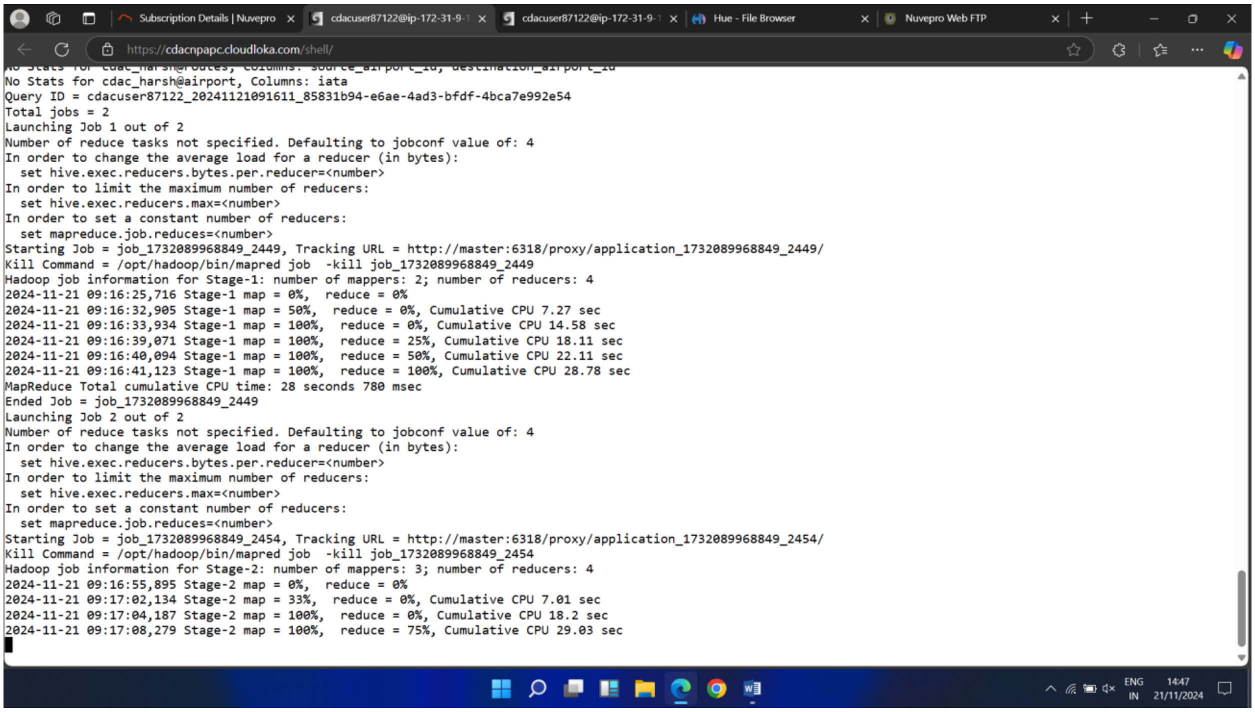
Q1.

1.

```
select src.name from airport src  join routes r on src.iata =
r.source_airport_id join airport dest on
r.destination_airport_id=dest.iata wher
e src.iata=dest.iata limit 10;
```



