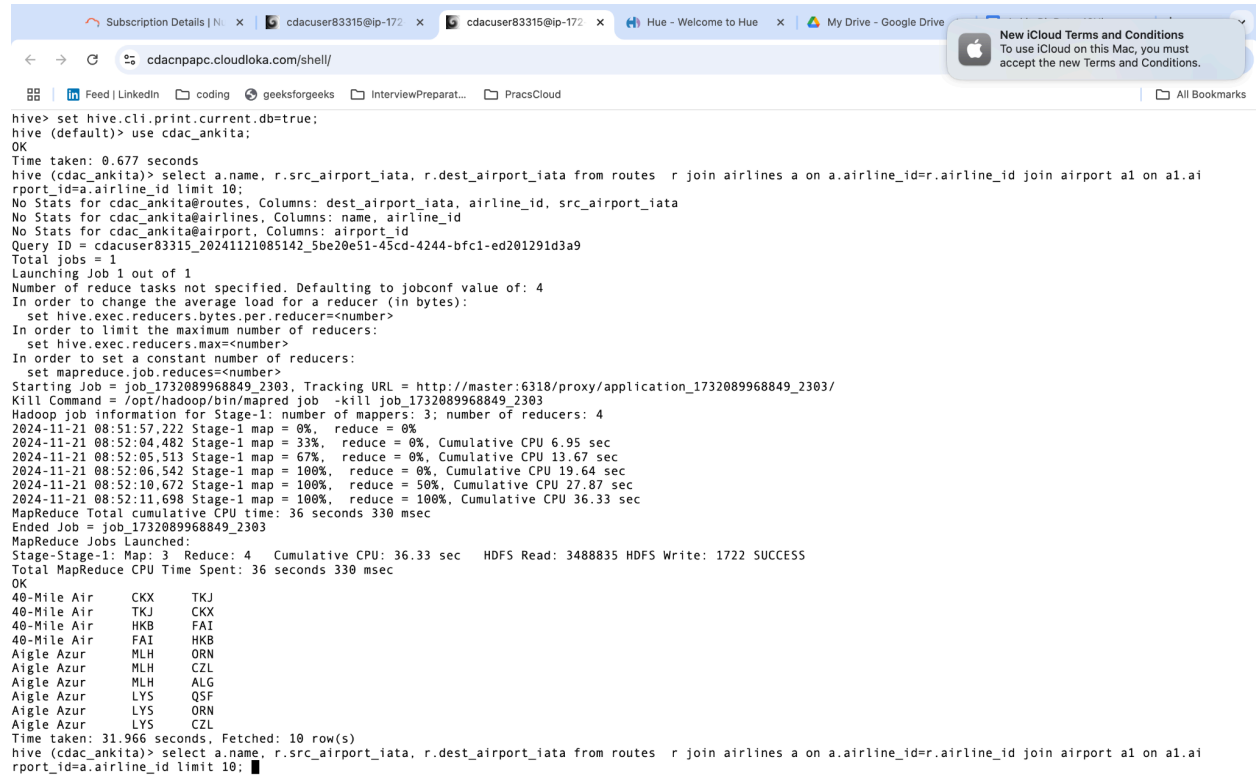


Hive

Ques1.

