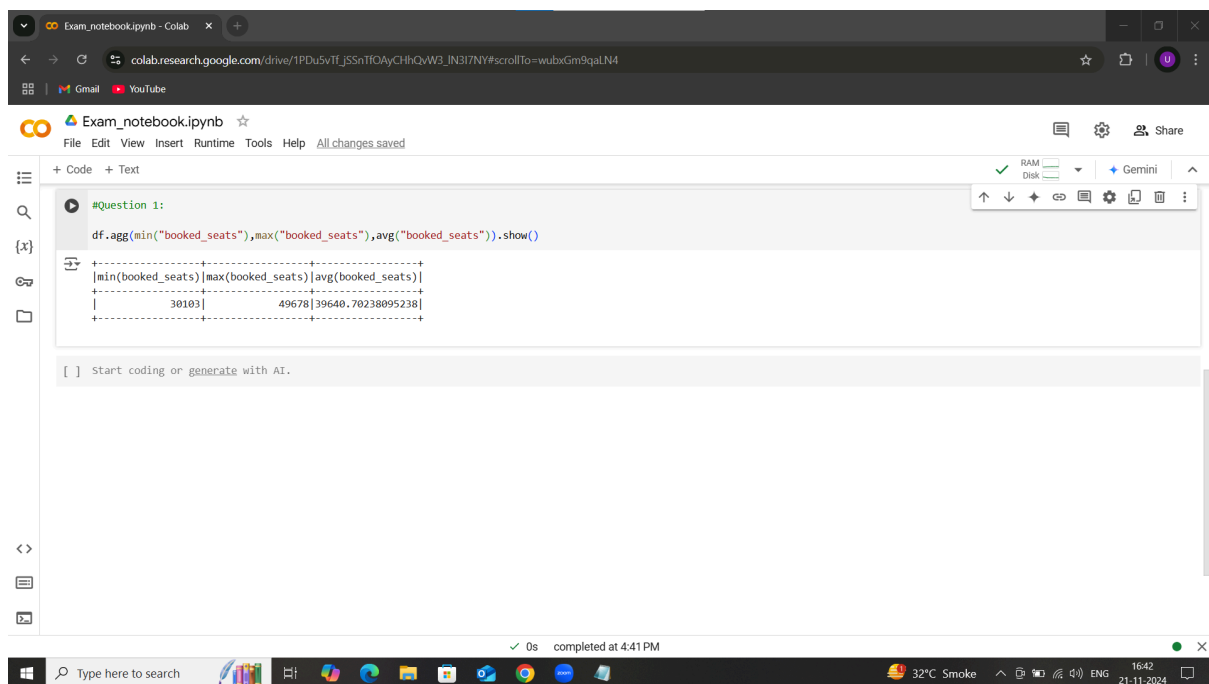


SPARK

USE RDD OR DATAFRAME:

1:

```
df.agg(min("booked_seats"),max("booked_seats"),avg("booked_seats")).show()
```



2:

```
df.filter(col("avg_rev")<290.0).count()
```

The screenshot shows a Google Colab notebook titled "Exam_notebook.ipynb". The code cell contains the following Python code:

```
#Question 2
df.filter(col("avg_rev")<290.0).count()
```

The output of the code is the integer value 9. The notebook interface includes a menu bar (File, Edit, View, Insert, Runtime, Tools, Help), a toolbar with icons for code, text, and search, and a status bar at the bottom indicating the execution time as 1s and completion at 4:44 PM.

3:

```
df.groupby("quart").agg(avg("booked_seats").alias("Average Booked Seats")).show()
```

The screenshot shows a Google Colab notebook titled "Exam_notebook.ipynb". The code cell contains the following Python code:

```
[14] #Question 3
df.groupby("quart").agg(avg("booked_seats").alias("Average Booked Seats")).show()
```

