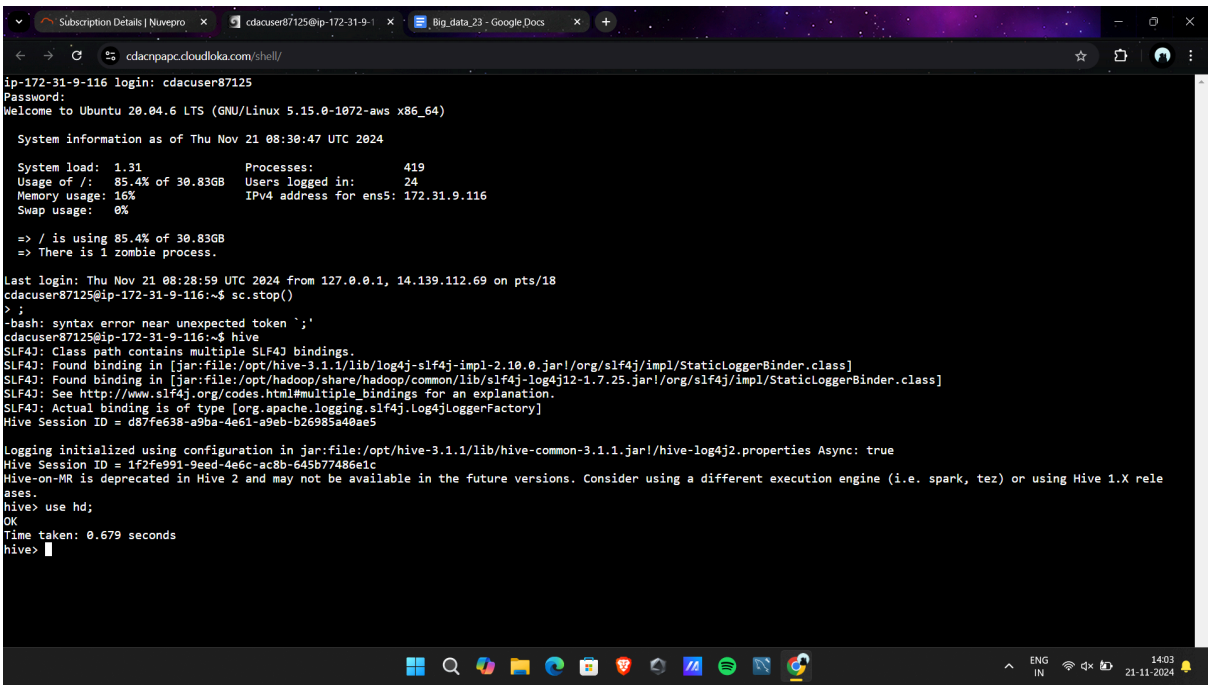
Hive

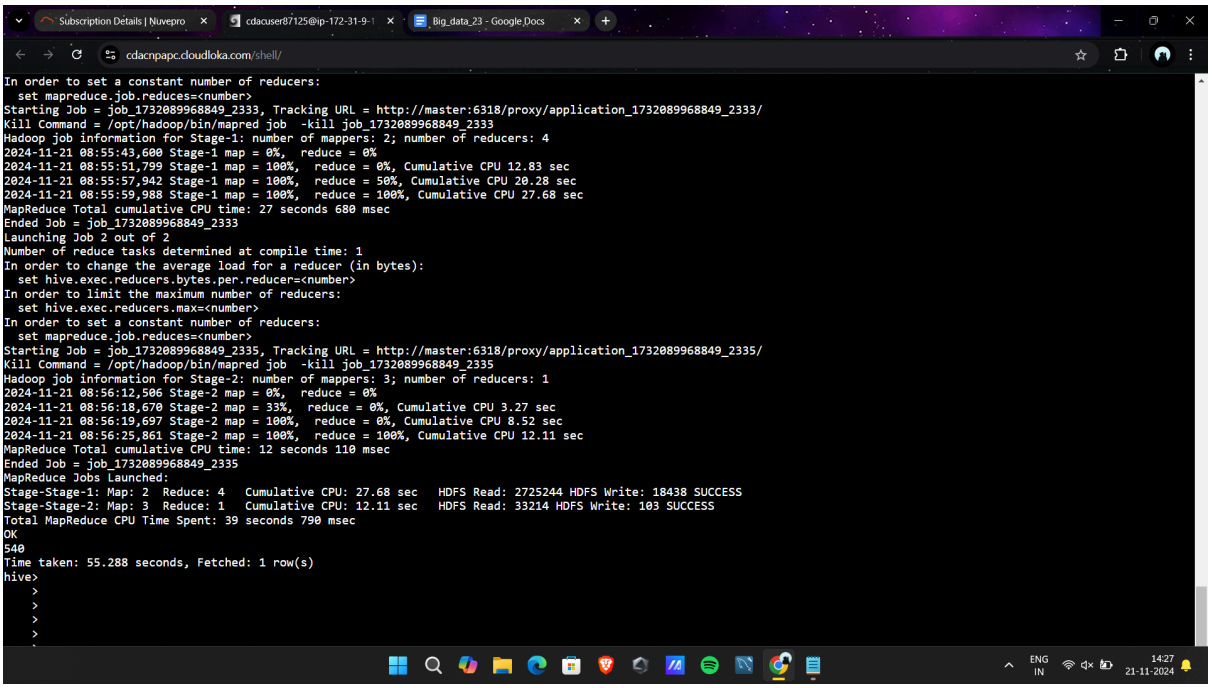
1.


