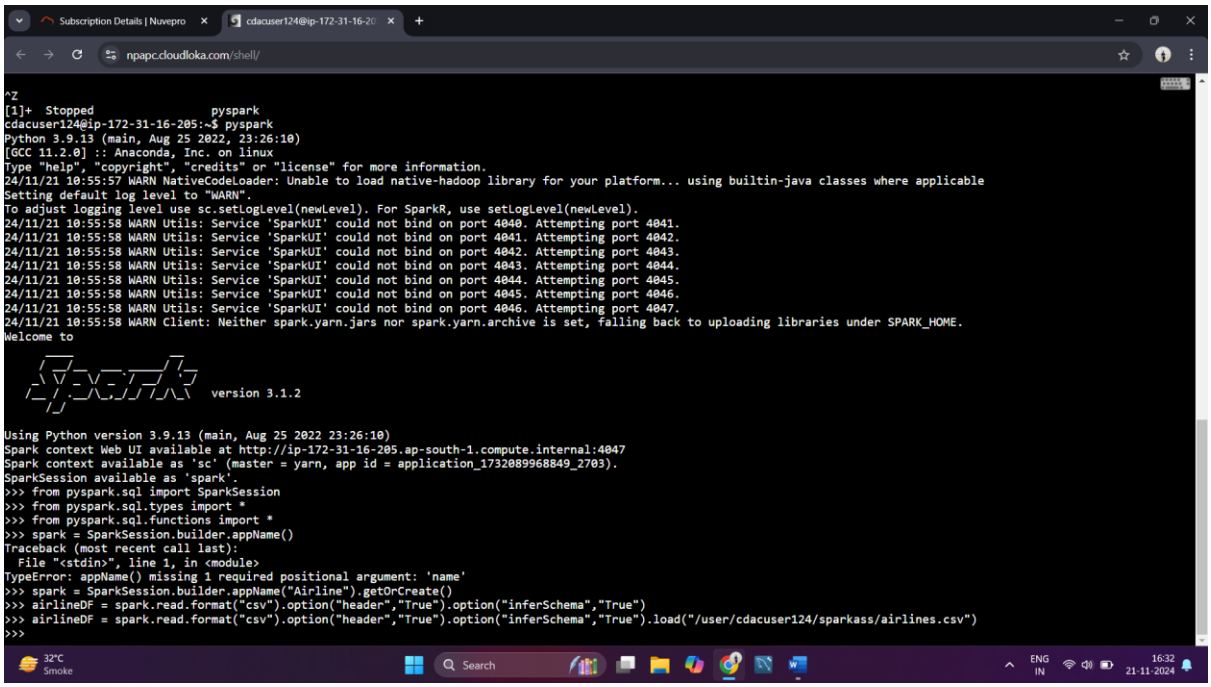
Spark

Q2. By using Dataframe

1.

→ starting pyspark → importing libraries → starting spark session → loading data into dataframe

```
from pyspark.sql import SparkSession

from pyspark.sql.types import *

from pyspark.sql.functions import *

spark = SparkSession.builder.appName("Airline").getOrCreate()

airlineDF = spark.read.format("csv").option("header","True").option("inferSchema","True").load("/user/cdacuser124/sparkass/airlines.csv")
```



Minimum:

```
airlineDF.groupBy("Year","Quarter").agg(min("booked_seats").alias("Minimum_booked_seats")).limit(1).show()
```

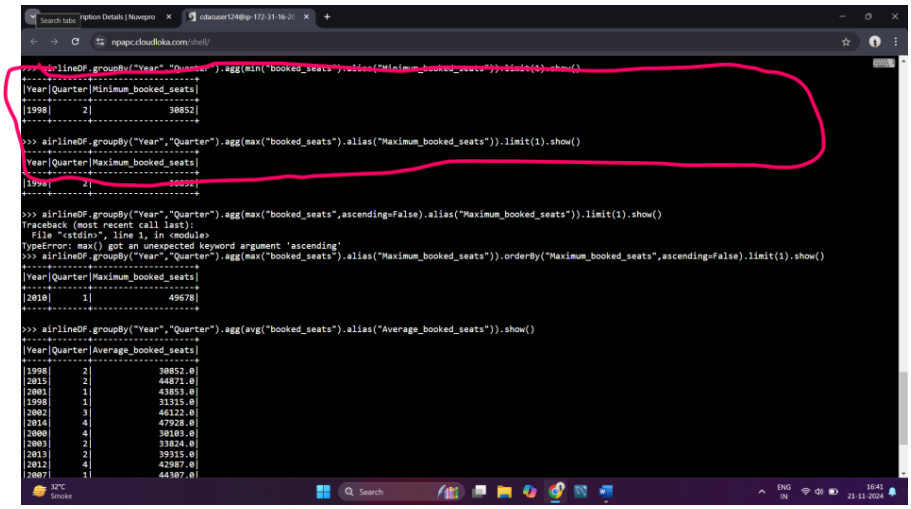
Maximum:

```
airlineDF.groupBy("Year","Quarter").agg(max("booked_seats").alias("Maximum_booked_seats")).limit(1).show()
```