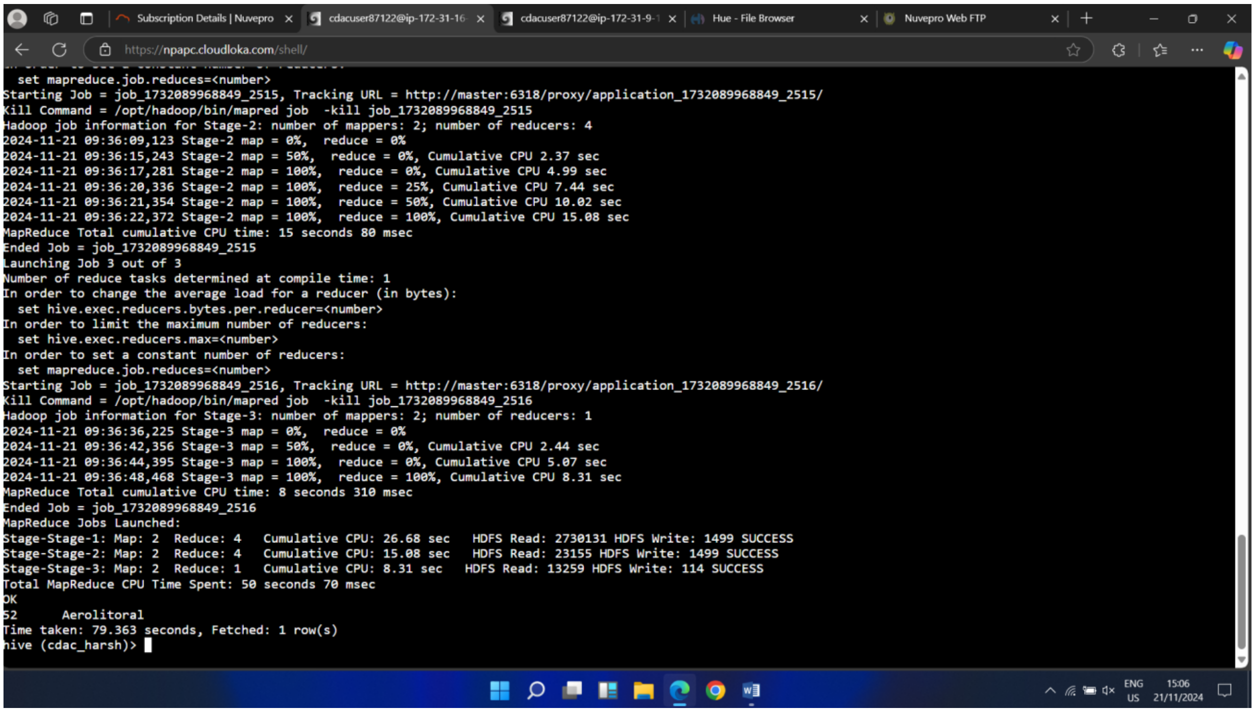
Subscription Details | Nuvepro
cdacuser87122@ip-172-31-9-1
cdacuser87122@ip-172-31-9-1
Hue - File Browser
Nuvepro Web FTP
https://cdacnpapc.cloudloka.com/shell/
ENG
IN
14:47
21/11/2024

```
No Stats for cdac_harsh@routes, Columns: source_airport_id, destination_airport_id
No Stats for cdac_harsh@airport, Columns: iata
Query ID = cdacuser87122_20241121091611_85831b94-e6ae-4ad3-bfdf-4bca7e992e54
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_2449, Tracking URL = http://master:6318/proxy/application_1732089968849_2449/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2449
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 09:16:25,716 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:16:32,905 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 7.27 sec
2024-11-21 09:16:33,934 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.58 sec
2024-11-21 09:16:39,071 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 18.11 sec
2024-11-21 09:16:40,094 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 22.11 sec
2024-11-21 09:16:41,123 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.78 sec
MapReduce Total cumulative CPU time: 28 seconds 780 msec
Ended Job = job_1732089968849_2449
Launching Job 2 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_2454, Tracking URL = http://master:6318/proxy/application_1732089968849_2454/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2454
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 4
2024-11-21 09:16:55,895 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:17:02,134 Stage-2 map = 33%,  reduce = 0%, Cumulative CPU 7.01 sec
2024-11-21 09:17:04,187 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 18.2 sec
2024-11-21 09:17:08,279 Stage-2 map = 100%,  reduce = 75%, Cumulative CPU 29.03 sec
```

Q1.

3.

```
select  count(*) as no_of_count , al.name  from routes r  join
airline al  on al.iata = r.source_airport_id   group by al.name
order by no_of_
count desc limit 1;
```

