Pranavsing Thakur

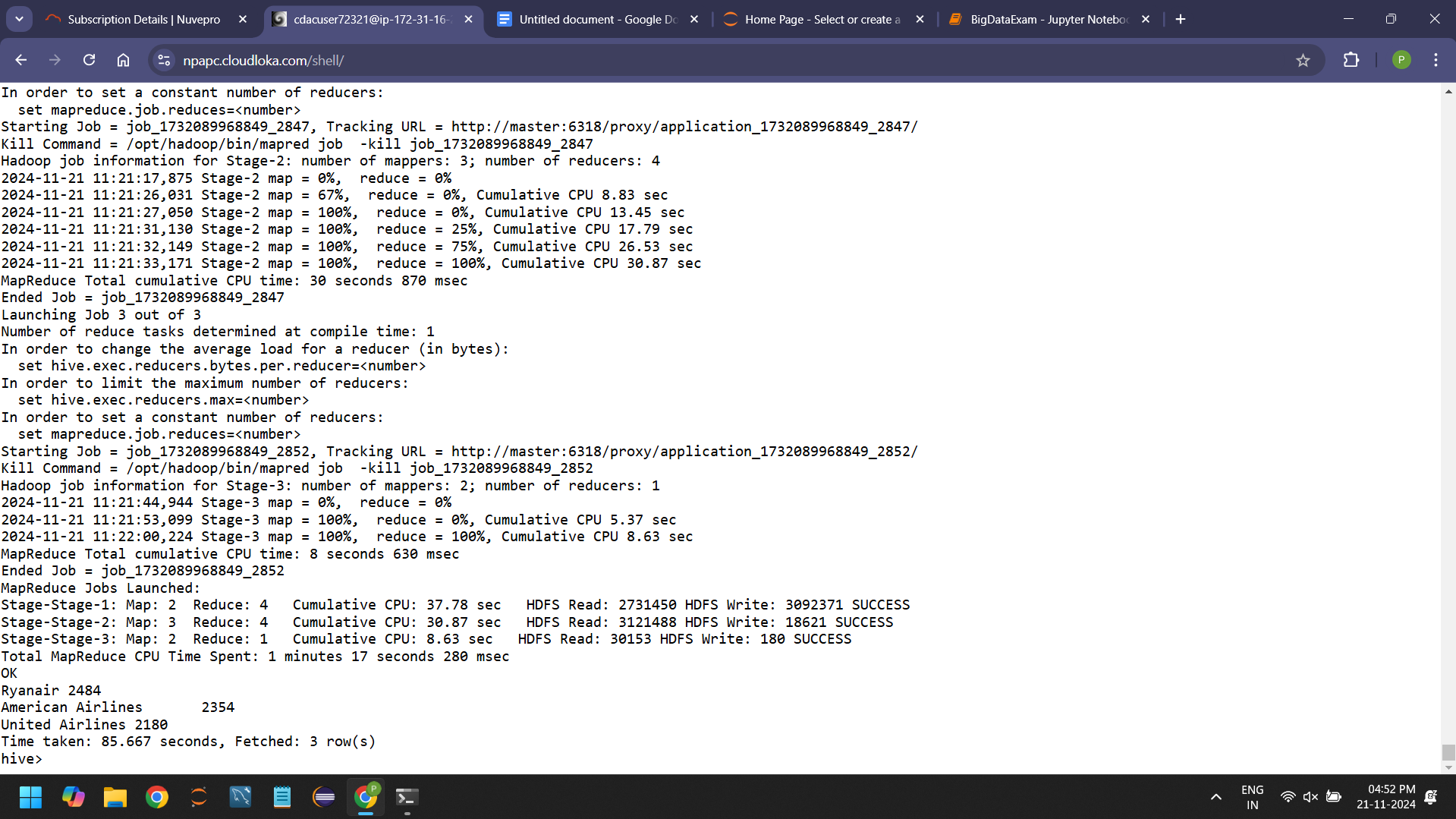
240840325037

Question 1:

2.

select ae.name , count(distinct(r.src\_airport\_id,'-',r.dest\_airport\_id)) from airline ae join route r on ae.airline\_id=r.airline\_id group by ae.name order by count(distinct(r.src\_airport\_id,'-',r.dest\_airport\_id)) desc limit 3;

