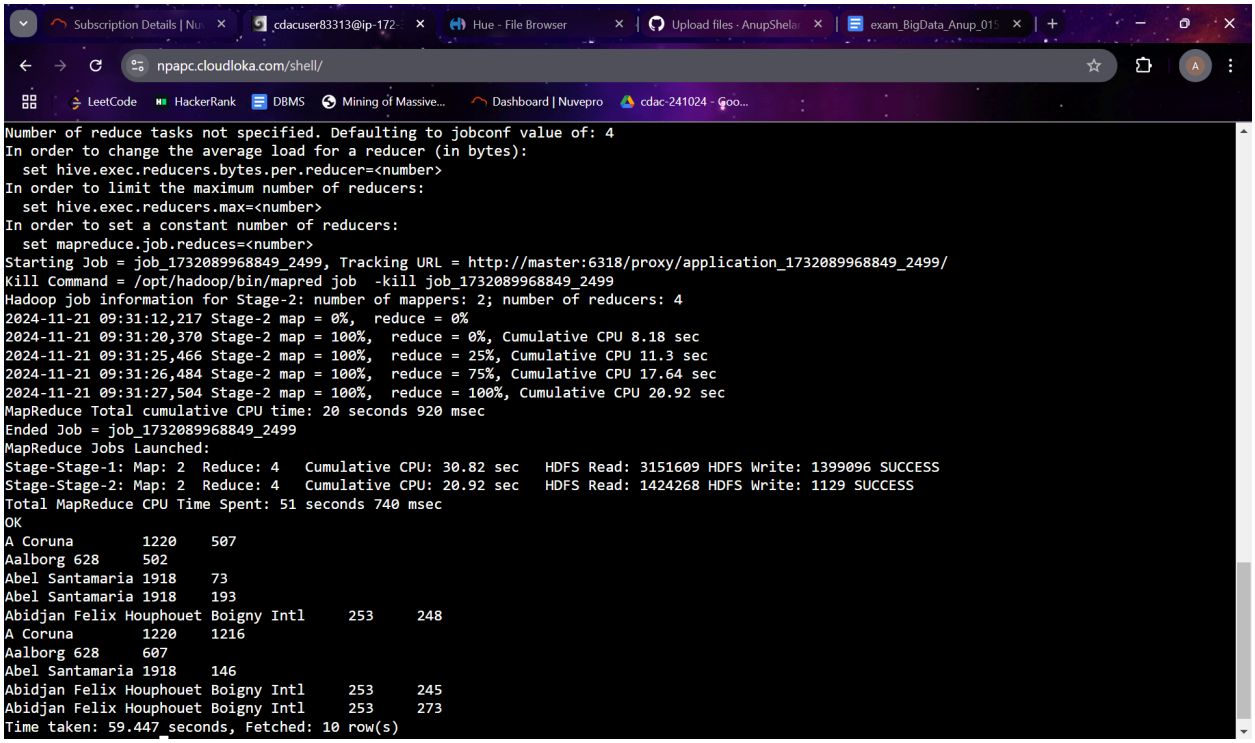
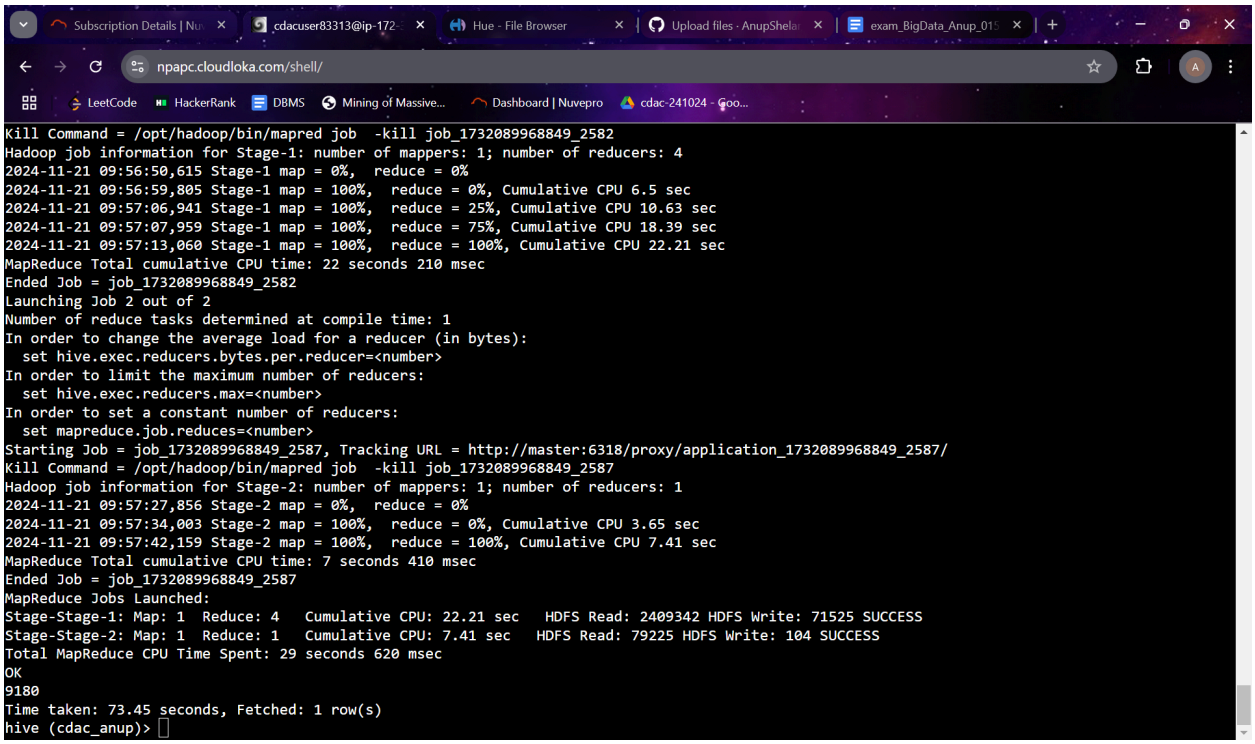
Hive

