

AQ1. select ap.name ,r.src_airport_iata from airports ap
join routes r
on r.src_airport_id=ap.airport_id
where r.src_airport_iata != dest_airport_iata limit 10;

```

Query ID = cdacuser3114_20241121110641_62adcc50-7faa-447d-9351-cbb1c856c917
Total jobs = 1
Launching Job 1 out of 1
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_2756, Tracking URL = http://master:6318/proxy/application_1732089968849_2756/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2756
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 11:06:54,281 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:07:02,484 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 6.48 sec
2024-11-21 11:07:03,506 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.97 sec
2024-11-21 11:07:09,640 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 26.15 sec
2024-11-21 11:07:10,658 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 30.26 sec
MapReduce Total cumulative CPU time: 30 seconds 260 msec
Ended Job = job_1732089968849_2756
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 30.26 sec HDFS Read: 3156418 HDFS Write: 1526 S
SUCCESS
Total MapReduce CPU Time Spent: 30 seconds 260 msec
OK
Madang MAG
Madang MAG
Madang MAG
Madang MAG
Madang MAG
Madang MAG
Madang MAG
Madang MAG
Wewak Intl WIK
Wewak Intl WIK
Time taken: 32.694 seconds, Fetched: 10 row(s)
hive (cdac_shivam)>

```

AQ2.

AQ3. select count(distinct(equipment)) from routes;

```
Subscription Details | Nuvepro x cdacuser3114@ip-172-31-16-2 x Hue x +
npapcloudloka.com/shell/
Gmail YouTube Maps W3Schools Online... Tutorials List - Javat... Job card - shivanga... Java Programming f...
2024-11-21 11:16:36,468 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.57 sec
MapReduce Total cumulative CPU time: 6 seconds 570 msec
Ended Job = job_1732089968849_2822
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.57 sec HDFS Read: 2389880 HDFS Write: 105 SUC
CESS
Total MapReduce CPU Time Spent: 6 seconds 570 msec
OK
67663
Time taken: 29.566 seconds, Fetched: 1 row(s)
hive (cdac_shivam)> select count(distinct(equipment)) from routes;
Query ID = cdacuser3114_20241121111713_5ac2d979-a246-46ce-8ebf-fda3f3610d17
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_2829, Tracking URL = http://master:6318/proxy/application_173208996
8849_2829/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2829
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 11:17:25,479 Stage-1 map = 0%, reduce = 0%
2024-11-21 11:17:33,629 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.74 sec
2024-11-21 11:17:41,788 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.75 sec
MapReduce Total cumulative CPU time: 7 seconds 750 msec
Ended Job = job_1732089968849_2829
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.75 sec HDFS Read: 2385309 HDFS Write: 104 SUC
CESS
Total MapReduce CPU Time Spent: 7 seconds 750 msec
OK
3946
Time taken: 31.321 seconds, Fetched: 1 row(s)
hive (cdac_shivam)>
```