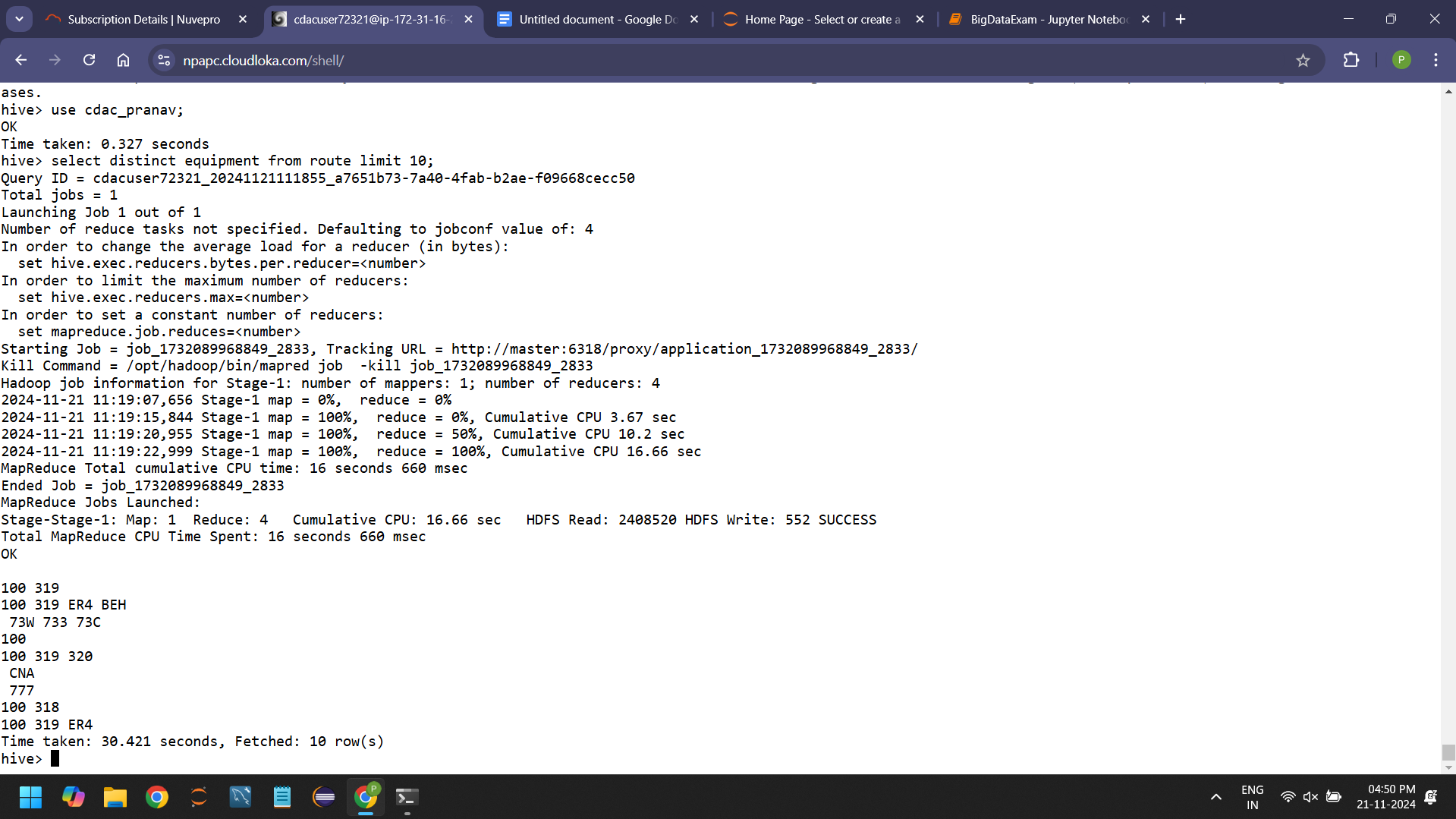




3.

select distinct equipment from route limit 10;





1.

Question 2:

1.

