Name: Khushal Pareta

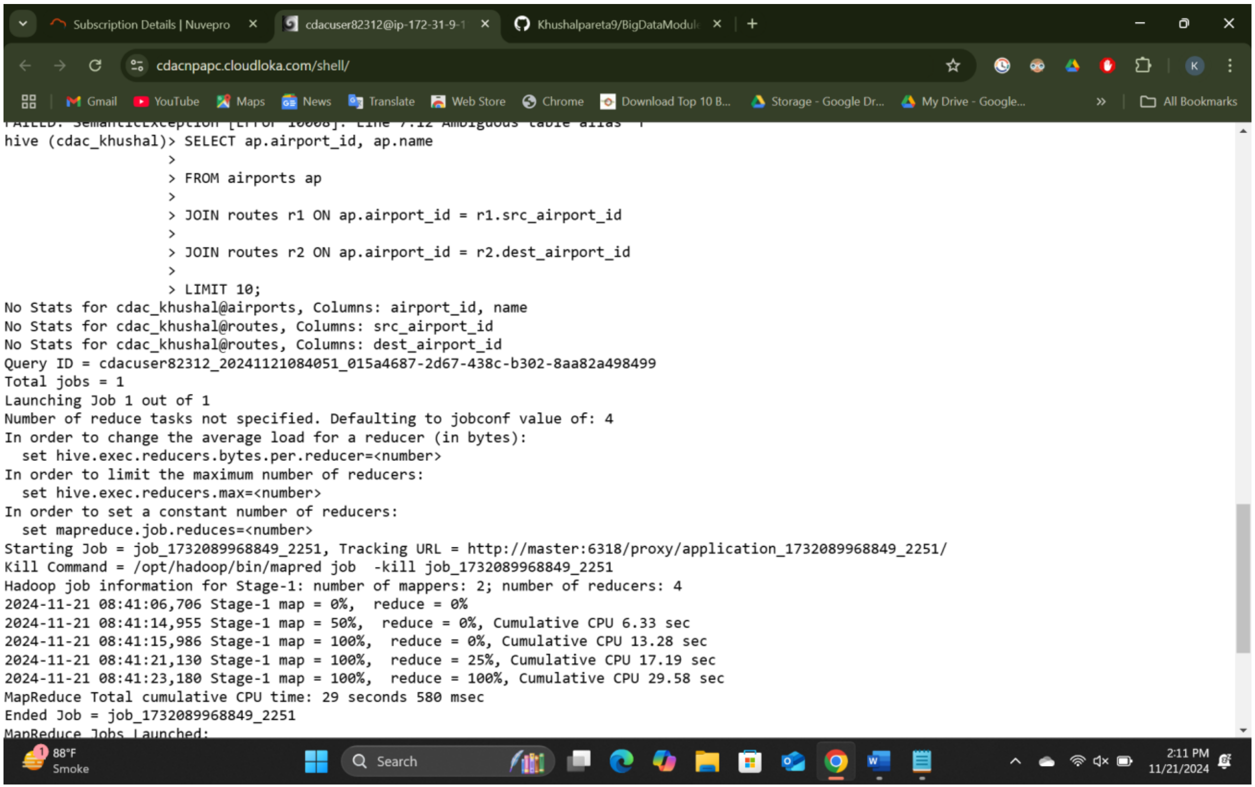
Student ID: 240840325031

#HIVE

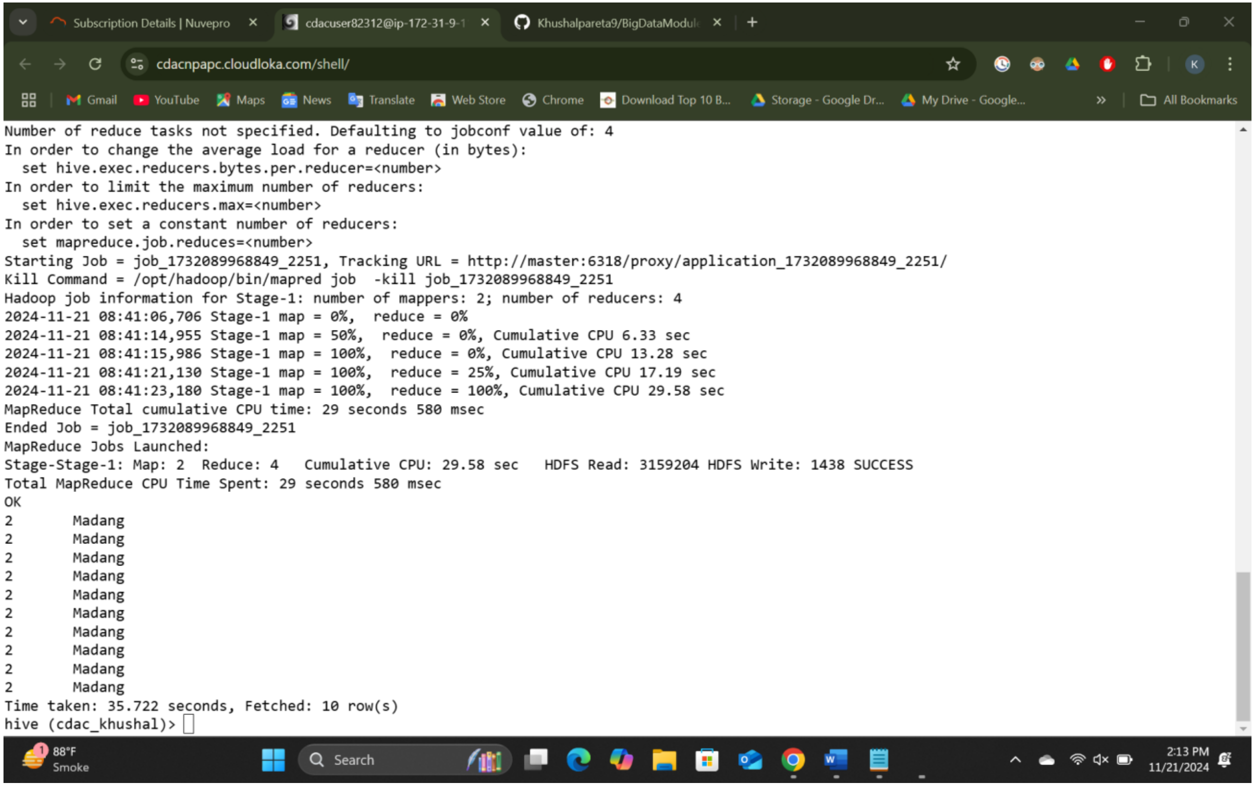
Question 1:

1-> airports that are listed as both source and destination

```
SELECT ap.airport_id, ap.name

FROM airports ap

JOIN routes r1 ON ap.airport_id = r1.src_airport_id

JOIN routes r2 ON ap.airport_id = r2.dest_airport_id

LIMIT 10;
```



```
FAILED: SemanticException [Error 10008]: Line 7:12 Ambiguous table alias 'r'
hive (cdac_khushal)> SELECT ap.airport_id, ap.name
                   >
                   > FROM airports ap
                   >
                   > JOIN routes r1 ON ap.airport_id = r1.src_airport_id
                   >
                   > JOIN routes r2 ON ap.airport_id = r2.dest_airport_id
                   >
                   > LIMIT 10;
No Stats for cdac_khushal@airports, Columns: airport_id, name
No Stats for cdac_khushal@routes, Columns: src_airport_id
No Stats for cdac_khushal@routes, Columns: dest_airport_id
Query ID = cdacuser82312_20241121084051_015a4687-2d67-438c-b302-8aa82a498499
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_2251, Tracking URL = http://master:6318/proxy/application_1732089968849_2251/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2251
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 08:41:06,706 Stage-1 map = 0%,  reduce = 0%
2024-11-21 08:41:14,955 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 6.33 sec
2024-11-21 08:41:15,986 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.28 sec
2024-11-21 08:41:21,130 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 17.19 sec
2024-11-21 08:41:23,180 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 29.58 sec
MapReduce Total cumulative CPU time: 29 seconds 580 msec
Ended Job = job_1732089968849_2251
MapReduce Jobs Launched:
```

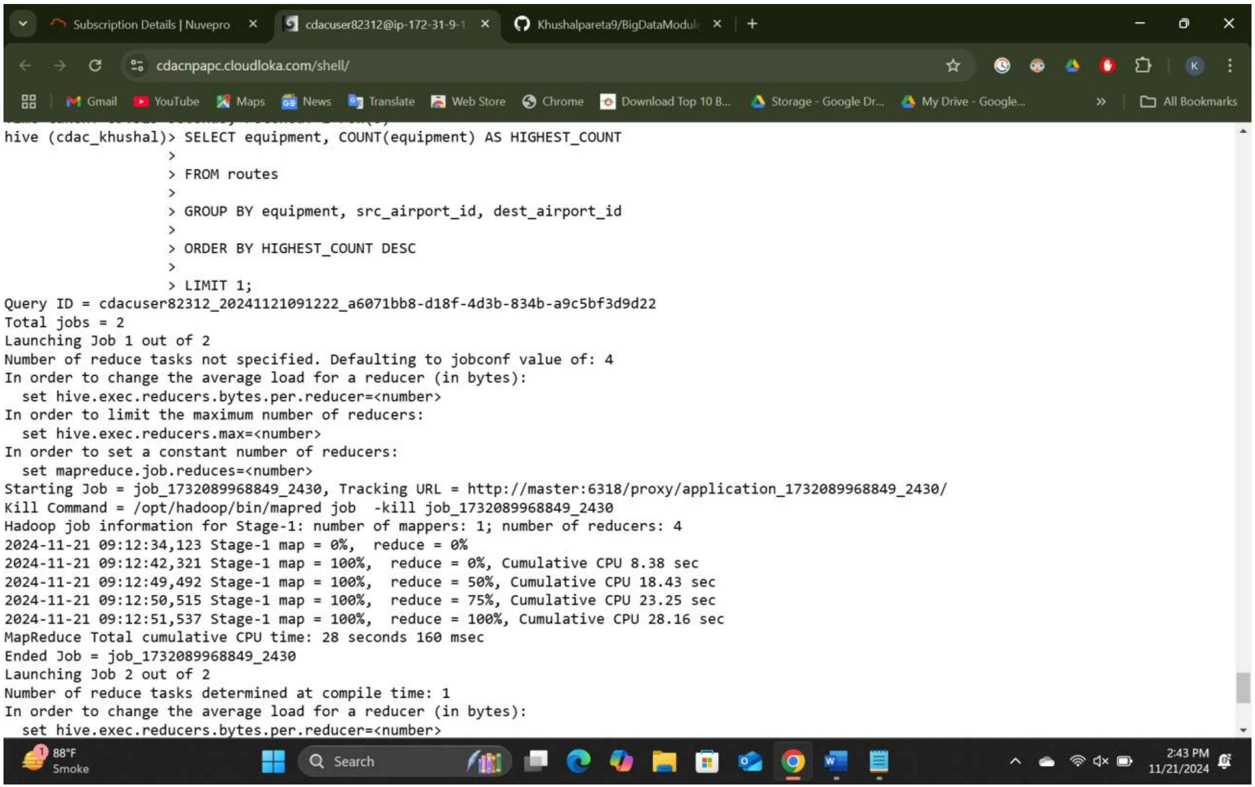


```
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2251, Tracking URL = http://master:6318/proxy/application_1732089968849_2251/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2251
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 08:41:06,706 Stage-1 map = 0%,  reduce = 0%
2024-11-21 08:41:14,955 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 6.33 sec
2024-11-21 08:41:15,986 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.28 sec
2024-11-21 08:41:21,130 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 17.19 sec
2024-11-21 08:41:23,180 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 29.58 sec
MapReduce Total cumulative CPU time: 29 seconds 580 msec
Ended Job = job_1732089968849_2251
MapReduce Jobs Launched:
```