B.Q1

B.Q2

BQ3. select * from routes where dest_airport_iata='ORD';

```
Subscription Details | Nuvepro x cdacuser3114@ip-172-31-16-2 x Hue x +
npapcloudloka.com/shell/
Gmail YouTube Maps W3Schools Online... Tutorials List - Javat... Job card - shivanga... Java Programming f...
OK
Time taken: 0.339 seconds
hive (cdac_shivam)> select * from routes where dest_airport_iata='ORD';
OK
3E 10739 BRL 5726 ORD 3830 0 CNC
3E 10739 DEC 4042 ORD 3830 0 CNC
AA 24 ABQ 4019 ORD 3830 Y 0 E75
AA 24 ALO 5718 ORD 3830 Y 0 ERD
AA 24 AMM 2170 ORD 3830 Y 0 340
AA 24 ART 3838 ORD 3830 Y 0 ERD
AA 24 ATL 3682 ORD 3830 Y 0 CR7 E75
AA 24 AUH 2179 ORD 3830 Y 0 777
AA 24 AUS 3673 ORD 3830 0 M83 M80
AA 24 AZO 4039 ORD 3830 Y 0 ER4 ERD
AA 24 BDL 3825 ORD 3830 Y 0 E75 ER4
AA 24 BMI 4037 ORD 3830 Y 0 ER4 ERD
AA 24 BNA 3690 ORD 3830 Y 0 ERD ER4 CR7 E75
AA 24 BOS 3448 ORD 3830 0 738
AA 24 BUF 3820 ORD 3830 Y 0 CR7 ER4 E75
AA 24 BWI 3849 ORD 3830 Y 0 E75 ER4 CR7
AA 24 CDG 1382 ORD 3830 0 763
AA 24 CHA 3578 ORD 3830 Y 0 ER4 ERD
AA 24 CHO 4015 ORD 3830 Y 0 CR7
AA 24 CID 4043 ORD 3830 Y 0 ER4
AA 24 CLE 3486 ORD 3830 Y 0 CR7 ER4 ERD
AA 24 CLT 3876 ORD 3830 0 321 320 CR9 E90
AA 24 CMH 3759 ORD 3830 Y 0 ER4 CR7 ERD
AA 24 CHI 4049 ORD 3830 Y 0 ER4 ERD
AA 24 COU 3719 ORD 3830 Y 0 ER4 ERD
AA 24 CUN 1852 ORD 3830 0 738
AA 24 CVG 3488 ORD 3830 Y 0 ER4
AA 24 CWA 4045 ORD 3830 Y 0 ER4 ERD
AA 24 DAY 3627 ORD 3830 Y 0 ER4
AA 24 DBQ 4388 ORD 3830 Y 0 ER4 ERD
AA 24 DCA 3520 ORD 3830 0 738 M83 M80
AA 24 DEN 3751 ORD 3830 0 M80 M83
AA 24 DFW 3670 ORD 3830 0 M80 M83
AA 24 DOH 2241 ORD 3830 0 777
AA 24 DSM 3729 ORD 3830 Y 0 E75 ERD ER4
```

BQ4.

Spark Q2 -1

```
Subscription Details | Nuvepro x cdacuser3114@ip-172-31-16-2 x cdacuser3114@ip-172-31-16-2 x Hue - File Browser x +
npapccloudloka.com/shell/
Gmail YouTube Maps W3Schools Online... Tutorials List - Javat... Job card - shivanga... Java Programming f...
All Bookmarks

[2001] 43399.5
[2005] 37652.5
[2000] 38594.0
[2010] 40935.25
[2011] 35661.75
[2008] 41724.25
[1999] 37500.0
+-----+
only showing top 20 rows

>>> df.groupBy("Year").agg(F.avg("booked_seats").alias("avg_booked_seat")).show()
+-----+
|Year|avg_booked_seat|
+-----+
[2003] 39038.25
[2007] 44074.75
[2015] 41359.5
[2006] 38447.25
[2013] 43419.0
[1997] 39493.0
[2014] 39955.75
[2004] 41200.0
[1996] 41805.75
[1998] 33919.5
[2012] 41519.0
[2009] 37577.0
[1995] 37130.0
[2001] 43399.5
[2005] 37652.5
[2000] 38594.0
[2010] 40935.25
[2011] 35661.75
[2008] 41724.25
[1999] 37500.0
+-----+
only showing top 20 rows

>>>
```

```
>>> df.groupBy("Year").agg(F.min("booked_seats").alias("min_booked_seat")).orderBy("min_booked_seat",ascending=True).show(10)
+-----+
|Year|min_booked_seat|
+-----+
[2000] 30103
[1995] 30388
[2011] 30562
[1998] 30852
[2004] 30877
[1999] 31256
[2005] 32003
[2002] 32406
[2009] 32491
[2006] 32621
+-----+
only showing top 10 rows

>>> df.groupBy("Year").agg(F.max("booked_seats").alias("max_booked_seat")).orderBy("max_booked_seat",ascending=False).show(10)
+-----+
|Year|max_booked_seat|
+-----+
[2010] 49678
[2013] 49143
[2004] 49022
[2000] 48159
[2014] 47928
[1996] 47808
[2007] 47758
[2005] 47608
[1999] 47453
[2008] 46885
+-----+
only showing top 10 rows

>>> █
```