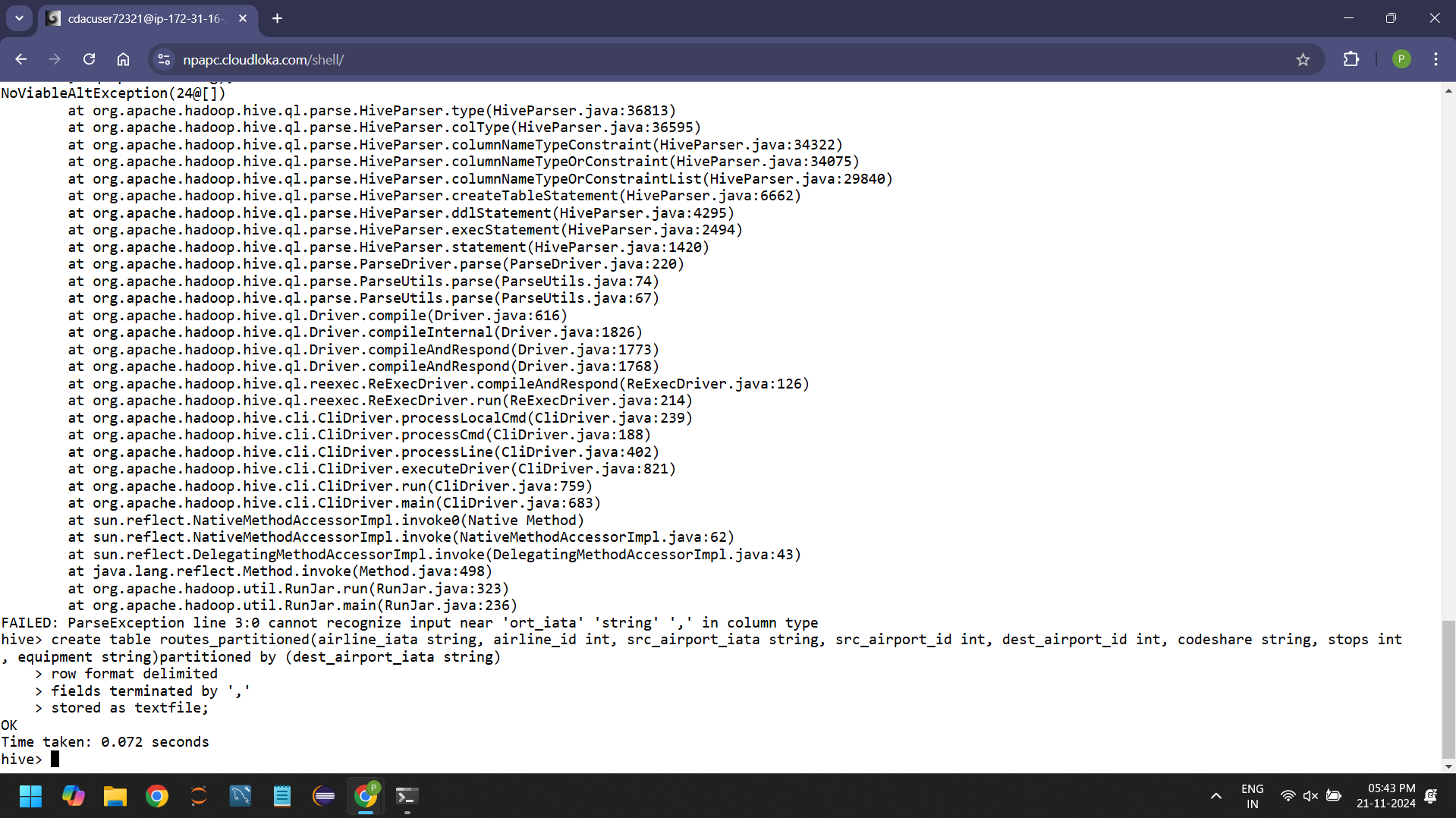
create table routes\_partitioned(airline\_iata string, airline\_id int, src\_airport\_iata string, src\_airport\_id int, dest\_airport\_id int, codeshare string, stops int, equipment string)

> partitioned by (dest\_airport\_iata string)

> row format delimited

> fields terminated by ','

> stored as textfile;



2.