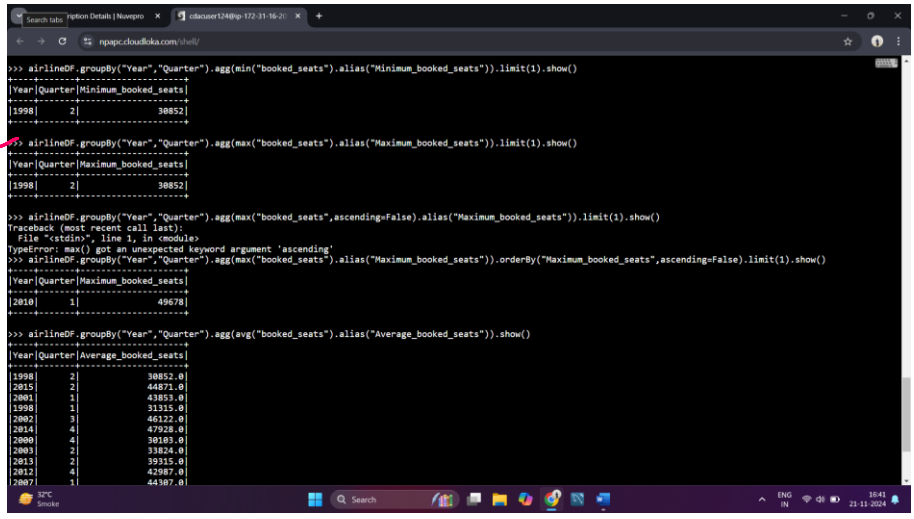
Average:

airlineDF.groupBy("Year","Quarter").agg(avg("booked_seats").alias("Average_booked_seats")).show()

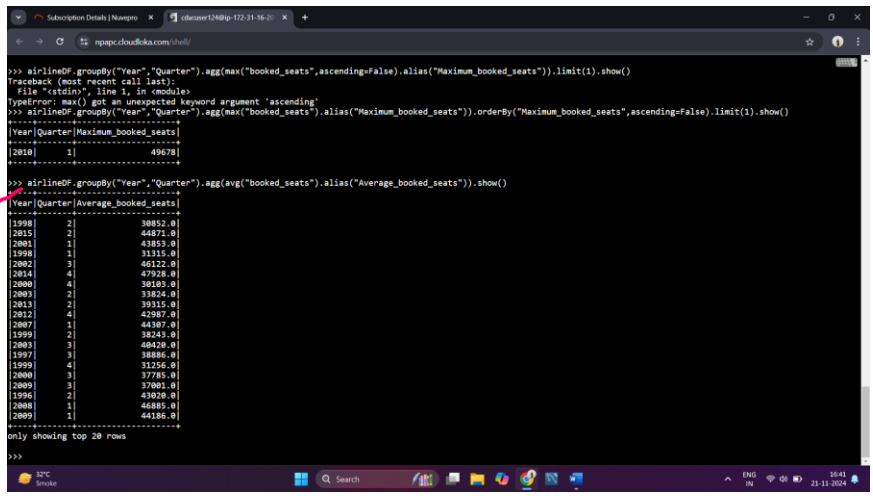
→ Minimum:



Maximum:



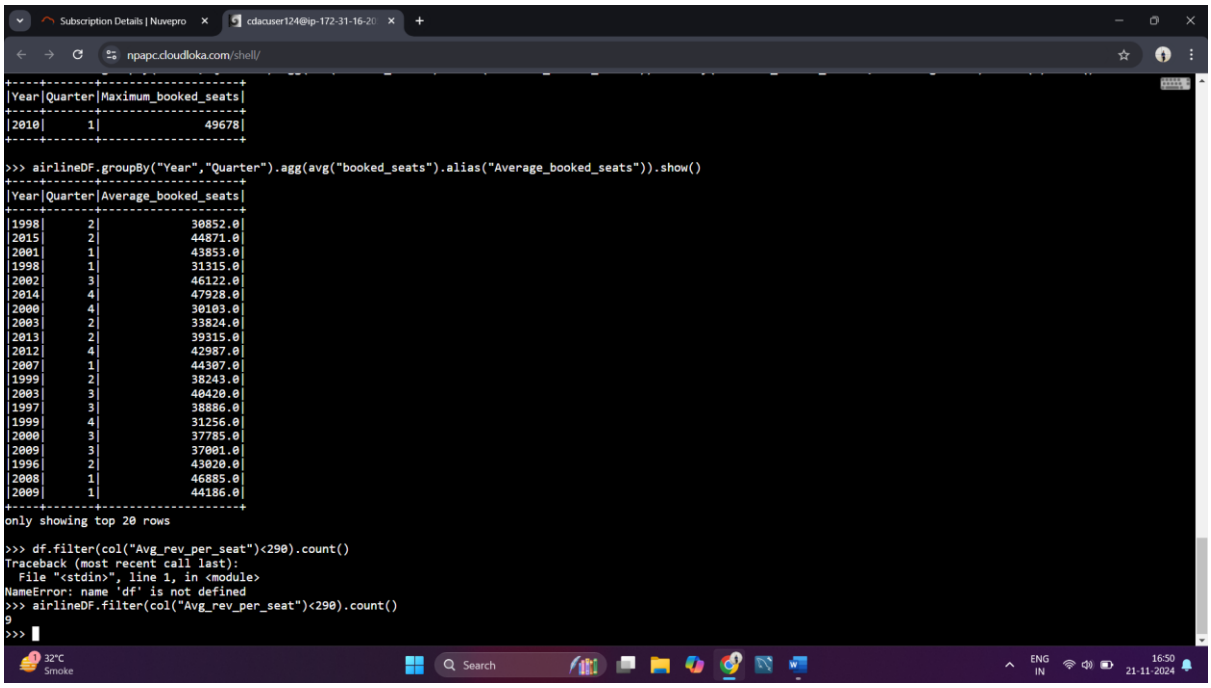
Average:



2.

→

airlineDF.filter(col("Avg_rev_per_seat")<290).count()