Subscription Details | Nuvepro
cdacuser87125@ip-172-31-9-1
Big_data_23 - Google Docs
cdacnpapc.cloudloka.com/shell/
ip-172-31-9-116 login: cdacuser87125
Password:
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-1072-aws x86_64)
System information as of Thu Nov 21 08:30:47 UTC 2024
System load: 1.31 Processes: 419
Usage of /: 85.4% of 30.83GB Users logged in: 24
Memory usage: 16% IPv4 address for ens5: 172.31.9.116
Swap usage: 0%
=> / is using 85.4% of 30.83GB
=> There is 1 zombie process.
Last login: Thu Nov 21 08:28:59 UTC 2024 from 127.0.0.1, 14.139.112.69 on pts/18
cdacuser87125@ip-172-31-9-116:~$ sc.stop()
> ;
-bash: syntax error near unexpected token `;'
cdacuser87125@ip-172-31-9-116:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive-3.1.1/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = d87fe638-a9ba-4e61-a9eb-b26985a40ae5
Logging initialized using configuration in jar:file:/opt/hive-3.1.1/lib/hive-common-3.1.1.jar!/hive-log4j2.properties Async: true
Hive Session ID = 1f2fe991-9eed-4e6c-ac8b-645b77486e1c
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
hive> use hd;
OK
Time taken: 0.679 seconds
hive>
ENG
IN
14:03
21-11-2024

2.

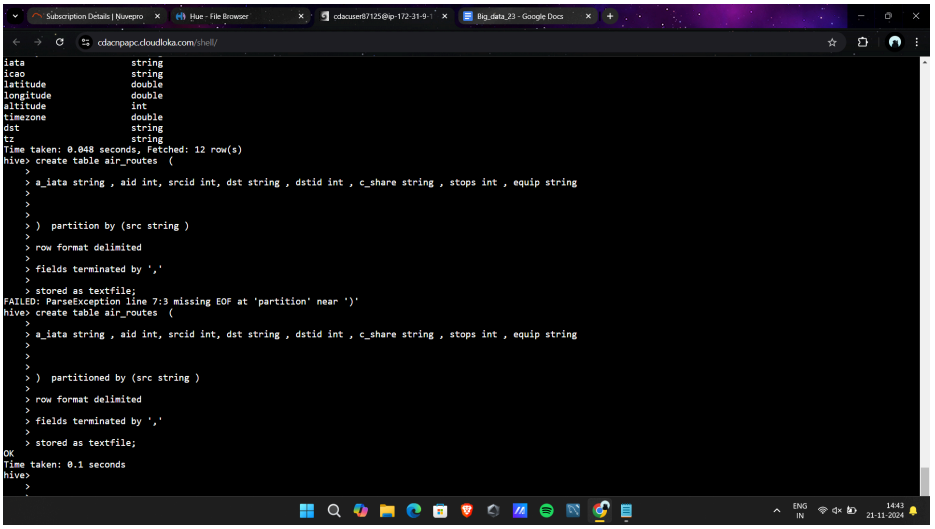
3

select count(distinct(a.name)) from airlines a
join routes r on a.a_id =r.a_id ;