1.ans

```
select a.name, r.src_airport_iata, r.dest_airport_iata from routes r
join airlines a on a.airline_id=r.airline_id join airport a1 on a1.airport_id=a.airline_id limit 10;
```



```
cdacuser83315@ip-172: x cdacuser83315@ip-172: x Hue - Welcome to Hue x My Drive - Google Drive
cdacnpac.cloudloka.com/shell/
Feed | LinkedIn coding geeksforgeeks InterviewPreparat... PracsCloud All Bookmarks

hive> set hive.cli.print.current.db=true;
hive (default)> use cdac_ankita;
OK
Time taken: 0.677 seconds
hive (cdac_ankita)> select a.name, r.src_airport_iata, r.dest_airport_iata from routes r join airlines a on a.airline_id=r.airline_id join airport a1 on a1.airport_id=a.airline_id limit 10;
No Stats for cdac_ankita@routes, Columns: dest_airport_iata, airline_id, src_airport_iata
No Stats for cdac_ankita@airlines, Columns: name, airline_id
No Stats for cdac_ankita@airport, Columns: airport_id
Query ID = cdacuser83315_20241121085142_5be20e51-45cd-4244-bfc1-ed201291d3a9
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_2303, Tracking URL = http://master:6318/proxy/application_1732089968849_2303/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2303
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 4
2024-11-21 08:51:57,222 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:52:04,482 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 6.95 sec
2024-11-21 08:52:05,513 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 13.67 sec
2024-11-21 08:52:06,542 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 19.64 sec
2024-11-21 08:52:10,672 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 27.87 sec
2024-11-21 08:52:11,698 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 36.33 sec
MapReduce Total cumulative CPU time: 36 seconds 330 msec
Ended Job = job_1732089968849_2303
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3 Reduce: 4 Cumulative CPU: 36.33 sec HDFS Read: 3488835 HDFS Write: 1722 SUCCESS
Total MapReduce CPU Time Spent: 36 seconds 330 msec
OK
40-Mile Air CKX TKJ
40-Mile Air TKJ CKX
40-Mile Air HKB FAI
40-Mile Air FAI HKB
Aigle Azur MLH ORN
Aigle Azur MLH CZL
Aigle Azur MLH ALG
Aigle Azur LYS QSF
Aigle Azur LYS ORN
Aigle Azur LYS CZL
Time taken: 31.966 seconds, Fetched: 10 row(s)
hive (cdac_ankita)> select a.name, r.src_airport_iata, r.dest_airport_iata from routes r join airlines a on a.airline_id=r.airline_id join airport a1 on a1.airport_id=a.airline_id limit 10;
```

2.

3.

```
select max(airline_iata), count(*) as no_routes from routes;
```

```
Subscription Details | Nuvep... x cdacuser83315@ip-172-31-0 x cdacuser83315@ip-172-31-0 x Hue - File Browser x My P...
cdacnppac.cloudloka.com/shell/
Feed | LinkedIn coding geeksforgeeks InterviewPreparat... PracsCloud
hive (cdac_ankita)> select distinct(a.name) from routes r join airline a on a.airline_id=r.airline_id where trim(upper(r.equ
FAILED: SemanticException [Error 10001]: Line 1:43 Table not found 'airline'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airline a on a.airline_id=r.airline_id;
FAILED: SemanticException [Error 10001]: Line 1:58 Table not found 'airline'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id;
FAILED: SemanticException [Error 10002]: Line 1:74 Invalid column reference 'airline_id'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id;
FAILED: SemanticException [Error 10128]: Line 1:24 Not yet supported place for UDAF 'max'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id group by airline_iata;
FAILED: SemanticException 1:107 SELECT DISTINCT and GROUP BY can not be in the same query. Error encountered near token 'airline_iata'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id group by name;
FAILED: SemanticException 1:107 SELECT DISTINCT and GROUP BY can not be in the same query. Error encountered near token 'name'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id group by airline_id;
FAILED: SemanticException 1:107 SELECT DISTINCT and GROUP BY can not be in the same query. Error encountered near token 'airline_id'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id;
FAILED: SemanticException [Error 10128]: Line 1:24 Not yet supported place for UDAF 'max'
hive (cdac_ankita)> select a.name,max(equipment) from routes r join airlines a on a.airline_id=r.airline_id group by airline_id;
FAILED: SemanticException Column airline_id Found in more than One Tables/Subqueries
FAILED: SemanticException 1:107 SELECT DISTINCT and GROUP BY can not be in the same query. Error encountered near token 'airline_iata'
hive (cdac_ankita)> select distinct(a.name),max(equipment) from routes r join airlines a on a.airline_id=r.airline_id ;
FAILED: SemanticException [Error 10128]: Line 1:24 Not yet supported place for UDAF 'max'
hive (cdac_ankita)> select max(airline_iata).count(*) as no_routes from routes;
Query ID = cdacuser83315_20241121101922_5fa23171-7d89-40fa-a4fb-1b5608415bfd
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_2661, Tracking URL = http://master:6318/proxy/application_1732089968849_2661/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2661
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 10:19:32.970 Stage-1 map = 0%, reduce = 0%
2024-11-21 10:19:41.168 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.31 sec
2024-11-21 10:19:49.370 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.69 sec
MapReduce Total cumulative CPU time: 6 seconds 690 msec
Ended Job = job_1732089968849_2661
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.69 sec HDFS Read: 2390588 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 690 msec
OK
ZM
67663
Time taken: 29.453 seconds, Fetched: 1 row(s)
hive (cdac_ankita)>
```