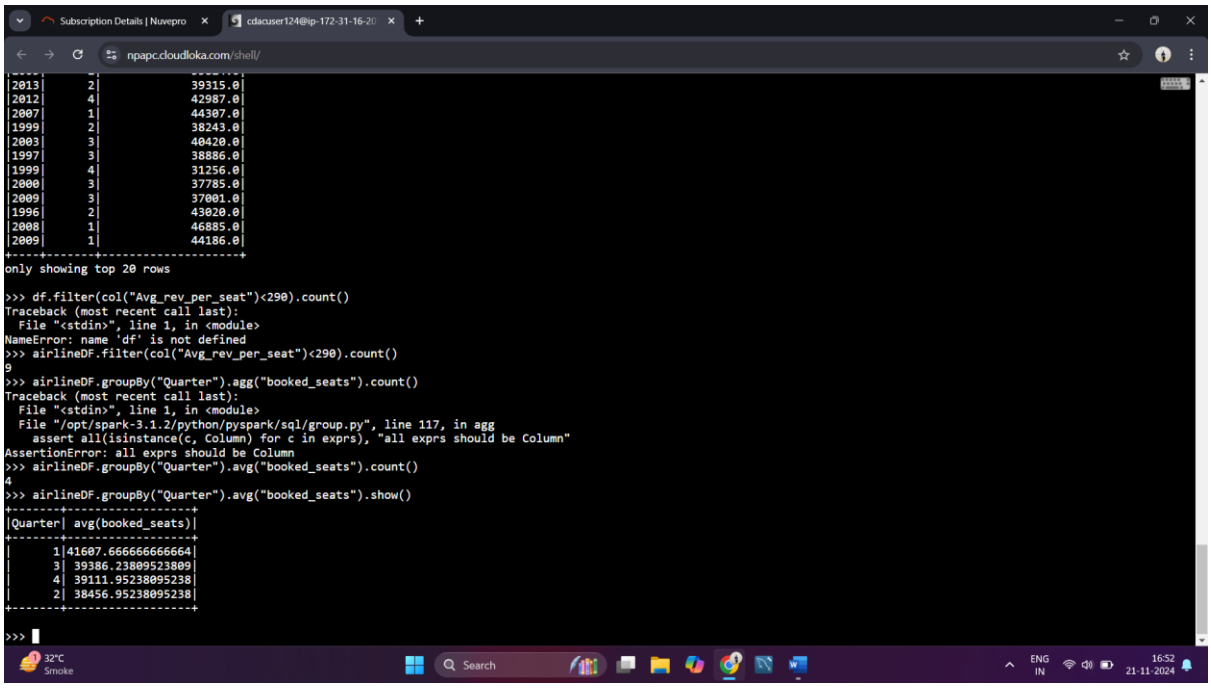


3.

→

airlineDF.groupBy("Quarter").avg("booked_seats").count()

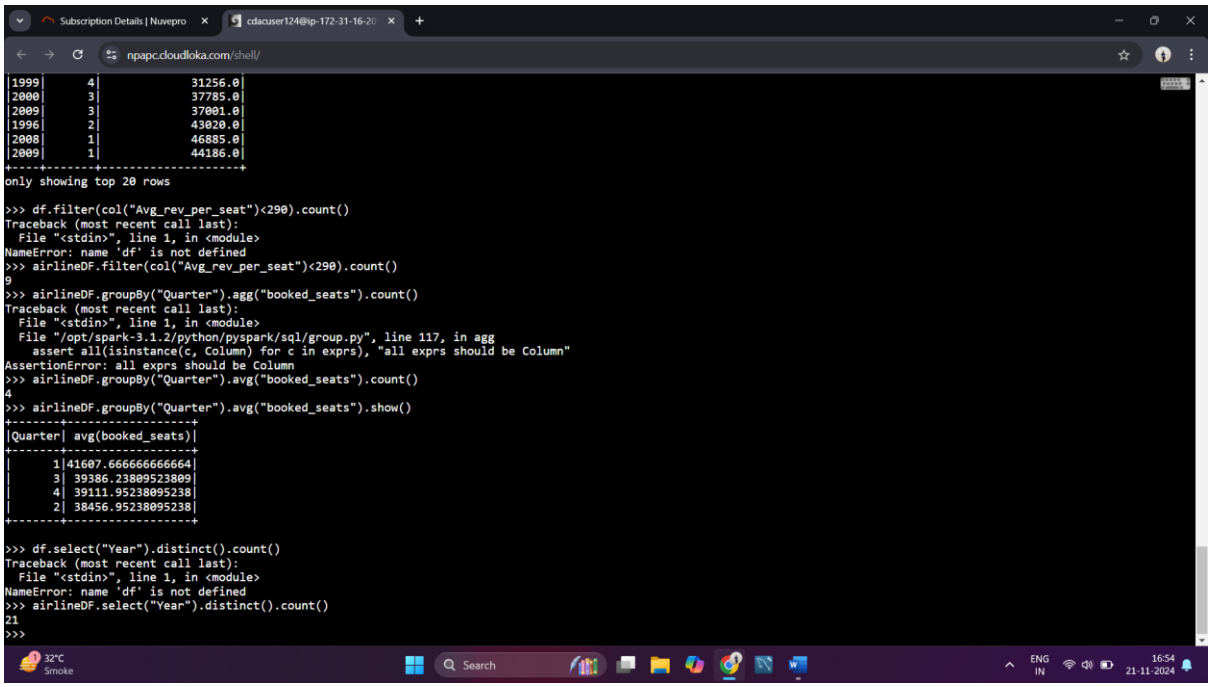
airlineDF.groupBy("Quarter").avg("booked_seats").show()



4.