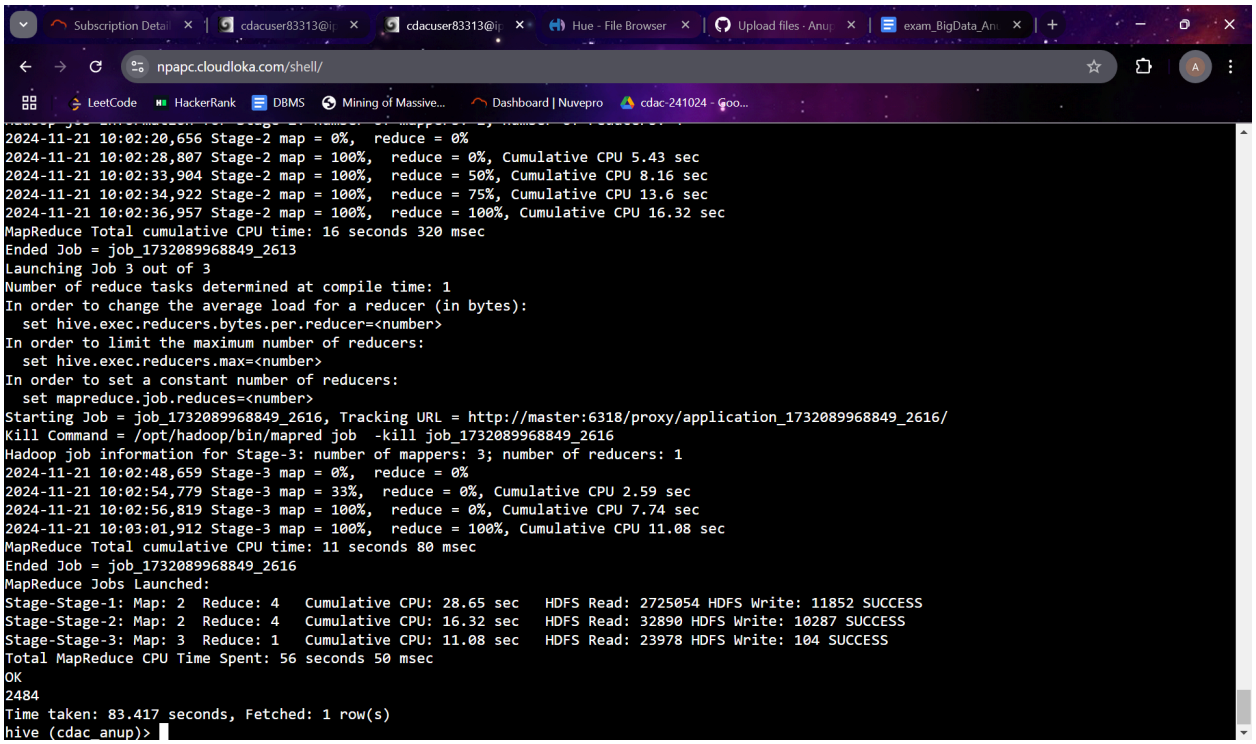
1) select a.name, r.src_airport_id, r.dest_airport_id from airport a join routes r on a.airport_id=r.src_airport_id limit 10;