insert overwrite table routes\_partitioned partition(dest\_airport\_id) select a.airline\_iata, a.airline\_id, a.src\_airport\_iata, a.src\_airport\_id, a.dest\_airport\_id, a.codeshare, a.stops, a.equipment, a.dest\_airport\_id from route a distribute by dest\_airport\_id;

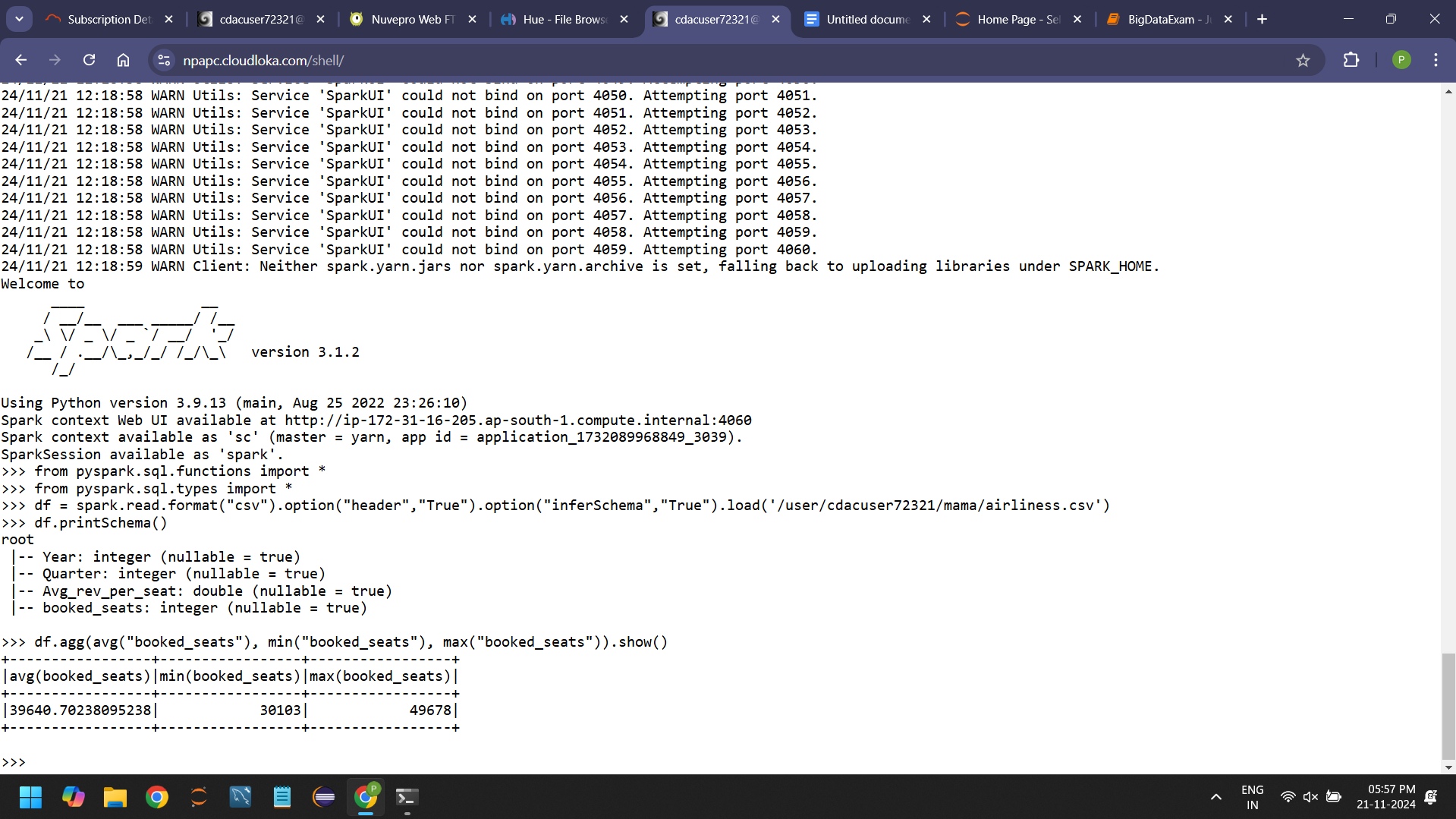
Cant upload screenshot because processing is taking to much time hive is slow