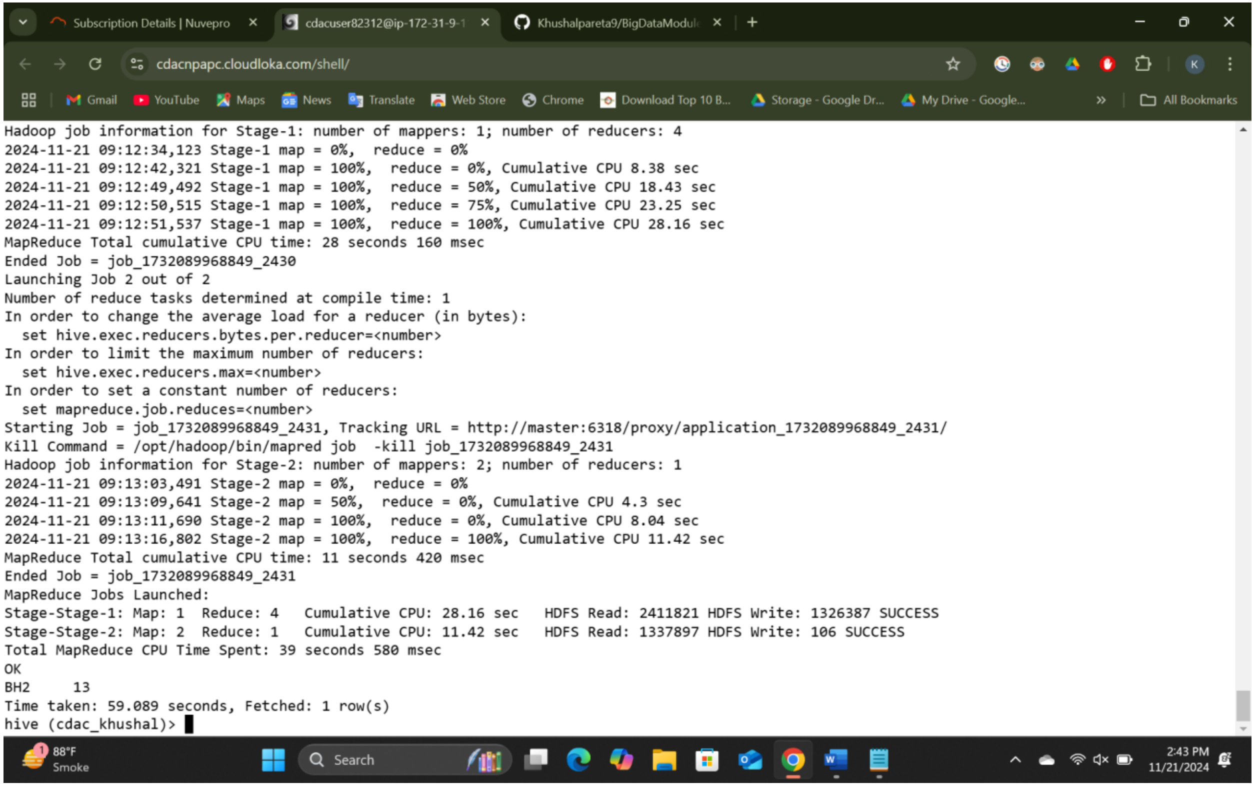
```
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 29.58 sec   HDFS Read: 3159204 HDFS Write: 1438 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 580 msec
OK
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
2       Madang
Time taken: 35.722 seconds, Fetched: 10 row(s)
hive (cdac_khushal)> 
```