2) select count(airline_id) as high from routes group by equipment order by high desc limit 1;



3) select count(r.airline_id) as high from routes r join airlines l on r.airline_id=l.airline_id group by(r.airline_id) order by high desc limit 1;



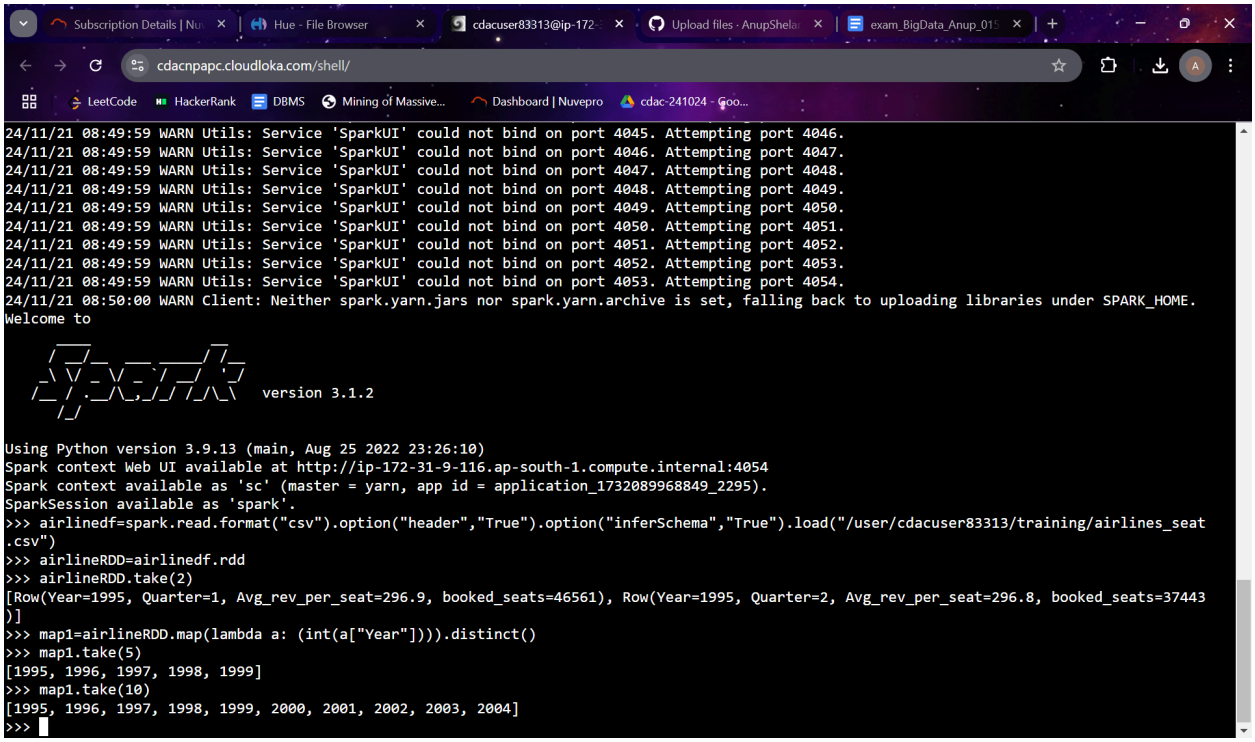
Question2:

1) create table routes23(airline_iata STRING, airline_id INT, src_airport_iata STRING, src_airport_id INT, dest_airport_iata STRING, dest_airport_id INT, codeshare STRING, stops INT, equipment STRING) partitioned by(src_airport_iata STRING) row format delimited by fields terminated by "," stored by textfile;

Pyspark

Question 1:

1)
2) Distinct year in airlines table



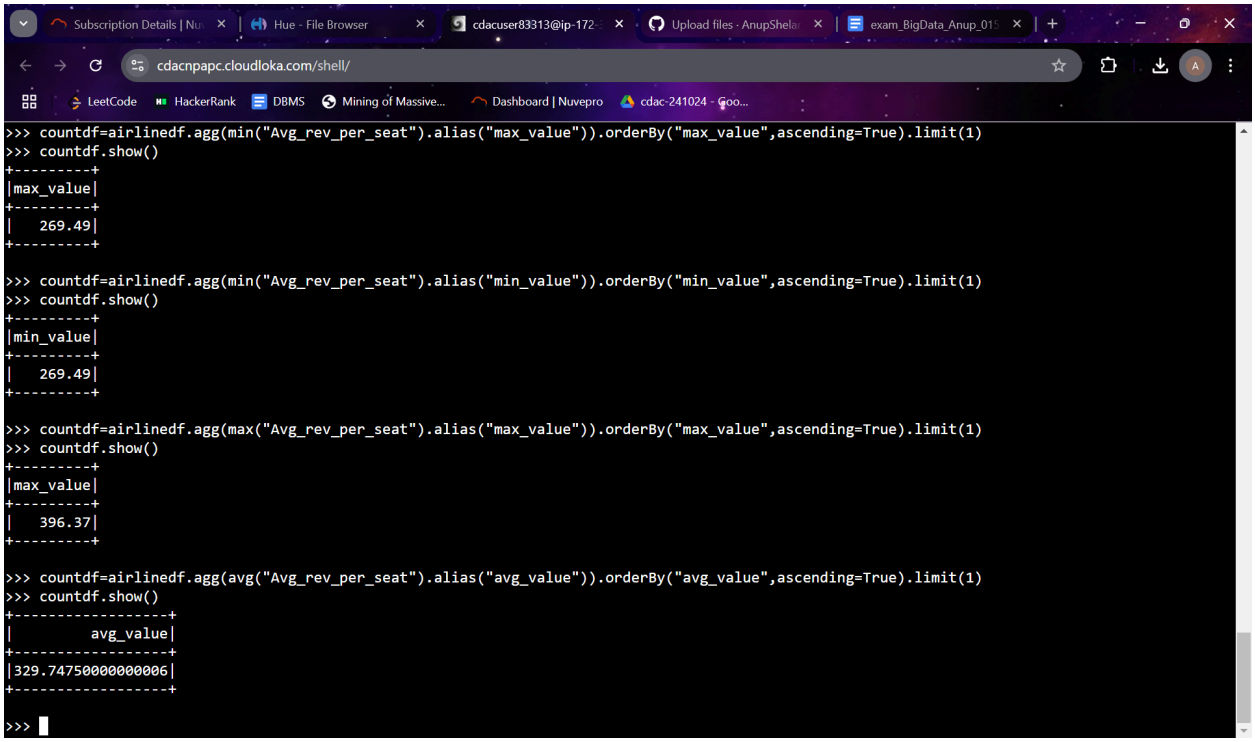
Question 2:

1)

1.Min value:
countdf=airlinedf.agg(min("Avg_rev_per_seat").alias("min_value")).orderBy("min_value", ascending=True).limit(1)

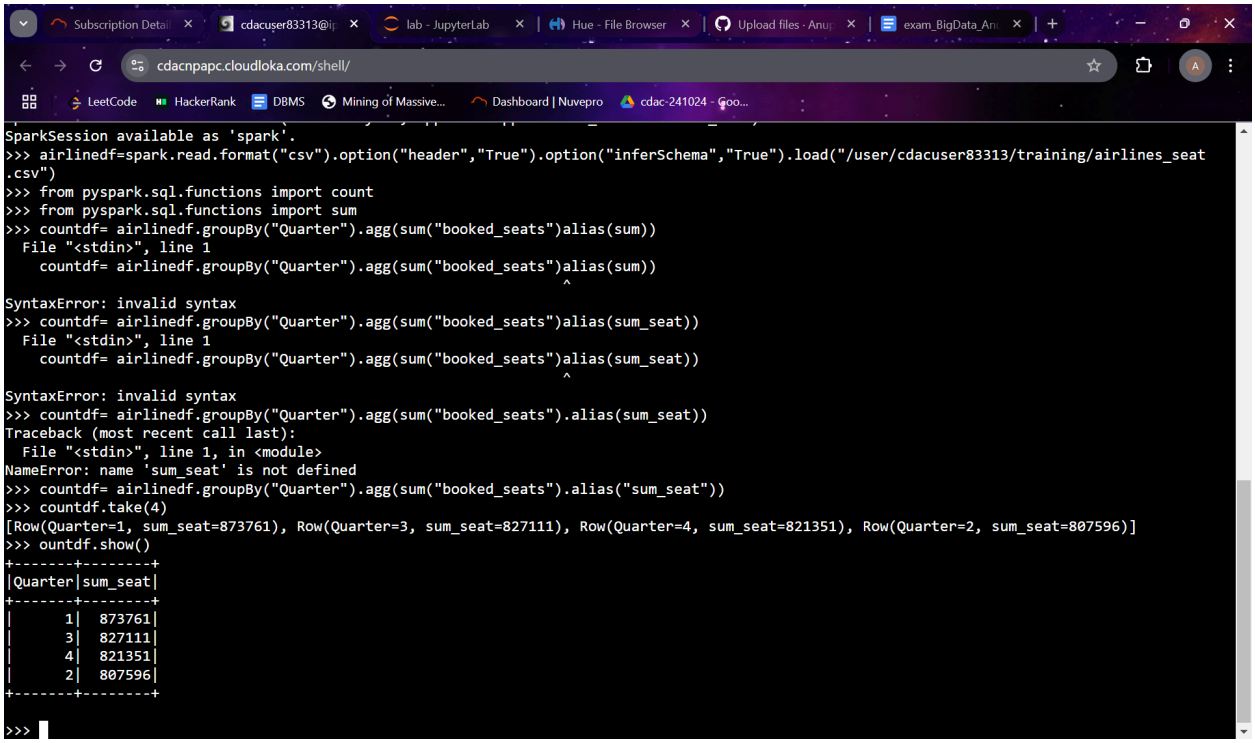
2.Maxvalue
countdf=airlinedf.agg(max("Avg_rev_per_seat").alias("max_value")).orderBy("max_value",ascending=True).limit(1)

