impl.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 1:239 cannot recognize input near ')' 'row' 'format' in
column type
hive> create table routes partitioned(airline_iata string, airline_id int, src_airp
ort_iata string, src_airport_id int, dest_airport_iata string, codeshare string, st
ops int, equipment string) partitioned by (dest_airport_id int) row format delimite
d fields terminated by ',' stored as textfile;
Ok
Time taken: 0.059 seconds
hive>
```



2) Query - `insert overwrite table routes_partitioned partition(dest_airport_id) select airline_iata,airline_id,src_airport_iata,src_airport_id,dest_airport_iata,dest_airport_id,codeshare,stops,equipment from routes;`

```
Untitled document - Google... cdacuser223@ip-172-31-1... Subscription Details | Nav... Hue - File Browser... cdacuser223@ip-172-31-1... GitHub...
npacc.cloudloka.com/shell/
PreparationBooks CCE
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 1:109 cannot recognize input near 'airline_id' 'int' ',' in selection target
hive> insert overwrite table routes_partitioned partition(dest_airport_id) select airline_iata,airline_id,src_airport_iata,src_airport_id,dest_airport_iata,dest_airport_id,
FAILED: ParseException line 1:191 missing EOF at ')' near 'equipment'
hive> insert overwrite table routes_partitioned partition(dest_airport_id) select airline_iata,airline_id,src_airport_iata,src_airport_id,dest_airport_iata,dest_airport_id,
_id,codeshare,stops,equipment from routes;
Query ID = cdacuser223_202411211122915_13c977b9-6f8d-4c5b-b912-91f1ba7875f8
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_3061, Tracking URL = http://master:6318/proxy/application_1732089968849_3061/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_3061
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 12:29:25,499 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:30:26,485 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 21.82 sec
2024-11-21 12:30:49,831 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 27.23 sec
2024-11-21 12:31:50,734 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 37.18 sec
2024-11-21 12:32:51,597 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 43.69 sec
2024-11-21 12:33:52,465 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 49.92 sec
2024-11-21 12:34:53,324 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 55.76 sec
```

Problem - Mapping and Reducing task was stuck as shown in the screenshot

The screenshot shows the Hue File Browser interface. The left sidebar displays a tree view of tables under the 'sumeet007' database, including 'airlines', 'airport', 'indianairports', 'new\_txnrecsbystate', 'nyse', 'routes', 'routes\_partitioned', 'txn\_orc', 'txn\_parquet', 'txnrecords', 'txnrecsbycat', 'txnrecsbycat2', 'txnrecsbycat3', and 'txnrecsbycat4'. The main panel shows the file browser for the path '/ user / hive / warehouse / sumeet007.db / routes\_partitioned / .hive-staging\_hive\_2024-11-21\_12-29-15\_021\_3288721526529729354-1'. A table lists the files in this directory:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">.</a>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:29 AM
<input type="checkbox"/>	<a href="#">-ext-10001</a>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:29 AM
<input type="checkbox"/>	<a href="#">_task_tmp.-ext-10002</a>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<a href="#">_tmp.-ext-10002</a>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:32 AM

At the bottom of the interface, the Windows taskbar is visible, showing the system clock as 05:02 PM on 21-11-2024.

Untitled document - Google... x cdacuser223@ip-172-31-1 x Subscription Details [New... x Hue - File Browser x cdacuser223@ip-172-31-1 x GitHub x + -

Not secure npapc.cloudloka.com:8132/hue/filebrowser/view-%2Fuser%2Fcdacuser223#/user/hive/warehouse/sumeet007.db/routes\_partitioned/hive-staging\_hive\_2024-11-21-12-29-15\_0... ☆ All Bookmarks

Search data and saved documents...

Jobs

File Browser

Search for file name Actions Delete forever Upload New

sumeet007 Tables (17) +

Filter...

- airlines
- airport
- indianaairports
- new\_txnrecsbystate
- nyse
- routes
- routes\_partitioned
- airline\_lata (string)
- airline\_id (int)
- src\_airport\_lata (string)
- src\_airport\_id (int)
- dest\_airport\_lata (string)
- codeshare (string)
- stops (int)
- equipment (string)
- dest\_airport\_id (int)
- txn\_orc
- txn\_parquet
- txnrecords
- txnrecsbycat
- txnrecsbycat2
- txnrecsbycat3
- txnrecsbycat4

Home

/ user / hive / warehouse / sumeet007.db / routes\_partitioned / .hive-staging\_hive\_2024-11-21-12-29-15\_021\_3288721526529729354-1 / \_task\_tmp.-ext-10002

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<b>f</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:29 AM
<input type="checkbox"/>	<b>.</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id= 73W 733 73C</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id= 777</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id= CNA</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:29 AM
<input type="checkbox"/>	<b>dest_airport_id=100</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:29 AM
<input type="checkbox"/>	<b>dest_airport_id=100 318</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id=100 319</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id=100 319 320</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id=100 319 ER4</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:30 AM
<input type="checkbox"/>	<b>dest_airport_id=100 319 ER4 BEH</b>		cdacuser223	hive	drwxr-xr-x	November 21, 2024 04:31 AM

ENG IN 06:02 PM 21-11-2024

3) Query - select \* from routes\_partitioned where dest\_airport\_id = "ORD"