→

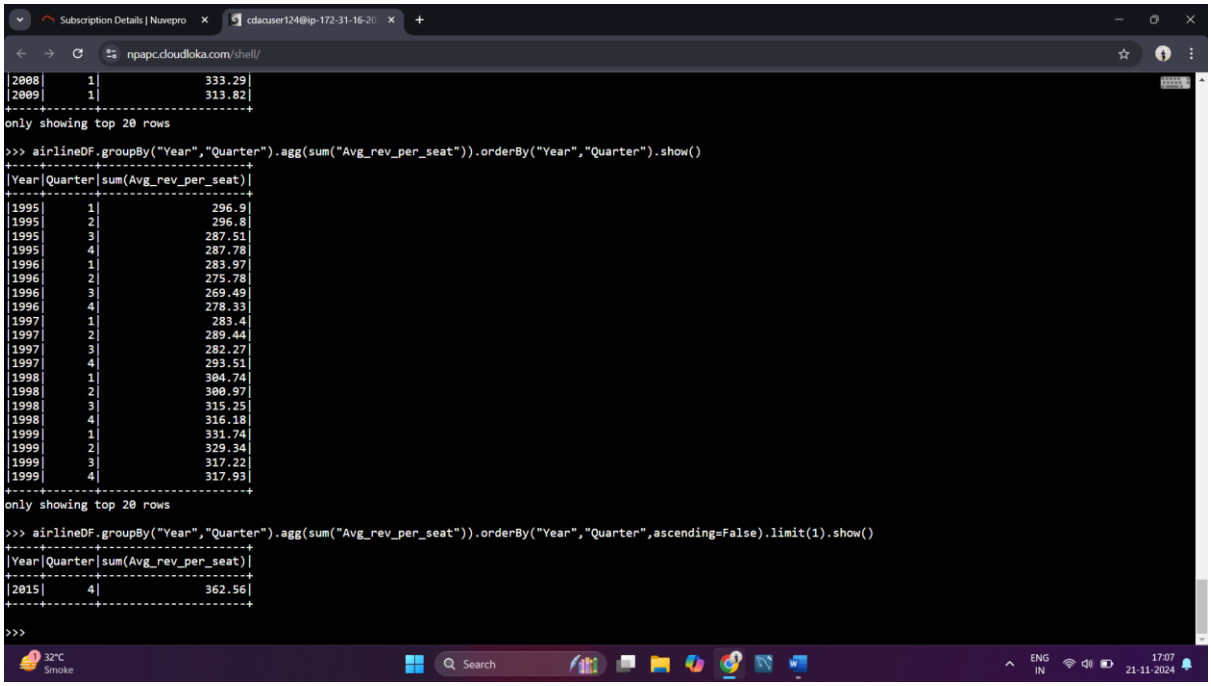
airlineDF.select("Year").distinct().count()



5.

→

airlineDF.groupBy("Year","Quarter").agg(sum("Avg_rev_per_seat")).orderBy("Year","Quarter",ascending=False).limit(1).show()



Hive

Q1.

1.