2 -> determine equipment that is used on highest number of routes

```
SELECT equipment, COUNT(equipment) AS HIGHEST_COUNT
FROM routes
GROUP BY equipment, src_airport_id, dest_airport_id
ORDER BY HIGHEST_COUNT DESC
LIMIT 1;
```



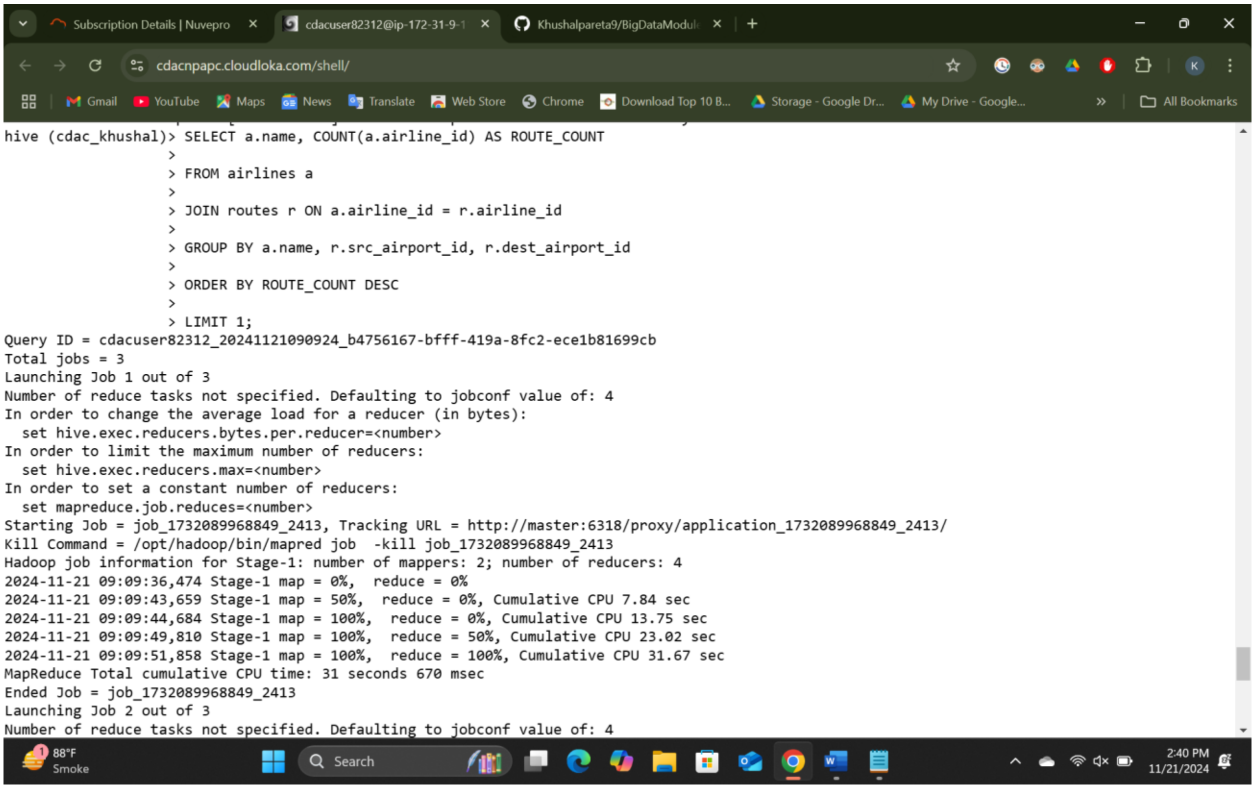
```
hive (cdac_khushal)> SELECT equipment, COUNT(equipment) AS HIGHEST_COUNT
                   >
                   > FROM routes
                   >
                   > GROUP BY equipment, src_airport_id, dest_airport_id
                   >
                   > ORDER BY HIGHEST_COUNT DESC
                   >
                   > LIMIT 1;
Query ID = cdacuser82312_20241121091222_a6071bb8-d18f-4d3b-834b-a9c5bf3d9d22
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_2430, Tracking URL = http://master:6318/proxy/application_1732089968849_2430/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2430
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 09:12:34,123 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:12:42,321 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.38 sec
2024-11-21 09:12:49,492 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 18.43 sec
2024-11-21 09:12:50,515 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 23.25 sec
2024-11-21 09:12:51,537 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.16 sec
MapReduce Total cumulative CPU time: 28 seconds 160 msec
Ended Job = job_1732089968849_2430
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>
```