Spark:

## Ques1

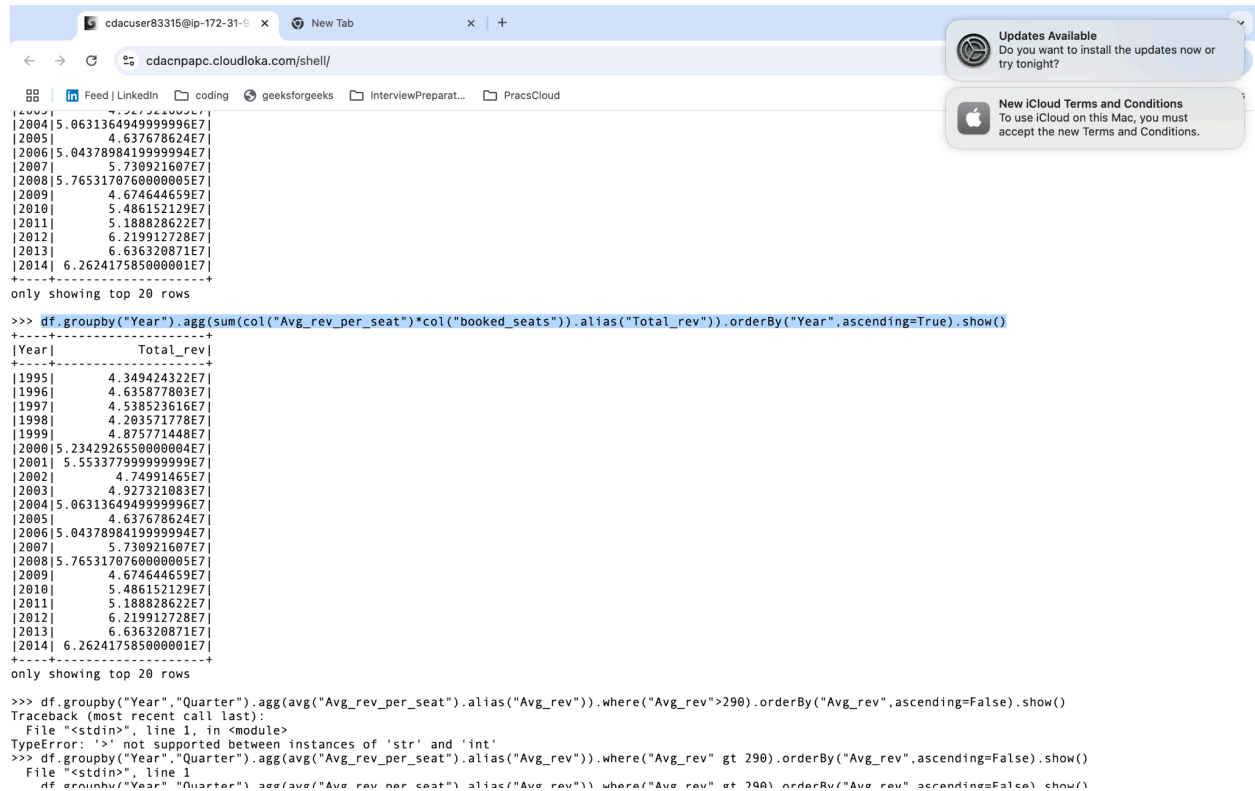
1.Ans

```
df.groupby("Year","booked_seat").where("booked_seats">40000).show()
```