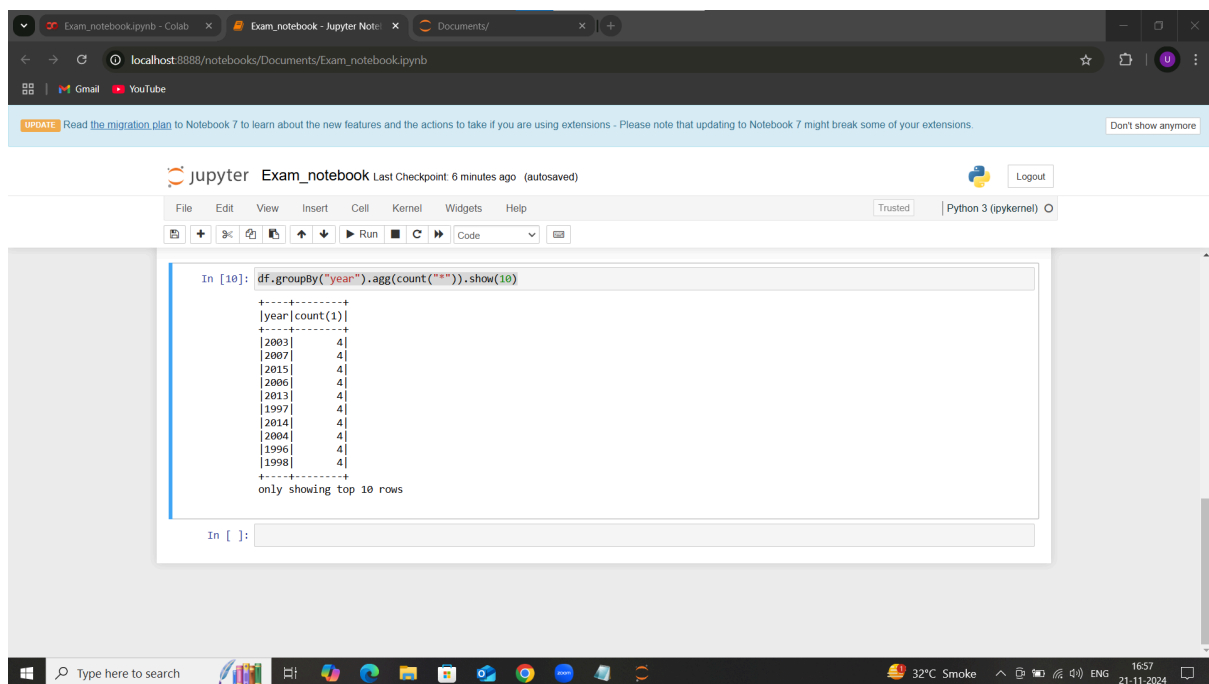
The output of the code is a text representation of a DataFrame:

```
+-----+
|quart|Average Booked Seats|
+-----+
| 1   | 41607.666666666664  |
| 3   | 39386.23809523809   |
| 4   | 39111.95238095238   |
| 2   | 38456.95238095238   |
+-----+
```

The notebook interface includes a menu bar (File, Edit, View, Insert, Runtime, Tools, Help), a toolbar with icons for code, text, and search, and a status bar at the bottom indicating the execution time as 0s and completion at 4:48 PM.