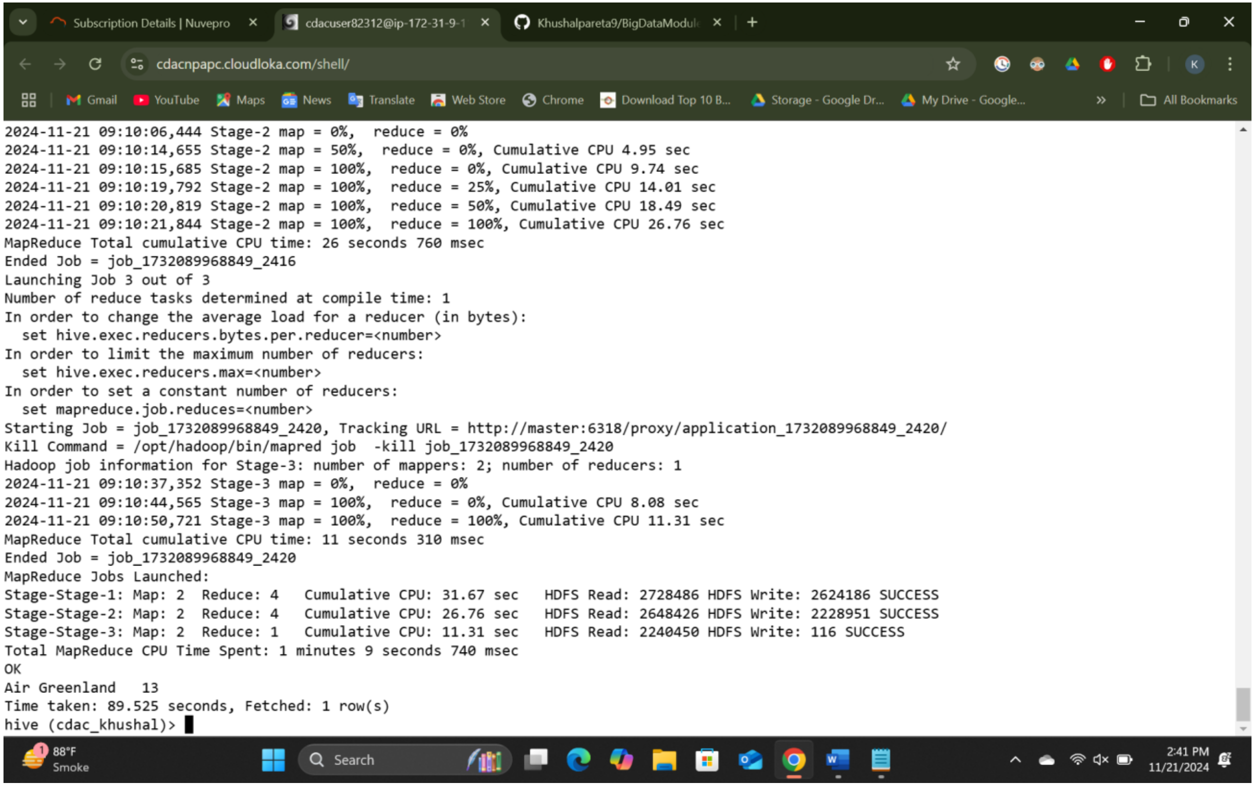
```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 09:12:34,123 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:12:42,321 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.38 sec
2024-11-21 09:12:49,492 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 18.43 sec
2024-11-21 09:12:50,515 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 23.25 sec
2024-11-21 09:12:51,537 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.16 sec
MapReduce Total cumulative CPU time: 28 seconds 160 msec
Ended Job = job_1732089968849_2430
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_2431, Tracking URL = http://master:6318/proxy/application_1732089968849_2431/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2431
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 09:13:03,491 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:13:09,641 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 4.3 sec
2024-11-21 09:13:11,690 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 8.04 sec
2024-11-21 09:13:16,802 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 11.42 sec
MapReduce Total cumulative CPU time: 11 seconds 420 msec
Ended Job = job_1732089968849_2431
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 28.16 sec   HDFS Read: 2411821 HDFS Write: 1326387 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 11.42 sec   HDFS Read: 1337897 HDFS Write: 106 SUCCESS
Total MapReduce CPU Time Spent: 39 seconds 580 msec
OK
BH2     13
Time taken: 59.089 seconds, Fetched: 1 row(s)
hive (cdac_khushal)> 
```

3 - > Airline which operates the highest number of routes and count of those routes

```
SELECT a.name, COUNT(a.airline_id) AS ROUTE_COUNT
FROM airlines a
JOIN routes r ON a.airline_id = r.airline_id
GROUP BY a.name, r.src_airport_id, r.dest_airport_id
ORDER BY ROUTE_COUNT DESC
LIMIT 1;
```



```
hive (cdac_khushal)> SELECT a.name, COUNT(a.airline_id) AS ROUTE_COUNT
                   >
                   > FROM airlines a
                   >
                   > JOIN routes r ON a.airline_id = r.airline_id
                   >
                   > GROUP BY a.name, r.src_airport_id, r.dest_airport_id
                   >
                   > ORDER BY ROUTE_COUNT DESC
                   >
                   > LIMIT 1;
Query ID = cdacuser82312_20241121090924_b4756167-bfff-419a-8fc2-ece1b81699cb
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_2413, Tracking URL = http://master:6318/proxy/application_1732089968849_2413/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2413
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 09:09:36,474 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:09:43,659 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 7.84 sec
2024-11-21 09:09:44,684 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.75 sec
2024-11-21 09:09:49,810 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 23.02 sec
2024-11-21 09:09:51,858 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 31.67 sec
MapReduce Total cumulative CPU time: 31 seconds 670 msec
Ended Job = job_1732089968849_2413
Launching Job 2 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
```



