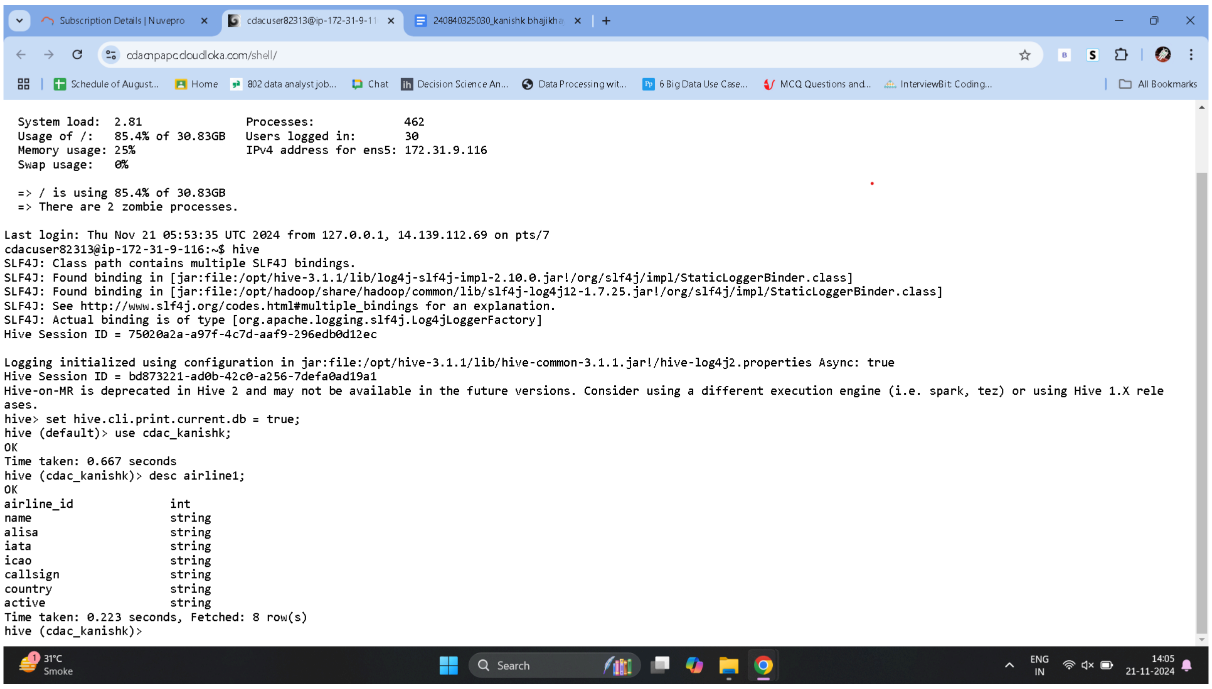
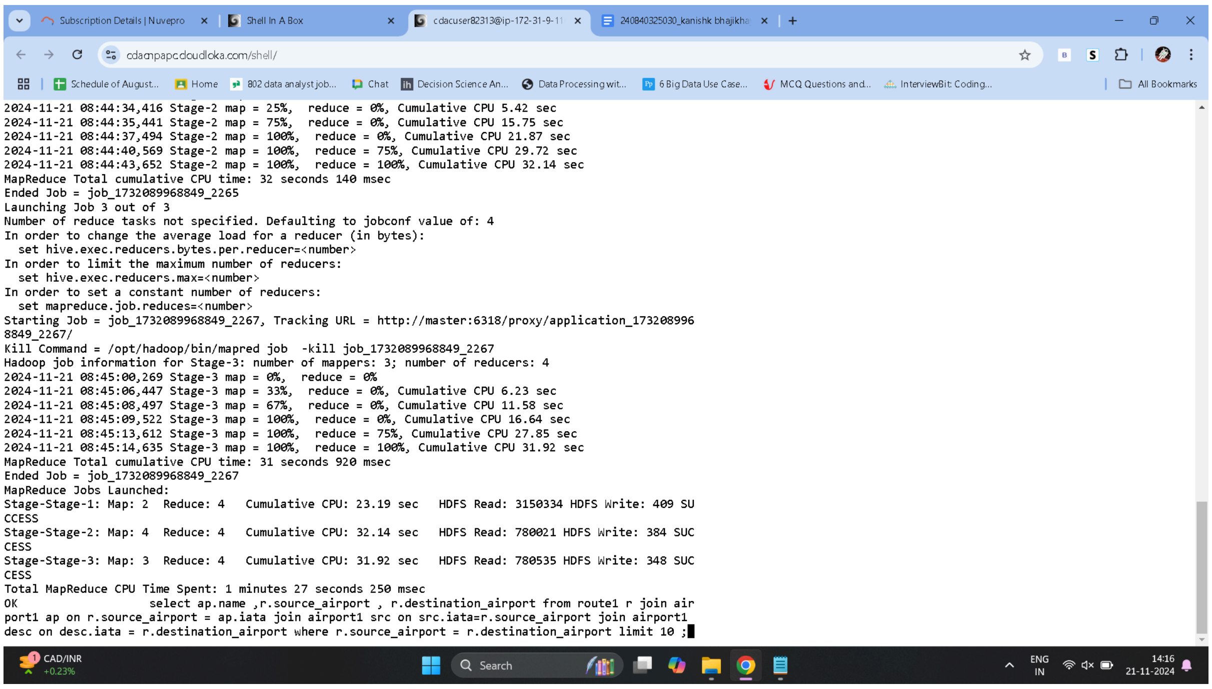


```
  System load:  2.81              Processes:            462
  Usage of /:   85.4% of 30.83GB  Users logged in:      30
  Memory usage: 25%               IPv4 address for ens5: 172.31.9.116
  Swap usage:   0%

  => / is using 85.4% of 30.83GB
  => There are 2 zombie processes.

Last login: Thu Nov 21 05:53:35 UTC 2024 from 127.0.0.1, 14.139.112.69 on pts/7
cdacuser82313@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 = 75020a2a-a97f-4c7d-aaf9-296edb0d12ec

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 = bd873221-ad0b-42c0-a256-7defa0ad19a1
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> set hive.cli.print.current.db = true;
hive (default)> use cdac_kanishk;
OK
Time taken: 0.667 seconds
hive (cdac_kanishk)> desc airline1;
OK
airline_id              int
name                    string
alisa                   string
iata                    string
icao                    string
callsign                string
country                 string
active                  string
Time taken: 0.223 seconds, Fetched: 8 row(s)
hive (cdac_kanishk)>
```

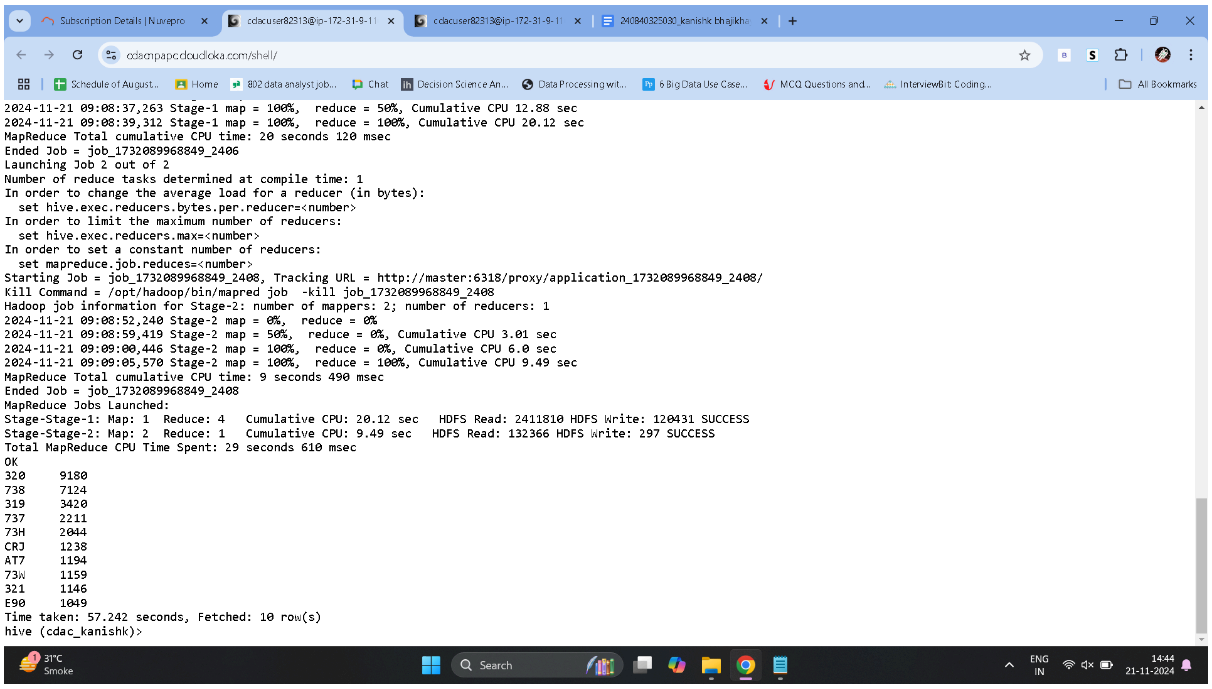