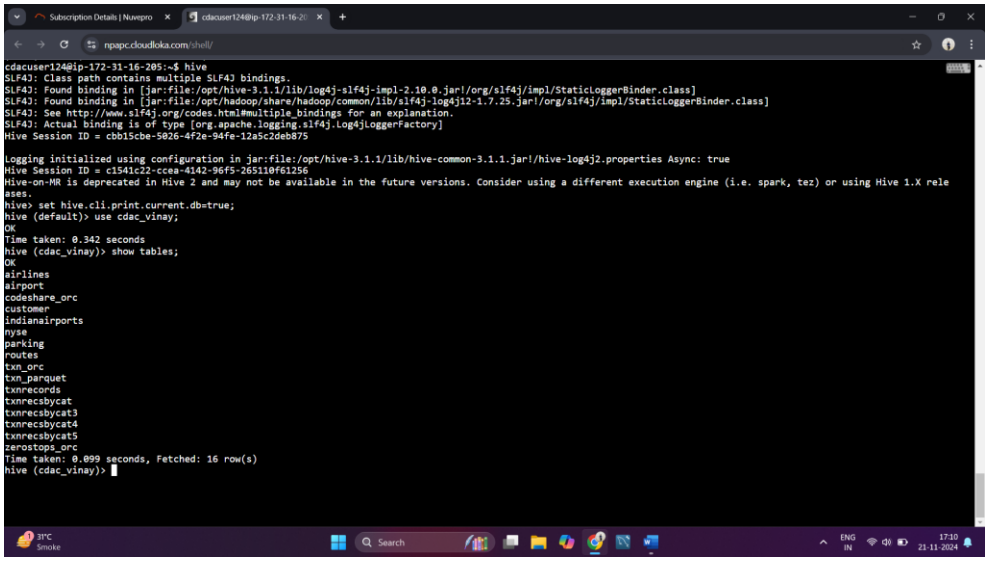
→

hive

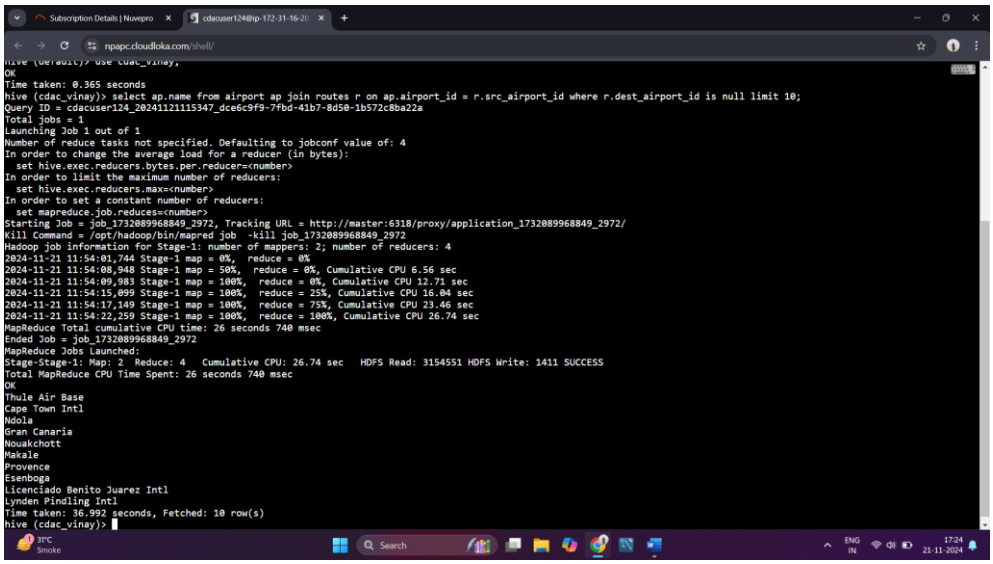
set hive.cli.print.current.db=true;

use cdac_vinay;

show tables;



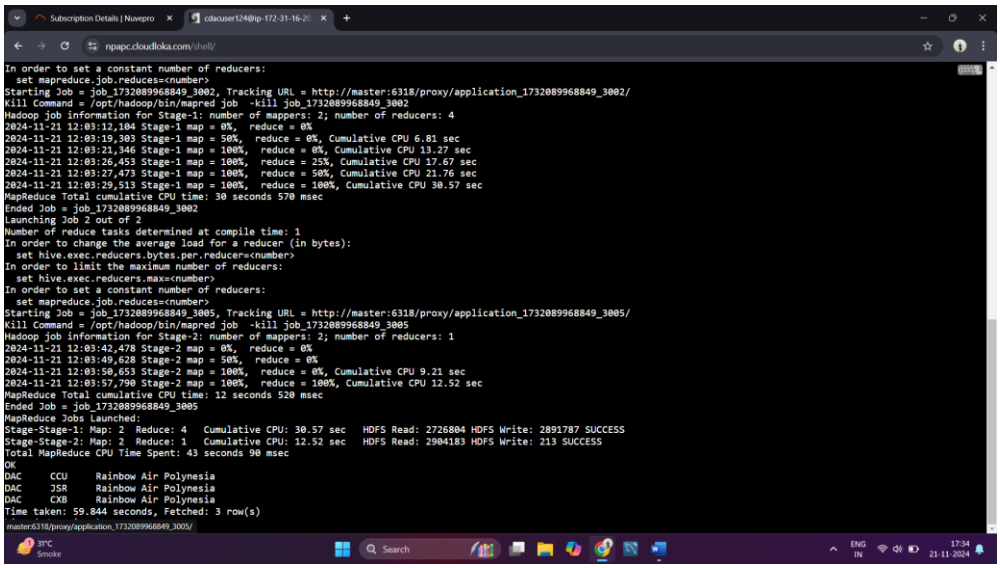
select ap.name from airport ap join routes r on ap.airport_id = r.src_airport_id where r.dest_airport_id is null limit 10;



2.