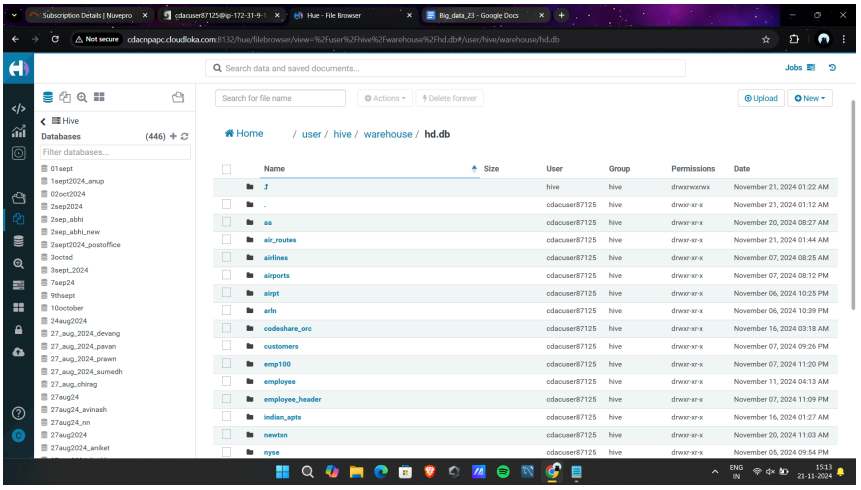
Question2

1.

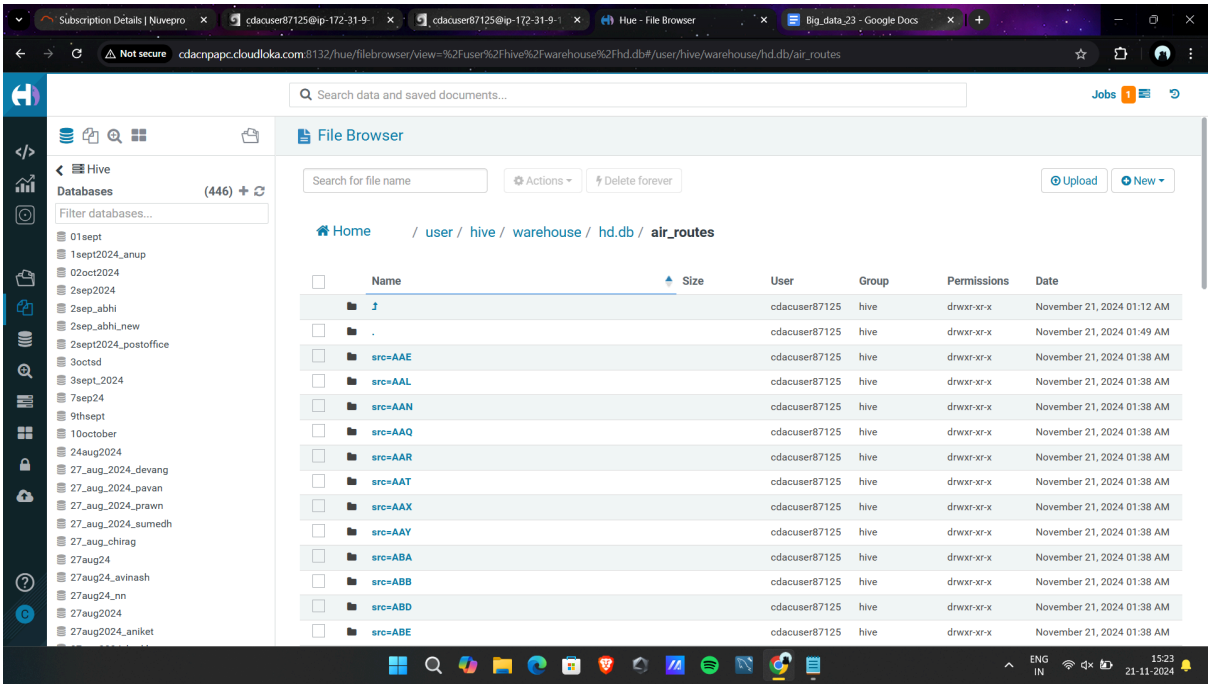


create table air_routes (
a_iata string , aid int, srcid int, dst string , dstid int , c_share string , stops int , equip string