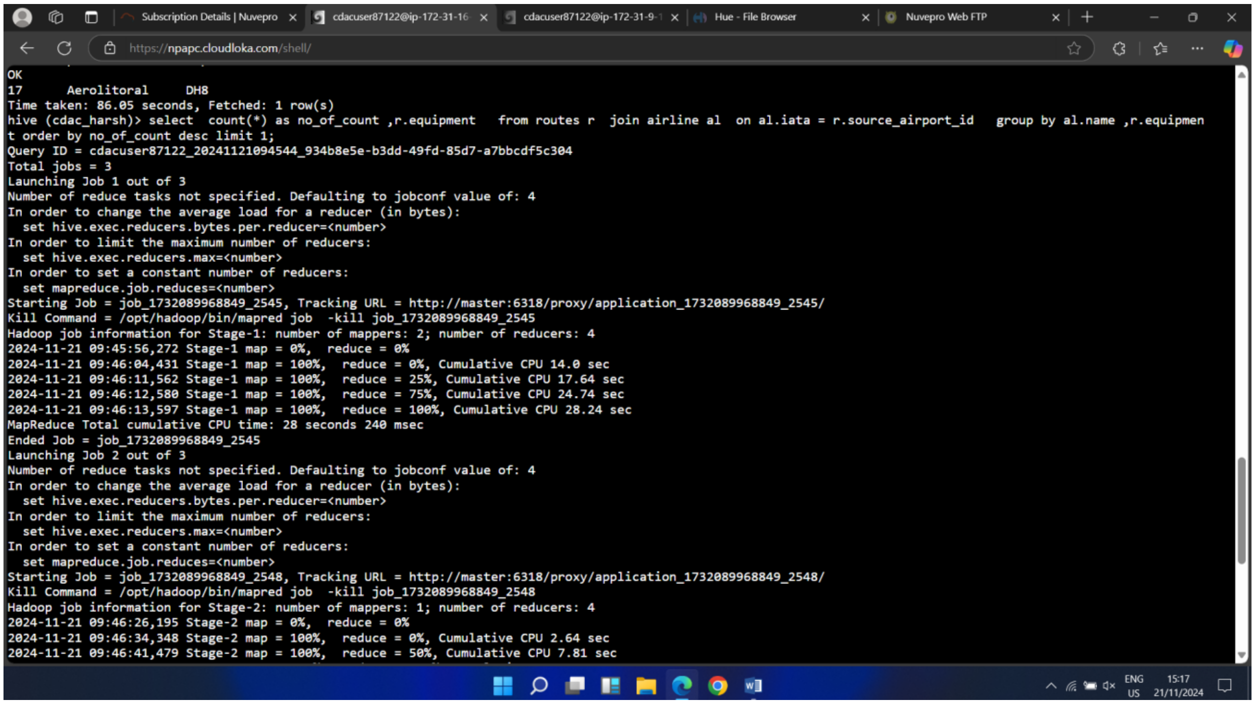

Subscription Details | Nuvepro
cdacuser87122@ip-172-31-16-
cdacuser87122@ip-172-31-9-1
Hue - File Browser
Nuvepro Web FTP
https://npapc.cloudloka.com/shell/
ENG
US
15:06
21/11/2024

```
  set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2515, Tracking URL = http://master:6318/proxy/application_1732089968849_2515/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2515
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 4
2024-11-21 09:36:09,123 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:36:15,243 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.37 sec
2024-11-21 09:36:17,281 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.99 sec
2024-11-21 09:36:20,336 Stage-2 map = 100%,  reduce = 25%, Cumulative CPU 7.44 sec
2024-11-21 09:36:21,354 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 10.02 sec
2024-11-21 09:36:22,372 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 15.08 sec
MapReduce Total cumulative CPU time: 15 seconds 80 msec
Ended Job = job_1732089968849_2515
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_2516, Tracking URL = http://master:6318/proxy/application_1732089968849_2516/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2516
Hadoop job information for Stage-3: number of mappers: 2; number of reducers: 1
2024-11-21 09:36:36,225 Stage-3 map = 0%,  reduce = 0%
2024-11-21 09:36:42,356 Stage-3 map = 50%,  reduce = 0%, Cumulative CPU 2.44 sec
2024-11-21 09:36:44,395 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.07 sec
2024-11-21 09:36:48,468 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 8.31 sec
MapReduce Total cumulative CPU time: 8 seconds 310 msec
Ended Job = job_1732089968849_2516
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 26.68 sec   HDFS Read: 2730131 HDFS Write: 1499 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 15.08 sec   HDFS Read: 23155 HDFS Write: 1499 SUCCESS
Stage-Stage-3: Map: 2  Reduce: 1   Cumulative CPU: 8.31 sec   HDFS Read: 13259 HDFS Write: 114 SUCCESS
Total MapReduce CPU Time Spent: 50 seconds 70 msec
OK
52      Aerolitoral
Time taken: 79.363 seconds, Fetched: 1 row(s)
hive (cdac_harsh)> 
```

Q1.

2.[redacted]select count(*) as no_of_count ,r.equipment   from routes r join airline al on al.iata = r.source_airport_id   group by al.name ,r

.equipment order by no_of_count desc limit 1;