3.avgvalue
countdf=airlinedf.agg(avg("Avg_rev_per_seat").alias("avg_value")).orderBy("avg_value", ascending=True).limit(1)

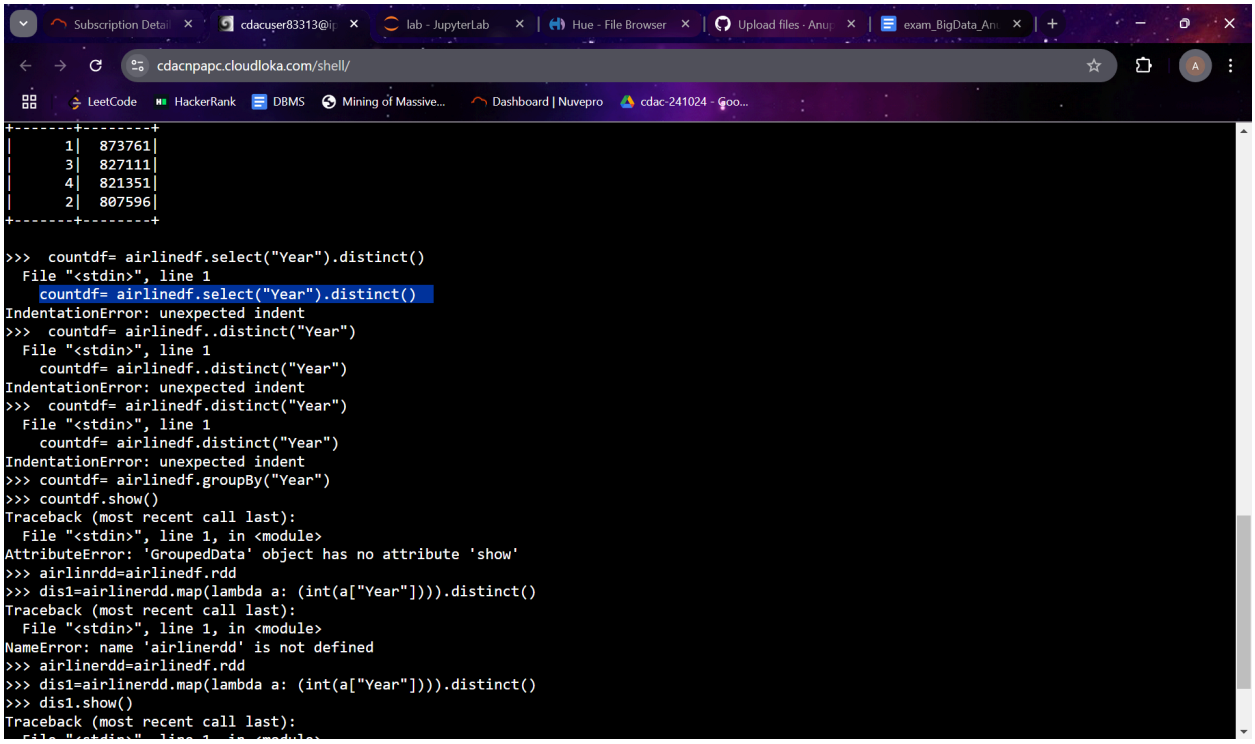


2)

3) countdf= airlinedf.groupBy("Quarter").agg(sum("booked_seats").alias("sum_seat"))



4) countdf= airlinedf.select("Year").distinct()



5)

```
year_rev=airlinedf.groupBy("Year").agg(sum("Avg_rev_per_seat").
alias("total_rev")).orderBy("total_rev").limit(3)
```