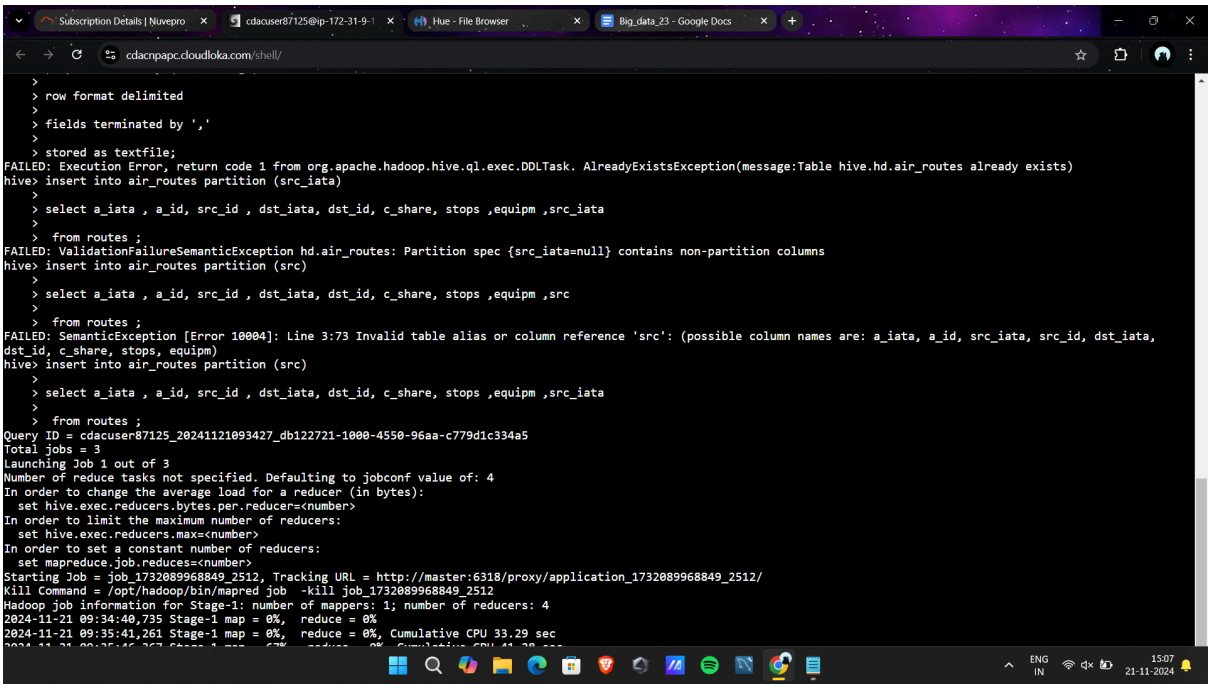
) partitioned by (src string)
row format delimited
fields terminated by ','
stored as textfile;



insert into air_routes partition (src)
select a_iata , a_id, src_id , dst_iata, dst_id, c_share, stops ,equipm ,src_iata
 from routes ;



Hive lagged



2.

```
insert into air_routes select * from routes where src_iata="JFK" ;
```

3.

```
Select  src , dst from air_routes where src='LAX' limit 10;
```

4.

```
explain air_routes;
```

Spark

Question 1

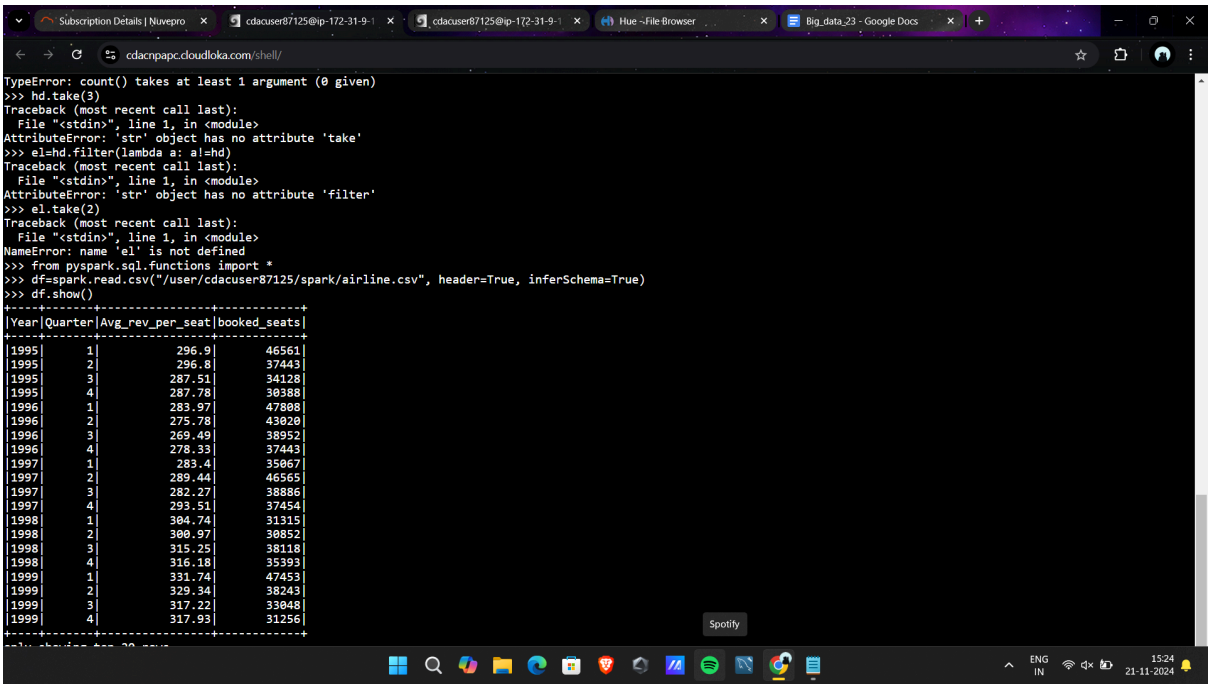
1.