→

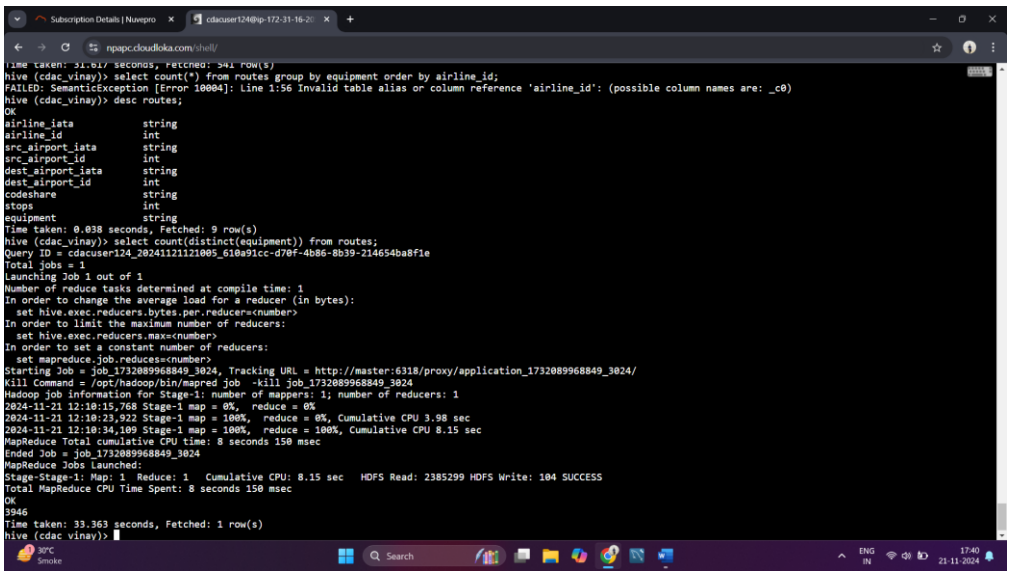
select r.src_airport_iata, r.dest_airport_iata, a.name from routes r join airlines a on r.airline_id = a.airline order by r.airline_id desc limit 3;



3.

→

select count(distinct(equipment)) from routes;

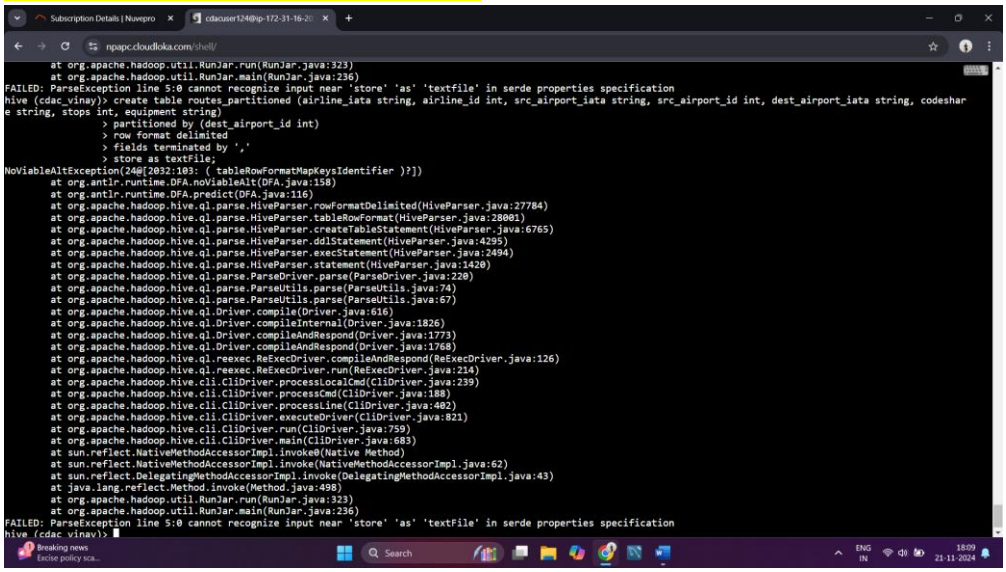


Q2.

1.