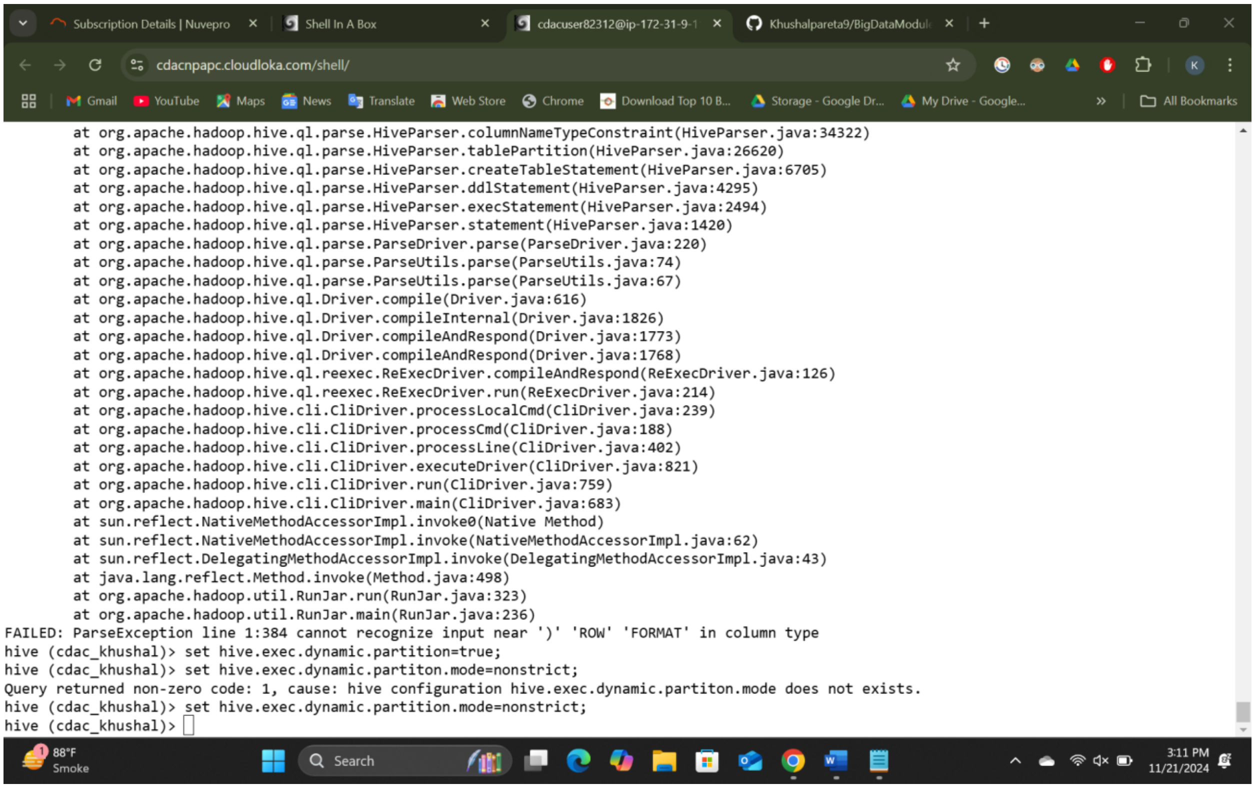
```
2024-11-21 09:10:06,444 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:10:14,655 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 4.95 sec
2024-11-21 09:10:15,685 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 9.74 sec
2024-11-21 09:10:19,792 Stage-2 map = 100%,  reduce = 25%, Cumulative CPU 14.01 sec
2024-11-21 09:10:20,819 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 18.49 sec
2024-11-21 09:10:21,844 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 26.76 sec
MapReduce Total cumulative CPU time: 26 seconds 760 msec
Ended Job = job_1732089968849_2416
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_2420, Tracking URL = http://master:6318/proxy/application_1732089968849_2420/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2420
Hadoop job information for Stage-3: number of mappers: 2; number of reducers: 1
2024-11-21 09:10:37,352 Stage-3 map = 0%,  reduce = 0%
2024-11-21 09:10:44,565 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 8.08 sec
2024-11-21 09:10:50,721 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 11.31 sec
MapReduce Total cumulative CPU time: 11 seconds 310 msec
Ended Job = job_1732089968849_2420
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 31.67 sec   HDFS Read: 2728486 HDFS Write: 2624186 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 26.76 sec   HDFS Read: 2648426 HDFS Write: 2228951 SUCCESS
Stage-Stage-3: Map: 2  Reduce: 1   Cumulative CPU: 11.31 sec   HDFS Read: 2240450 HDFS Write: 116 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 9 seconds 740 msec
OK
Air Greenland   13
Time taken: 89.525 seconds, Fetched: 1 row(s)
hive (cdac_khushal)>
```

Question 2:

1 -> Create a partition table for the source_airport, write a sql query to create this table and insert data into it.