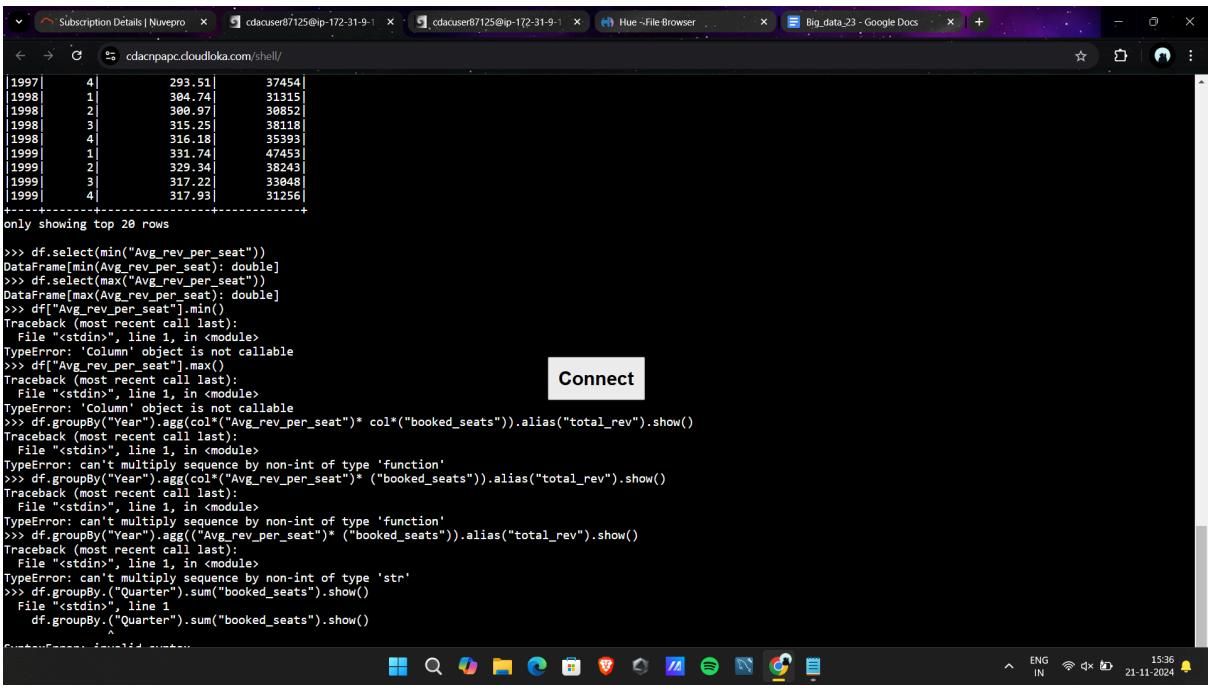
```
aird=sc.textfile("/user/cdacuser87125/spark/airline.csv")
aird.count()

header=aird.first()
el=aird.filter(lambda a :a != header)

split=el.map(lambda a:a.split(","))

aa=split.map(lambda a: str(a[0])+" "+ a[1], float(a[2]) )

arrange=aa.reduceByKey( lambda a,b :a+b)

main=aa.sortByKey(lambda a: a[0])
```

Question 2

1.

```
>>> from pyspark.sql.functions import *
>>>
df=spark.read.csv("/user/cdacuser87125/spark/airline.csv",
header=True, inferSchema=True)
```





2.

df.filter(“Avg_rev_per_seat”)>290
df.agg(count(“Avg_rev_per_seat”))

3.

df.groupBy(“Year”).sum(“booked_seats)

4.

df.select(“Year”).distinct()