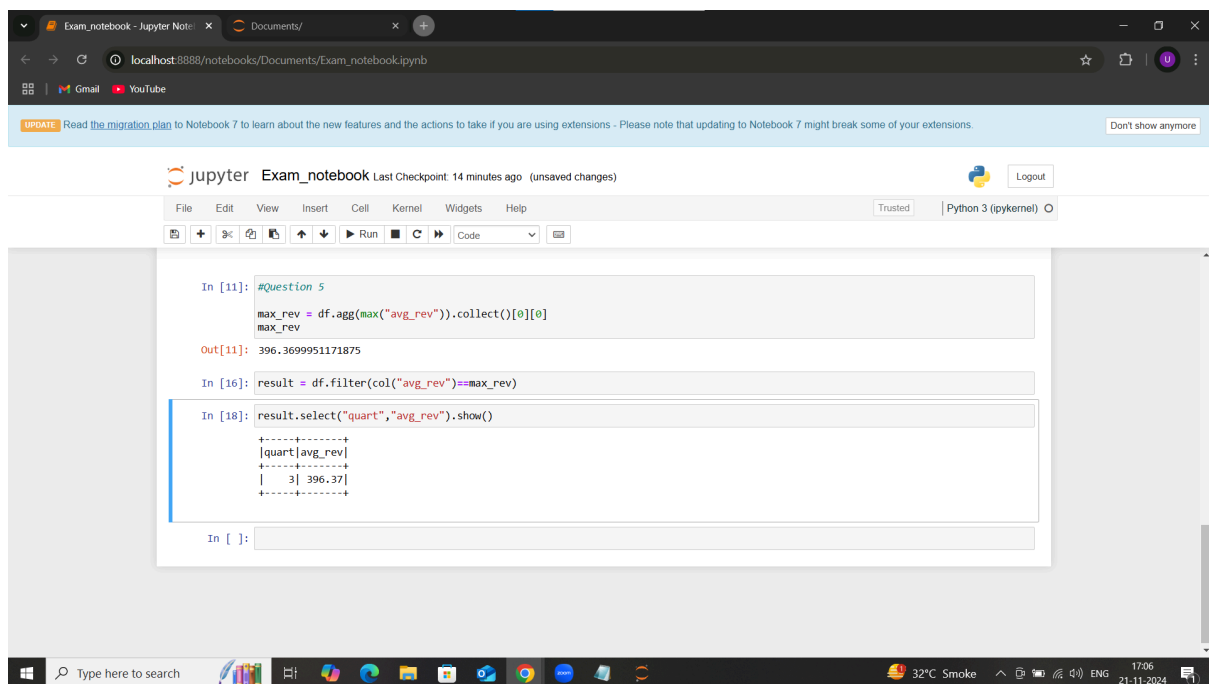
4:

```
df.groupby("year").agg(count("*")).show(10)
```



5:

```
max_rev = df.agg(max("avg_rev")).collect()[0][0]
result = df.filter(col("avg_rev")==max_rev)
result.select("quart","avg_rev").show()
```



USE RDD ONLY

1:

Code:

```
#loading file
datardd = sc.textFile("user/cdacuser/training/airlines.csv")

#checking for header
datardd.take(5)

#eliminating the header
head = datardd.first()
eliminate = datardd.filter(lambda a: a!=head)
```

HIVE:

Question 1:

1:

Query:

```
select src_airport_id from routes where src_airport_id not in (select dest_airport_id
from routes) limit 5;
```

Output:

```
Subscription Details | Nuvepro x Hue - File Browser x Nuvepro Web FTP x cdcuser222@ip-172-31-16-20 x exam_doc - Google Docs x +
npapcloudloka.com/shell/
Gmail YouTube

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_2971, Tracking URL = http://master:6318/proxy/application_1732089968849_2971/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2971
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 11:53:59,950 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:54:07,109 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.61 sec
2024-11-21 11:54:13,245 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.75 sec
MapReduce Total cumulative CPU time: 4 seconds 750 msec
Ended Job = job_1732089968849_2971
Launching Job 4 out of 4
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_2978, Tracking URL = http://master:6318/proxy/application_1732089968849_2978/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2978
Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 4
2024-11-21 11:54:33,271 Stage-2 map = 0%, reduce = 0%
2024-11-21 11:54:38,387 Stage-2 map = 33%, reduce = 0%, Cumulative CPU 2.14 sec
2024-11-21 11:54:40,431 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 6.73 sec
2024-11-21 11:54:45,535 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 19.27 sec
2024-11-21 11:54:46,559 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 23.63 sec
MapReduce Total cumulative CPU time: 23 seconds 630 msec
Ended Job = job_1732089968849_2978
MapReduce Jobs Launched:
Stage-Stage-4: Reduce: 4 Cumulative CPU: 9.34 sec HDFS Read: 23732 HDFS Write: 384 SUCCESS
Stage-Stage-3: Reduce: 1 Cumulative CPU: 2.65 sec HDFS Read: 6500 HDFS Write: 115 SUCCESS
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.75 sec HDFS Read: 9629 HDFS Write: 96 SUCCESS
Stage-Stage-2: Map: 3 Reduce: 4 Cumulative CPU: 23.63 sec HDFS Read: 31898 HDFS Write: 348 SUCCESS
Total MapReduce CPU Time Spent: 40 seconds 370 msec
OK
Time taken: 91.742 seconds
hive>
```