SET hive.exec.dynamic.partition=true;

SET hive.exec.dynamic.partiton.mode=nonstrict;



```
        at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeConstraint(HiveParser.java:34322)
        at org.apache.hadoop.hive.ql.parse.HiveParser.tablePartition(HiveParser.java:26620)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:6705)
        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 1:384 cannot recognize input near ')' 'ROW' 'FORMAT' in column type
hive (cdac_khushal)> set hive.exec.dynamic.partition=true;
hive (cdac_khushal)> set hive.exec.dynamic.partiton.mode=nonstrict;
Query returned non-zero code: 1, cause: hive configuration hive.exec.dynamic.partiton.mode does not exists.
hive (cdac_khushal)> set hive.exec.dynamic.partition.mode=nonstrict;
hive (cdac_khushal)>
```

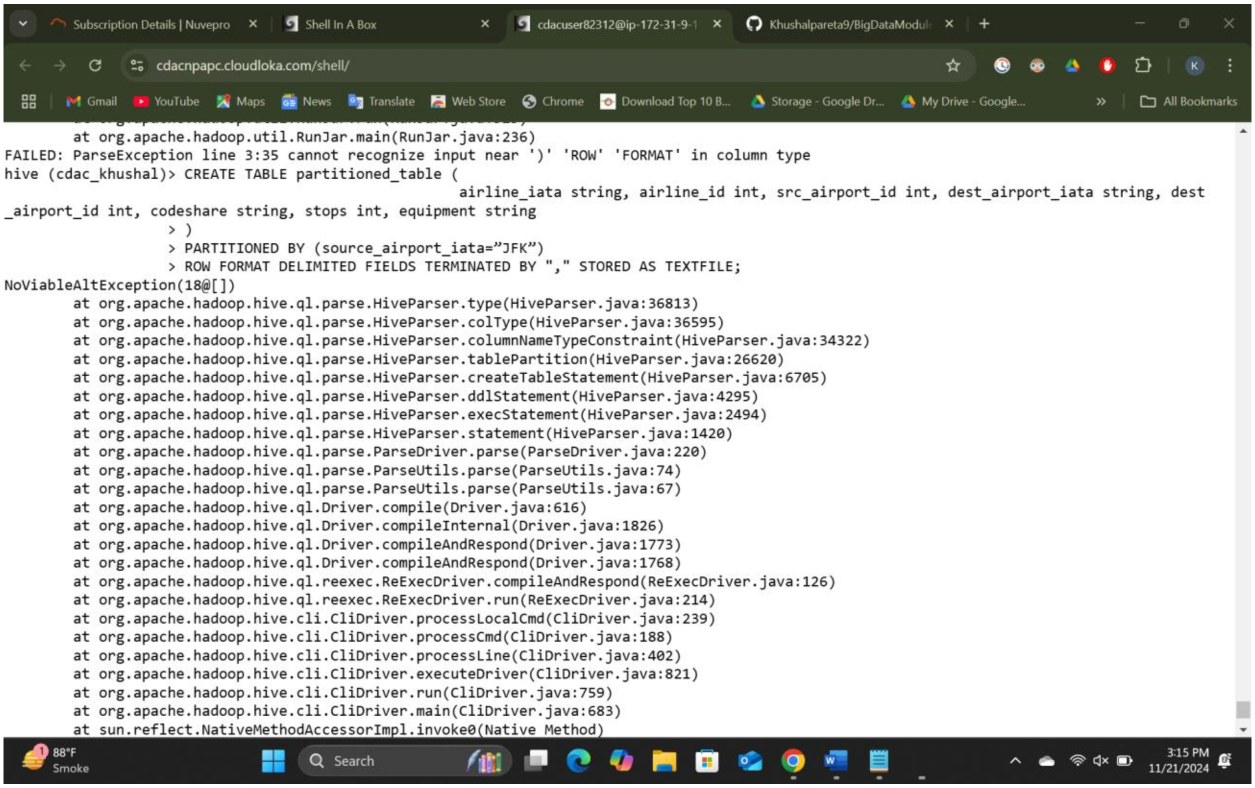
CREATE TABLE partitioned_table (

airline_iata string, airline_id int, src_airport_id int, dest_airport_iata string, dest_airport_id int, codeshare string, stops int, equipment string

)

PARTITIONED BY (source_airport_iata=‘JFK’)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ STORED AS TEXTFILE;



```
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 3:35 cannot recognize input near ')' 'ROW' 'FORMAT' in column type
hive (cdac_khushal)> CREATE TABLE partitioned_table (
                                                    airline_iata string, airline_id int, src_airport_id int, dest_airport_iata string, dest
_airport_id int, codeshare string, stops int, equipment string
                   > )
                   > PARTITIONED BY (source_airport_iata="JFK")
                   > ROW FORMAT DELIMITED FIELDS TERMINATED BY "," STORED AS TEXTFILE;
NoViableAltException(18@[])
        at org.apache.hadoop.hive.ql.parse.HiveParser.type(HiveParser.java:36813)
        at org.apache.hadoop.hive.ql.parse.HiveParser.colType(HiveParser.java:36595)
        at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeConstraint(HiveParser.java:34322)
        at org.apache.hadoop.hive.ql.parse.HiveParser.tablePartition(HiveParser.java:26620)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:6705)
        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)
```

2 ->

INSERT INTO TABLE partitioned_table SELECT * FROM airports WHERE src_airport_iata='JFK'

3 ->

SELECT * FROM partitioned_table WHERE src_airport_iata='LAX';