SPARK

Question 1:

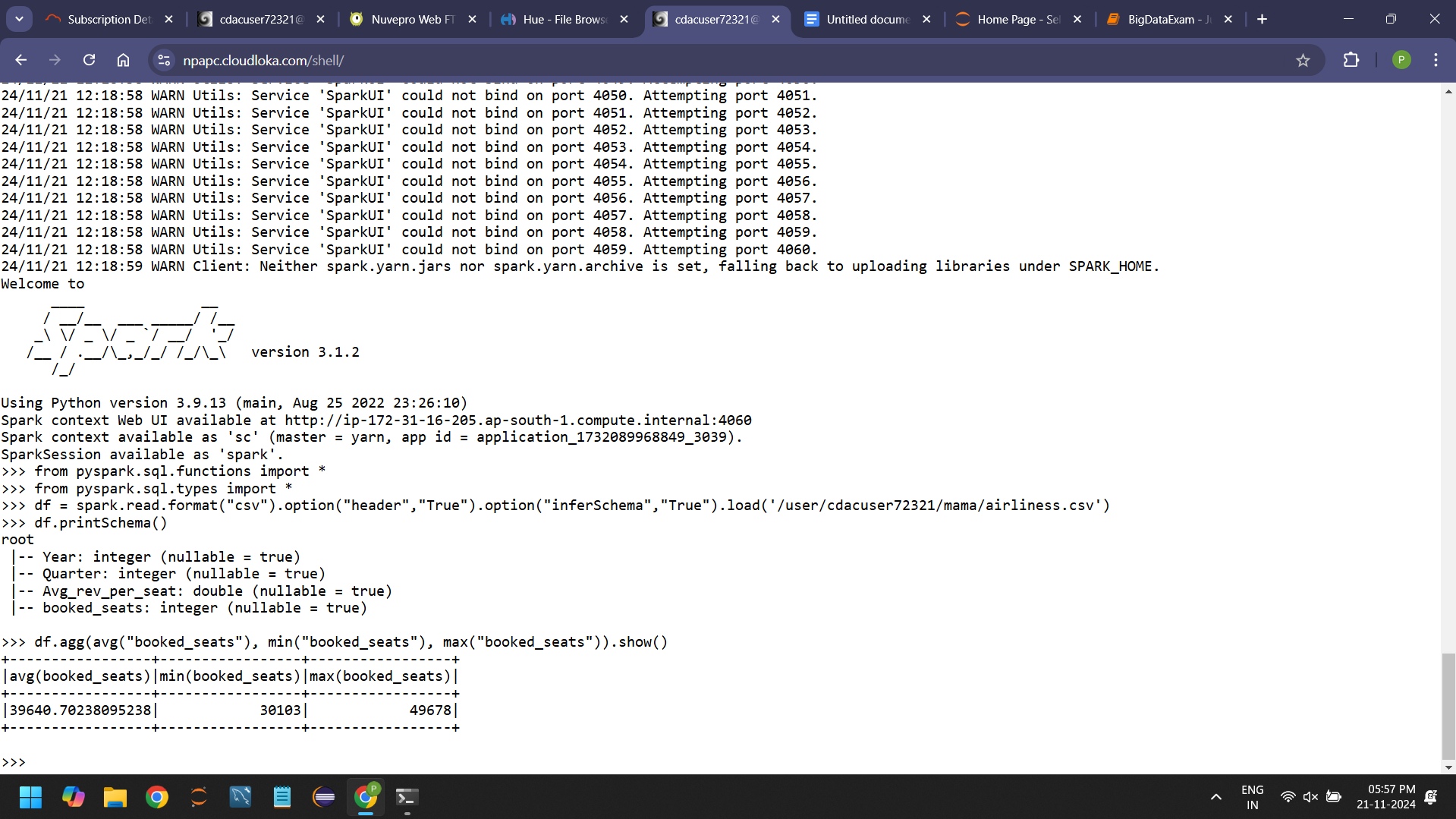
1.

df.agg(avg("booked\_seats"),min("booked\_seats"),max("booked\_seats")).show()