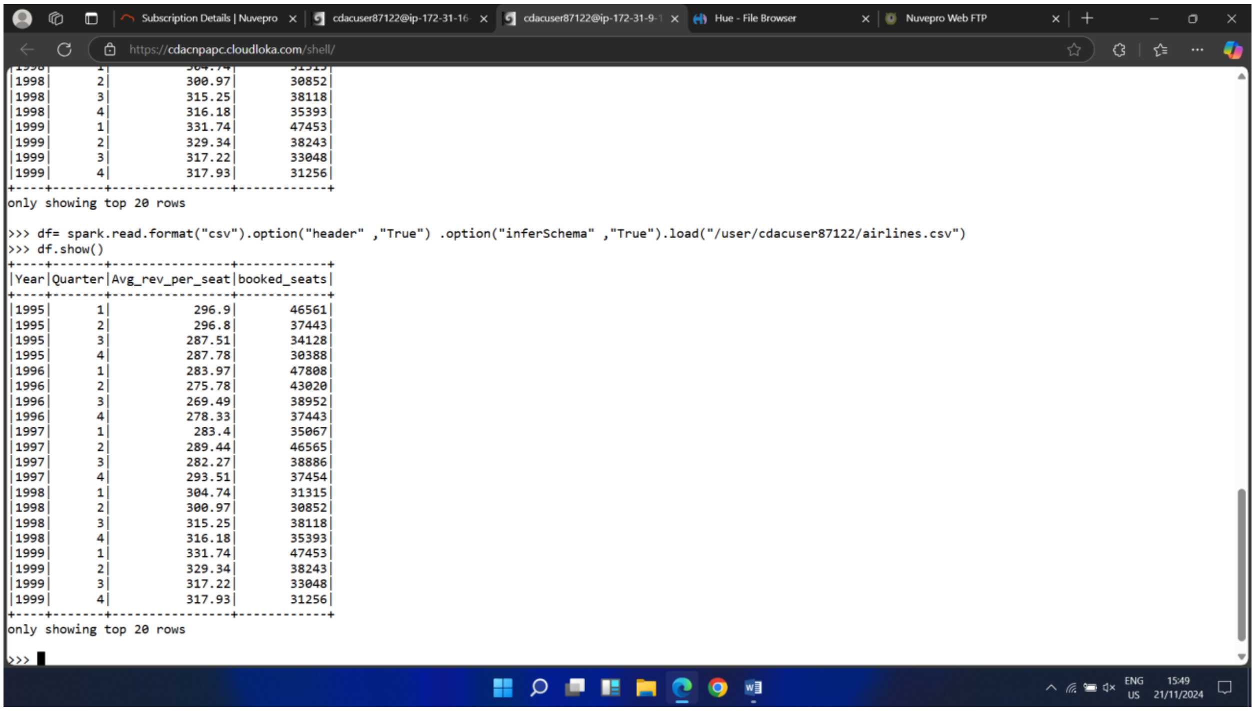

Subscription Details | Nuvepro
cdacuser87122@ip-172-31-16-
cdacuser87122@ip-172-31-9-1
Hue - File Browser
Nuvepro Web FTP
https://npapc.cloudloka.com/shell/
ENG
US
15:17
21/11/2024

```
OK
17      Aerolitoral     DH8
Time taken: 86.05 seconds, Fetched: 1 row(s)
hive (cdac_harsh)> select  count(*) as no_of_count ,r.equipment   from routes r  join airline al  on al.iata = r.source_airport_id   group by al.name ,r.equipmen
t order by no_of_count desc limit 1;
Query ID = cdacuser87122_20241121094544_934b8e5e-b3dd-49fd-85d7-a7bbcdf5c304
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_2545, Tracking URL = http://master:6318/proxy/application_1732089968849_2545/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2545
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 09:45:56,272 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:46:04,431 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 14.0 sec
2024-11-21 09:46:11,562 Stage-1 map = 100%,  reduce = 25%, Cumulative CPU 17.64 sec
2024-11-21 09:46:12,580 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 24.74 sec
2024-11-21 09:46:13,597 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.24 sec
MapReduce Total cumulative CPU time: 28 seconds 240 msec
Ended Job = job_1732089968849_2545
Launching Job 2 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_2548, Tracking URL = http://master:6318/proxy/application_1732089968849_2548/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2548
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 4
2024-11-21 09:46:26,195 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:46:34,348 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.64 sec
2024-11-21 09:46:41,479 Stage-2 map = 100%,  reduce = 50%, Cumulative CPU 7.81 sec
```

Data Frame

df= spark.read.format("csv").option("header" ,"True") .option("inferSchema" ,"True").load("/user/cdacuser87122/airlines.csv")

>>> df.show()



```
|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|
|1999|      4|          317.93|      31256|
+----+-------+----------------+-----------+
only showing top 20 rows

>>> df= spark.read.format("csv").option("header" ,"True") .option("inferSchema" ,"True").load("/user/cdacuser87122/airlines.csv")
>>> df.show()
+----+-------+----------------+-----------+
|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|
|1999|      4|          317.93|      31256|
+----+-------+----------------+-----------+
only showing top 20 rows

>>> 
```

Subscription Details | Nuvepro
cdacuser87122@ip-172-31-16-
cdacuser87122@ip-172-31-9-1
Hue - File Browser
Nuvepro Web FTP
https://cdacnpapc.cloudloka.com/shell/
ENG
US
15:49
21/11/2024

Q.2

1.

Find_insight =df.agg(min(“avg_rev_per_seat”) ,max(“avg_rev_per_seat”)avg(“avg_rev_per_seat)”)

Find_insight

2.

df.groupBy(“avg_rev_per_seat”>290).agg(count(“avg_rev_per_seat”)

3.

df.groupBy(“Quarter”)agg(sum(“booked_seat”).show()

4.

df.groupBy(“Year”).show()

5.

```
df.groupBy(“Year”).agg(sum(“avg_rev_per_seat”).limit(10)
```

RDD

1.