## 2. Ans

Using df show distinct years data

```
df.groupby("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats"))).alias("Total_rev").orderBy("Year",ascending=True).show()
```



```
cdacuser83315@ip-172-31-10 x New Tab x +
cdacnpapc.cloudloka.com/shell/
Feed | LinkedIn | coding | geeksforgeeks | InterviewPreparat... | PracsCloud
[2004] 5.0631364949999996E7 | 4.637678624E7 |
[2005] 5.0437898419999994E7 | 5.738921687E7 |
[2006] 5.7653170760000005E7 | 4.674644659E7 |
[2007] 5.738921687E7 | 5.486152129E7 |
[2008] 5.7653170760000005E7 | 5.188828622E7 |
[2009] 4.674644659E7 | 6.219912728E7 |
[2010] 5.486152129E7 | 6.636320871E7 |
[2011] 5.188828622E7 | 6.262417585000001E7 |
[2012] 6.219912728E7 |
[2013] 6.636320871E7 |
[2014] 6.262417585000001E7 |
+-----+
only showing top 20 rows

>>> df.groupby("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")).alias("Total_rev")).orderBy("Year",ascending=True).show()
+-----+-----+
|Year|      Total_rev|
+-----+-----+
[1995] 4.349424322E7 |
[1996] 4.635877803E7 |
[1997] 4.538523616E7 |
[1998] 4.203571778E7 |
[1999] 4.875771448E7 |
[2000] 5.2342926550000004E7 |
[2001] 5.553377999999999E7 |
[2002] 4.74991465E7 |
[2003] 4.927321803E7 |
[2004] 5.0631364949999996E7 |
[2005] 5.0437898419999994E7 |
[2006] 5.7653170760000005E7 |
[2007] 5.738921687E7 |
[2008] 5.7653170760000005E7 |
[2009] 4.674644659E7 |
[2010] 5.486152129E7 |
[2011] 5.188828622E7 |
[2012] 6.219912728E7 |
[2013] 6.636320871E7 |
[2014] 6.262417585000001E7 |
+-----+-----+
only showing top 20 rows

>>> df.groupby("Year","Quarter").agg(avg("Avg_rev_per_seat").alias("Avg_rev")).where("Avg_rev">290).orderBy("Avg_rev",ascending=False).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: '>' not supported between instances of 'str' and 'int'
>>> df.groupby("Year","Quarter").agg(avg("Avg_rev_per_seat").alias("Avg_rev")).where("Avg_rev" gt 290).orderBy("Avg_rev",ascending=False).show()
  File "<stdin>", line 1
    df.groupby("Year","Quarter").agg(avg("Avg_rev_per_seat").alias("Avg_rev")).where("Avg_rev" gt 290).orderBy("Avg_rev",ascending=False).show()
    ^
SyntaxError: invalid syntax
```

## Ques2:

1 ans

```
df.agg(min("Avg_rev_per_seat").alias("Min_rev")).show()
df.agg(max("Avg_rev_per_seat").alias("Max_rev")).show()
```

```
df.agg(avg("Avg_rev_per_seat").alias("Avg_rev")).show()
```

```
cdacuser83315@ip-172-31-0 x +  
← → ↻ 🌐 cdacnpapc.cloudloka.com/shell/  
  
📁 Feed | 🔗 LinkedIn 📂 coding 🔄 geeksforgeeks 📁 InterviewPreparat... 📁 PracsCloud  
| 📖 All Bookmarks  
  
|2008|      1| 333.29|  
|2009|      1| 313.82|  
+-----+  
only showing top 20 rows  
  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>> df.agg(min("Avg_rev_per_seat").alias("Min_rev")).show()  
+-----+  
|Min_rev|  
+-----+  
| 269.49|  
+-----+  
  
>>> df.agg(max("Avg_rev_per_seat").alias("Max_rev")).show()  
+-----+  
|Max_rev|  
+-----+  
| 396.37|  
+-----+  
  
>>> df.agg(avg("Avg_rev_per_seat").alias("Avg_rev")).show()  
+-----+  
|      Avg_rev|  
+-----+  
|329.7475000000006|  
+-----+  
  
>>>
```

2.