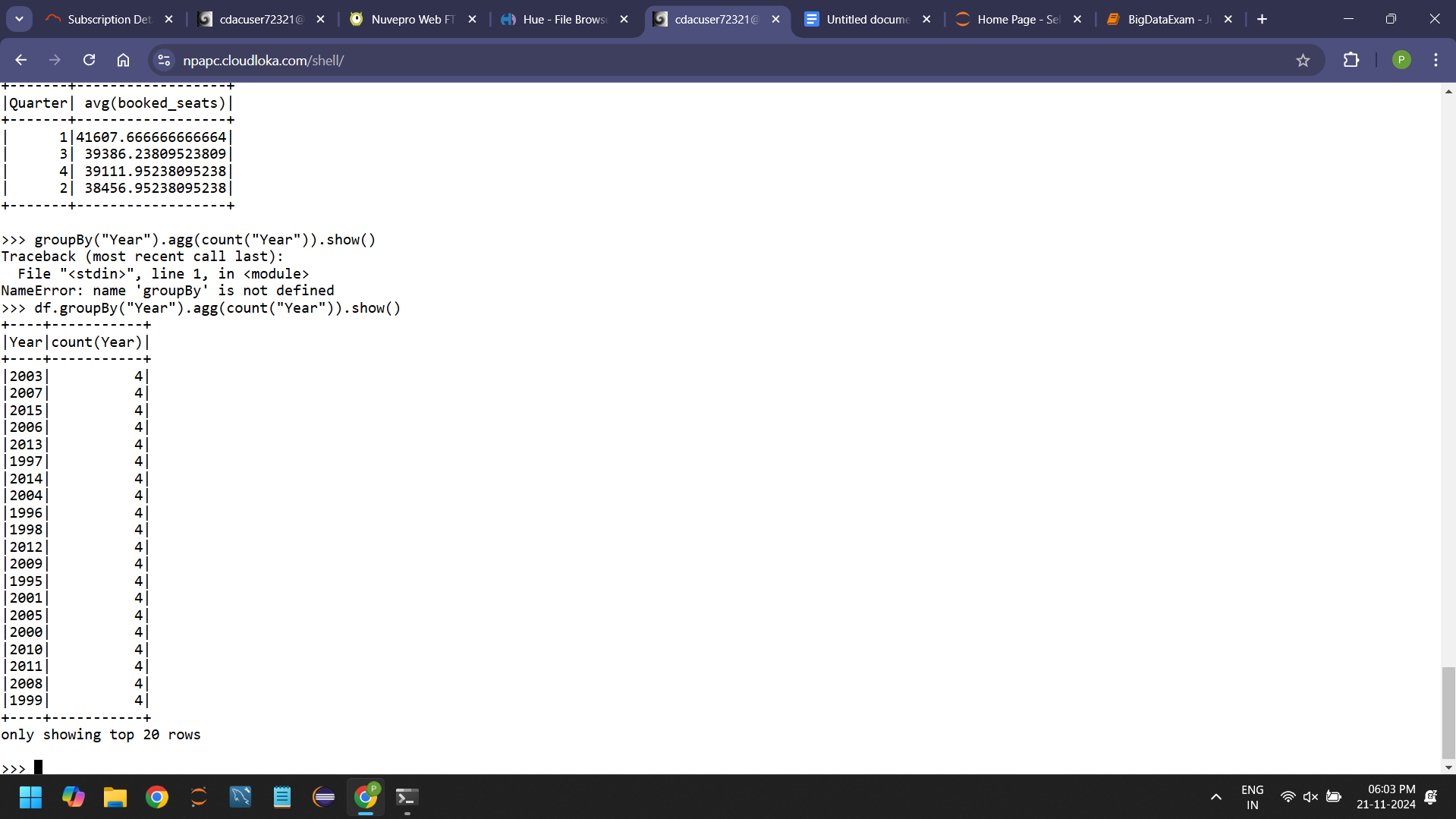
3.

df.groupBy("Quarter").avg("booked\_seats").show()



4.

df.groupBy("Year").agg(count("Year")).show()



5.

df.groupBy("Year","Quarter").agg(sum(col("avg\_rev\_per\_seat")\*col("booked\_seats")).alias("total\_revenue")).orderBy("total\_revenue",ascending=True).limit(1).show()