Q2-2

```
Subscription Details | Nuvepro x cdacuser3114@ip-172-31-16-2 x cdacuser3114@ip-172-31-16-2 x Hue - File Browser x +
npapc.cloudloka.com/shell/
Gmail YouTube Maps W3Schools Online... Tutorials List - Javat... Job card - shivanga... Java Programming f...
[1996] 276.8925
[1997] 287.155
[1995] 292.2475
[2004] 305.875
[2005] 307.185
[1998] 309.285
[2009] 310.61
[2002] 312.525
[2003] 315.4675
[2001] 319.7975
[1999] 324.0575
[2007] 325.14
[2006] 328.3
[2010] 335.8325
[2000] 339.0325
[2008] 345.1575
[2011] 363.63250000000005
[2012] 374.675
[2015] 377.1275
[2013] 382.0025
+-----+
only showing top 20 rows

>>> df.groupby("Year").agg(F.avg("Avg_rev_per_seat").alias("Avg_rev_per_seat")).orderBy("Avg_rev_per_seat"<290).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: '<' not supported between instances of 'str' and 'int'
>>> df.groupby("Year").agg(F.avg("Avg_rev_per_seat").alias("Avg_rev_per_seat")).orderBy("Avg_rev_per_seat",ascending=True).show(2)
+-----+
|Year|Avg_rev_per_seat|
+-----+
|1996| 276.8925|
|1997| 287.155|
+-----+
only showing top 2 rows

>>>
```

Q2.3

```
+-----+
only showing top 2 rows

>>> df.groupby("Quarter").agg(F.avg("booked_seat").alias("total_booked_seat")).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 118, in agg
    jdf = self._jgd.agg(exprs[0]._jc,
  File "/opt/spark-3.1.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark-3.1.2/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve ''Quarter'' given input columns: [Avg_rev_per_seat, Quarter, Year, booked_seats];
'Aggregate [Quarter], [Quarter, avg('booked_seat') AS total_booked_seat#294]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupby("Quarter").agg(F.avg("booked_seat").alias("total_booked_seat")).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-3.1.2/python/pyspark/sql/group.py", line 118, in agg
    jdf = self._jgd.agg(exprs[0]._jc,
  File "/opt/spark-3.1.2/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark-3.1.2/python/pyspark/sql/utils.py", line 117, in deco
    raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve ''booked_seat'' given input columns: [Avg_rev_per_seat, Quarter, Year, booked_seats];
'Aggregate [Quarter#17], [Quarter#17, avg('booked_seat') AS total_booked_seat#300]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupby("Quarter").agg(F.avg("booked_seats").alias("avg_booked_seat")).show()
+-----+
|Quarter| avg_booked_seat|
+-----+
| 1|41607.666666666664|
| 3| 39386.23809523809|
| 4| 39111.95238095238|
| 2| 38456.95238095238|
+-----+

>>>
```

Q2.4

```
Subscription Details | Nuvepro x cdacuser3114@ip-172-31-16-2 x cdacuser3114@ip-172-31-16-2 x Hue - File Browser x +
npapc.cloudloka.com/shell/
Gmail YouTube Maps W3Schools Online... Tutorials List - Javat... Job card - shivanga... Java Programming f... All Bookmarks

raise converted from None
pyspark.sql.utils.AnalysisException: cannot resolve ''Revenue'' given input columns: [Avg_rev_per_seat, Quarter, Year, booked_seats];
'Project ['Revenue, (Avg_rev_per_seat#18 * 'booked_seat) AS (Avg_rev_per_seat * booked_seat)#329]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupBy("Year", "Quarter").agg(((F.col("Avg_rev_per_seat")*F.col("booked_seat")).alias("Total_revenu")).limit(1).show()
...
File "<stdin>", line 3
^
SyntaxError: invalid syntax
>>> df.groupBy("Year").agg(F.count("Quarter").alias("row_count")).show()
+-----+
|Year|row_count|
+-----+
|2003|         4|
|2007|         4|
|2015|         4|
|2006|         4|
|2013|         4|
|1997|         4|
|2014|         4|
|2004|         4|
|1996|         4|
|1998|         4|
|2012|         4|
|2009|         4|
|1995|         4|
|2001|         4|
|2005|         4|
|2000|         4|
|2010|         4|
|2011|         4|
|2008|         4|
|1999|         4|
+-----+
only showing top 20 rows
```

Q2.5

```
SyntaxError: invalid syntax
>>> df.groupBy("Year", "Quarter").agg(((F.col("Avg_rev_per_seat")*F.col("booked_seat")).alias("Total_revenu")).limit(1).show()
...;
```