2:

Query:

```
select airline_id, count(*) from routes group by airline_id order by count(*) desc limit 3;
```

Output:

```
Subscription Details | Nuvepro x Hue - File Browser x Nuvepro Web FTP x cdcuser222@ip-172-31-16-20 x exam_doc - Google Docs x +
npapcloudloka.com/shell/
Gmail YouTube

set hive.exec.reducers.bytes.per.reducer=<number>
order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2945, Tracking URL = http://master:6318/proxy/application_1732089968849_2945/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2945
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
24-11-21 11:45:44,138 Stage-1 map = 0%, reduce = 0%
24-11-21 11:45:52,299 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.93 sec
24-11-21 11:45:59,429 Stage-1 map = 100%, reduce = 25%, Cumulative CPU 8.75 sec
24-11-21 11:46:00,448 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 14.09 sec
24-11-21 11:46:01,467 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 16.79 sec
MapReduce Total cumulative CPU time: 16 seconds 790 msec
Ended Job = job_1732089968849_2945
Launching Job 2 out of 2
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_2947, Tracking URL = http://master:6318/proxy/application_1732089968849_2947/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2947
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
24-11-21 11:46:14,497 Stage-2 map = 0%, reduce = 0%
24-11-21 11:46:20,617 Stage-2 map = 50%, reduce = 0%, Cumulative CPU 2.65 sec
24-11-21 11:46:22,659 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 5.36 sec
24-11-21 11:46:25,721 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.44 sec
MapReduce Total cumulative CPU time: 8 seconds 440 msec
Ended Job = job_1732089968849_2947
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 4 Cumulative CPU: 16.79 sec HDFS Read: 2408745 HDFS Write: 11872 SUCCESS
Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 8.44 sec HDFS Read: 23402 HDFS Write: 151 SUCCESS
Total MapReduce CPU Time Spent: 25 seconds 230 msec

36 2484
37 2354
39 2180
Time taken: 57.415 seconds, Fetched: 3 row(s)
```

3:

Query:

Select count(distinct(equipment)) from routes;

Output:

```
Subscription Details | Nuvepro x Hue - File Browser x Nuvepro Web FTP x cdacuser222@ip-172-31-16-20 x exam_doc - Google Docs x +
npapcloudloka.com/shell/
Gmail YouTube
routes
Time taken: 0.025 seconds, Fetched: 5 row(s)
hive (airlinedb)> select * from routes limit 10;
OK
2B 410 AER 2965 KZN 2990 0 CR2
2B 410 ASF 2966 KZN 2990 0 CR2
2B 410 ASF 2966 MRV 2962 0 CR2
2B 410 CEK 2968 KZN 2990 0 CR2
2B 410 CEK 2968 OVB 4078 0 CR2
2B 410 DME 4029 KZN 2990 0 CR2
2B 410 DME 4029 NBC 6969 0 CR2
2B 410 DME 4029 TKG NULL 0 CR2
2B 410 DME 4029 UUA 6160 0 CR2
2B 410 EGO 6156 KGO 2952 0 CR2
Time taken: 1.255 seconds, Fetched: 10 row(s)
hive (airlinedb)> select count(distinct(equipment)) from routes;
Query ID = cdacuser222_20241121120348_fcd94d88-272c-4353-8cef-e68ab565b4ab
Total jobs = 1
Launching Job 1 out of 1
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_3006, Tracking URL = http://master:6318/proxy/application_1732089968849_3006/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_3006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 12:04:00,729 Stage-1 map = 0%, reduce = 0%
2024-11-21 12:04:08,880 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.65 sec
2024-11-21 12:04:17,035 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.46 sec
MapReduce Total cumulative CPU time: 8 seconds 460 msec
Ended Job = job_1732089968849_3006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.46 sec HDFS Read: 2385214 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 460 msec
OK
3946
Time taken: 33.293 seconds, Fetched: 1 row(s)
hive (airlinedb)> 
```

Question 2:

1.

Query:

create table routes_partitioned
(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 int)
partitioned by (dest_airport_iata) row format delimited
fields terminated by "," stored as textfile;