→

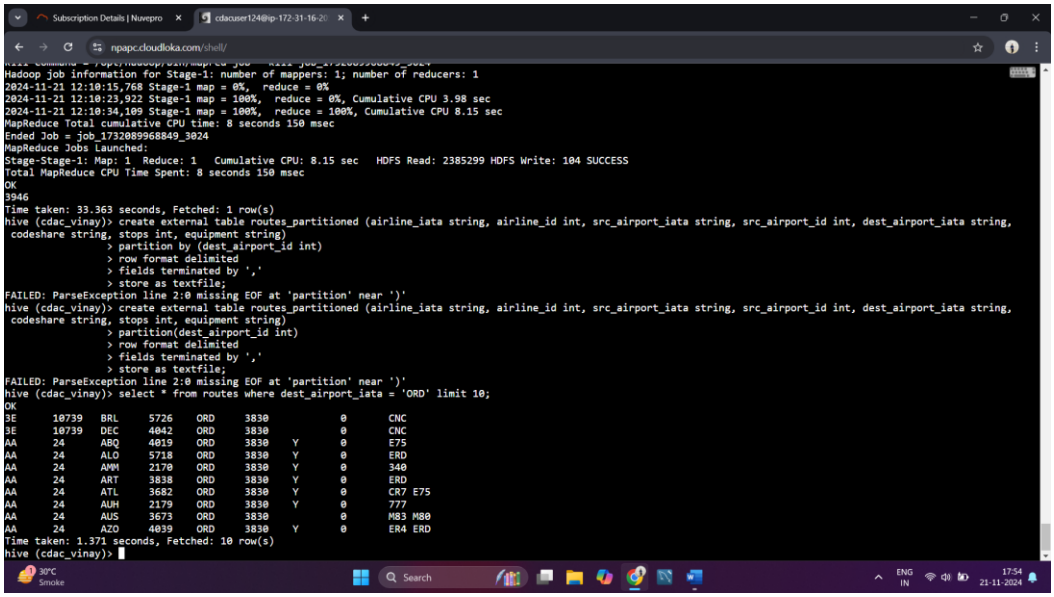
```
create table routes_partitioned (airline_iata string, airline_id int, src_airport_iata string,
src_airport_id int, dest_airport_iata string, codeshare string, stops int, equipment string)

partitioned by (dest_airport_id int)

 row format delimited

fields terminated by ','

store as textFile;

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.exec.dynamic.partition=true;
```



3.

→

```
select * from routes where dest_airport_iata = ‘ORD’ limit 10;
```


Subscription Details | Nuvepro
npapc.cloudloka.com/shell/
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:10:15,768 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:10:23,922 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.98 sec
2024-11-21 12:10:34,109 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.15 sec
MapReduce Total cumulative CPU time: 8 seconds 150 msec
Ended Job = job_1732089968849_3024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.15 sec HDFS Read: 2385299 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 150 msec
OK
3946
Time taken: 33.363 seconds, Fetched: 1 row(s)
hive (cdac_vinay)> create external 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)
> partition by (dest_airport_id int)
> row format delimited
> fields terminated by ','
> store as textfile;
FAILED: ParseException line 2:0 missing EOF at 'partition' near ')'
hive (cdac_vinay)> create external 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)
> partition(dest_airport_id int)
> row format delimited
> fields terminated by ','
> store as textfile;
FAILED: ParseException line 2:0 missing EOF at 'partition' near ')'
hive (cdac_vinay)> select * from routes where dest_airport_iata = 'ORD' limit 10;
OK
3E 10739 BRL 5726 ORD 3830 0 CNC
3E 10739 DEC 4042 ORD 3830 0 CNC
AA 24 ABQ 4019 ORD 3830 Y 0 E75
AA 24 ALO 5718 ORD 3830 Y 0 ERD
AA 24 AMW 2170 ORD 3830 Y 0 340
AA 24 ART 3838 ORD 3830 Y 0 ERD
AA 24 ATL 3682 ORD 3830 Y 0 CR7 E75
AA 24 AUH 2179 ORD 3830 Y 0 777
AA 24 AUS 3673 ORD 3830 0 M83 M80
AA 24 AZO 4039 ORD 3830 Y 0 ER4 ERD
Time taken: 1.371 seconds, Fetched: 10 row(s)
hive (cdac_vinay)>
30°C
Smoke
Search
ENG
IN
17:54
21-11-2024

4.

→