3.

```
df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).ord
erBy("Year",ascending=True).show()
```

```
cdacuser83315@ip-172-31-0 x New Tab x +  
cdacnpapc.cloudloka.com/shell/  
Feed | LinkedIn coding geeksforgeeks InterviewPreparat... PracsCloud  
[1997| 2| 46565|  
[1998| 1| 31315|  
[1998| 4| 35393|  
[1998| 2| 30852|  
[1998| 3| 38118|  
[1999| 1| 47453|  
[1999| 2| 38243|  
[1999| 3| 33048|  
[1999| 4| 31256|  
+-----+  
only showing top 20 rows  
  
>>> df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).orderBy("Year",ascending=True).show()  
+---+-----+  
|Year|Booked_seats|  
+---+-----+  
[1995| 148520|  
[1996| 167223|  
[1997| 157972|  
[1998| 135678|  
[1999| 150000|  
[2000| 154376|  
[2001| 173598|  
[2002| 152195|  
[2003| 156153|  
[2004| 164800|  
[2005| 150610|  
[2006| 153789|  
[2007| 176299|  
[2008| 166897|  
[2009| 150308|  
[2010| 163741|  
[2011| 142647|  
[2012| 166076|  
[2013| 173676|  
[2014| 159823|  
+-----+  
only showing top 20 rows  
  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>  
>>>
```

```
4. df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).ord
   erBy("Year",ascending=True).show()
```

```
cdacuser83315@ip-172-31-0 x New Tab
cdacnpapc.cloudloka.com/shell/
Feed | LinkedIn | coding | geeksforgeeks | InterviewPreparat... | PracsCloud
only showing top 20 rows

>>> df.groupby("Year").orderBy("Year",ascending=True).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'orderBy'
>>> df.groupby("Year").orderBy("Year",ascending=True).show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'orderBy'
>>> df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).orderBy("Year",ascending=True).show()
  File "<stdin>", line 1
df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).orderBy("Year",ascending=True).show()
^
SyntaxError: invalid syntax
>>> df.groupby("Year").agg(sum("booked_seats").alias("Booked_seats")).orderBy("Year",ascending=True).show()
+-----+
|Year|Booked_seats|
+-----+
|1995|    148520|
|1996|    167223|
|1997|    157972|
|1998|    135678|
|1999|    150000|
|2000|    154376|
|2001|    173598|
|2002|    152195|
|2003|    156153|
|2004|    164800|
|2005|    150610|
|2006|    153789|
|2007|    176299|
|2008|    166897|
|2009|    150308|
|2010|    163741|
|2011|    142647|
|2012|    166076|
|2013|    173676|
|2014|    159823|
+-----+
only showing top 20 rows

>>> df.groupby("Year").show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'show'
>>> 
```

5.

```
df.groupby("Year").agg(sum(col("Avg_rev_per_seat")*col("booked_seats")
)).alias("Total_rev")).orderBy("Year",ascending=True).show()
```

cdacuser83315@ip-172-31-0 xNew Tab

cdacnpapc.cloudloka.com/shell/

Updates Available  
Do you want to install the updates now or try tonight?

New iCloud Terms and Conditions  
To use iCloud on this Mac, you must accept the new Terms and Conditions.

Feed | LinkedIn | coding | geeksforgeeks | InterviewPreparat... | PracsCloud

[2000] 5.2342926550000004E7 |

[2001] 5.5533779999999999E7 |

[2002] 4.74991465E7 |

[2003] 4.927321083E7 |

[2004] 5.0631364949999996E7 |

[2005] 4.637678624E7 |

[2006] 5.0437898419999994E7 |

[2007] 5.730921607E7 |

[2008] 5.7653170760000005E7 |

[2009] 4.674644659E7 |

[2010] 5.486152129E7 |

[2011] 5.188828622E7 |

[2012] 6.219912728E7 |

[2013] 6.636320871E7 |

[2014] 6.262417585000001E7 |

+-----+

only showing top 20 rows

>>> df.groupby("Year").agg(sum(col("Avg\_rev\_per\_seat")\*col("booked\_seats")).alias("Total\_rev")).orderBy("Year",ascending=True).show()

+-----+

[Year] Total\_rev

+-----+

[1995] 4.349424322E7 |

[1996] 4.635877803E7 |

[1997] 4.538523616E7 |

[1998] 4.203571778E7 |

[1999] 4.875771448E7 |

[2000] 5.2342926550000004E7 |

[2001] 5.5533779999999999E7 |

[2002] 4.74991465E7 |

[2003] 4.927321083E7 |

[2004] 5.0631364949999996E7 |

[2005] 4.637678624E7 |

[2006] 5.0437898419999994E7 |

[2007] 5.730921607E7 |

[2008] 5.7653170760000005E7 |

[2009] 4.674644659E7 |

[2010] 5.486152129E7 |

[2011] 5.188828622E7 |

[2012] 6.219912728E7 |

[2013] 6.636320871E7 |

[2014] 6.262417585000001E7 |

+-----+

only showing top 20 rows

>>>