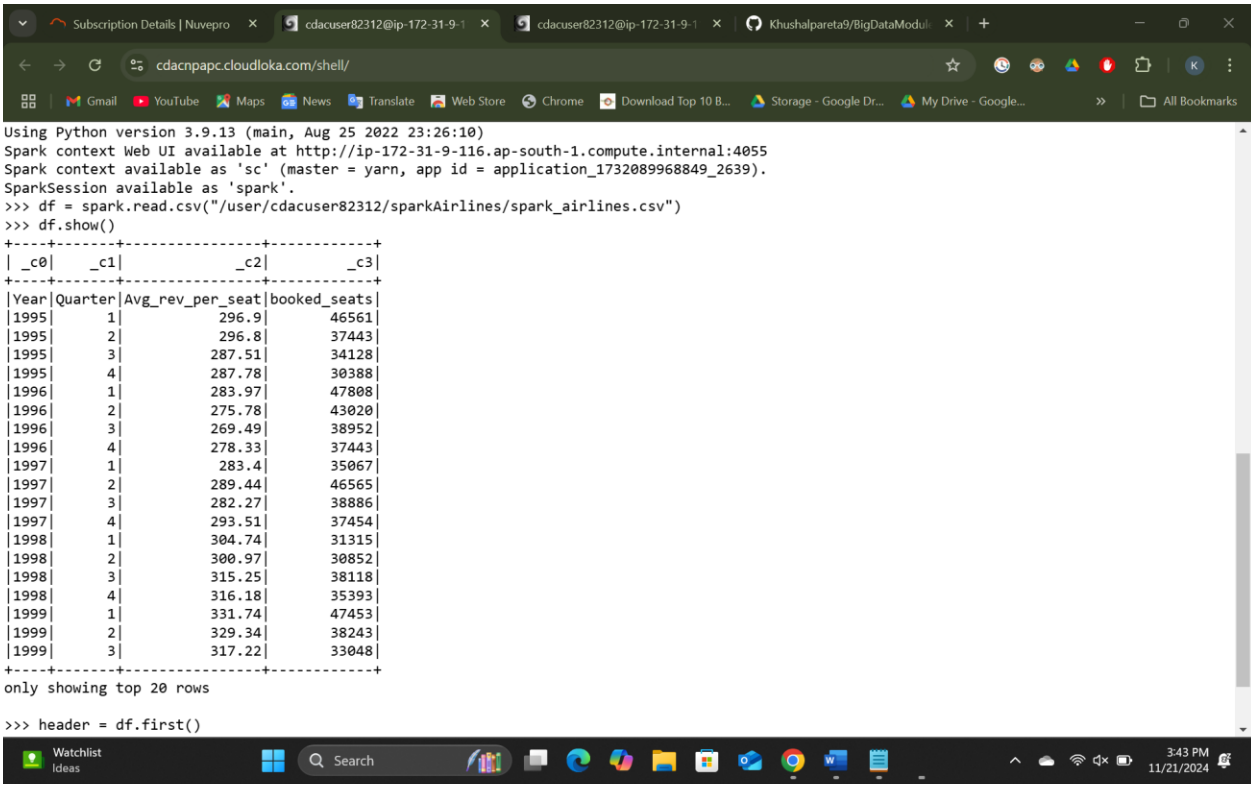
4 ->

SHOW paritions partitioned_table

#SPARK

Question 1:

df = spark.read.csv('/user/cdacuser82312/sparkAirlines/spark_airlines.csv', header=True, inferSchema=True)



```
Using Python version 3.9.13 (main, Aug 25 2022 23:26:10)
Spark context Web UI available at http://ip-172-31-9-116.ap-south-1.compute.internal:4055
Spark context available as 'sc' (master = yarn, app id = application_1732089968849_2639).
SparkSession available as 'spark'.
>>> df = spark.read.csv("/user/cdacuser82312/sparkAirlines/spark_airlines.csv")
>>> df.show()
+----+-------+---------------+-----------+
| _c0|    _c1|            _c2|        _c3|
+----+-------+---------------+-----------+
|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|
+----+-------+---------------+-----------+
only showing top 20 rows

>>> header = df.first()
```

1 ->

From pyspark.sql.functions import sum

df.groupBy('Year', 'Quarter').agg(count(sum('booked_seats')) > 40000).show()

2 ->

df.groupBy('Year').show()

QUESTION 2:

1->

From pyspark.sql.functions import sum, avg, min

df.groupBy('Year', 'Quarter').agg(sum('Avg_rev_per_seat').alias('TotalRevenuePerSeat'), avg('Avg_rev_per_seat').alias('AverageRevenuePerSeat'), min('Avg_rev_per_seat').alias('MinimumRevenuePerSeat')).show()

2 ->

```
  df.groupBy(‘Year’, ‘Quarter’).agg(count(avg_rev_per_seat).alias(‘Total_Count’) > 290)).show()
```

3 ->

```
From pyspark.sql.functions import sum
df.groupBy(‘Year’).agg(sum(‘booked_seats’).alias(‘Total_Booked_Seats)).show()
```

4 ->

```
 df.groupBy(‘year’).show()
```

5 ->

```
From pyspark.sql.functions import sum
 df.groupBy(‘Year’).agg(sum(‘avg_rev_per_seat’).alias(‘Total_Avg_Revenue’)).show()
```