Question 1:

```
1 select ap.name ,r.source_airport , r.destination_airport from
route1 r join air
port1 ap on r.source_airport = ap.iata join airport1 src on
src.iata=r.source_airport join airport1
desc on desc.iata = r.destination_airport where r.source_airport =
r.destination_airport limit 10 ;
```



```
2024-11-21 08:44:34,416 Stage-2 map = 25%,  reduce = 0%, Cumulative CPU 5.42 sec
2024-11-21 08:44:35,441 Stage-2 map = 75%,  reduce = 0%, Cumulative CPU 15.75 sec
2024-11-21 08:44:37,494 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 21.87 sec
2024-11-21 08:44:40,569 Stage-2 map = 100%,  reduce = 75%, Cumulative CPU 29.72 sec
2024-11-21 08:44:43,652 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 32.14 sec
MapReduce Total cumulative CPU time: 32 seconds 140 msec
Ended Job = job_1732089968849_2265
Launching Job 3 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_2267, Tracking URL = http://master:6318/proxy/application_173208996
8849_2267/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2267
Hadoop job information for Stage-3: number of mappers: 3; number of reducers: 4
2024-11-21 08:45:00,269 Stage-3 map = 0%,  reduce = 0%
2024-11-21 08:45:06,447 Stage-3 map = 33%,  reduce = 0%, Cumulative CPU 6.23 sec
2024-11-21 08:45:08,497 Stage-3 map = 67%,  reduce = 0%, Cumulative CPU 11.58 sec
2024-11-21 08:45:09,522 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 16.64 sec
2024-11-21 08:45:13,612 Stage-3 map = 100%,  reduce = 75%, Cumulative CPU 27.85 sec
2024-11-21 08:45:14,635 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 31.92 sec
MapReduce Total cumulative CPU time: 31 seconds 920 msec
Ended Job = job_1732089968849_2267
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 23.19 sec   HDFS Read: 3150334 HDFS Write: 409 SU
CCESS
Stage-Stage-2: Map: 4  Reduce: 4   Cumulative CPU: 32.14 sec   HDFS Read: 780021 HDFS Write: 384 SUC
CESS
Stage-Stage-3: Map: 3  Reduce: 4   Cumulative CPU: 31.92 sec   HDFS Read: 780535 HDFS Write: 348 SUC
CESS
Total MapReduce CPU Time Spent: 1 minutes 27 seconds 250 msec
OK                   select ap.name ,r.source_airport , r.destination_airport from route1 r join air
port1 ap on r.source_airport = ap.iata join airport1 src on src.iata=r.source_airport join airport1
desc on desc.iata = r.destination_airport where r.source_airport = r.destination_airport limit 10 ;
```

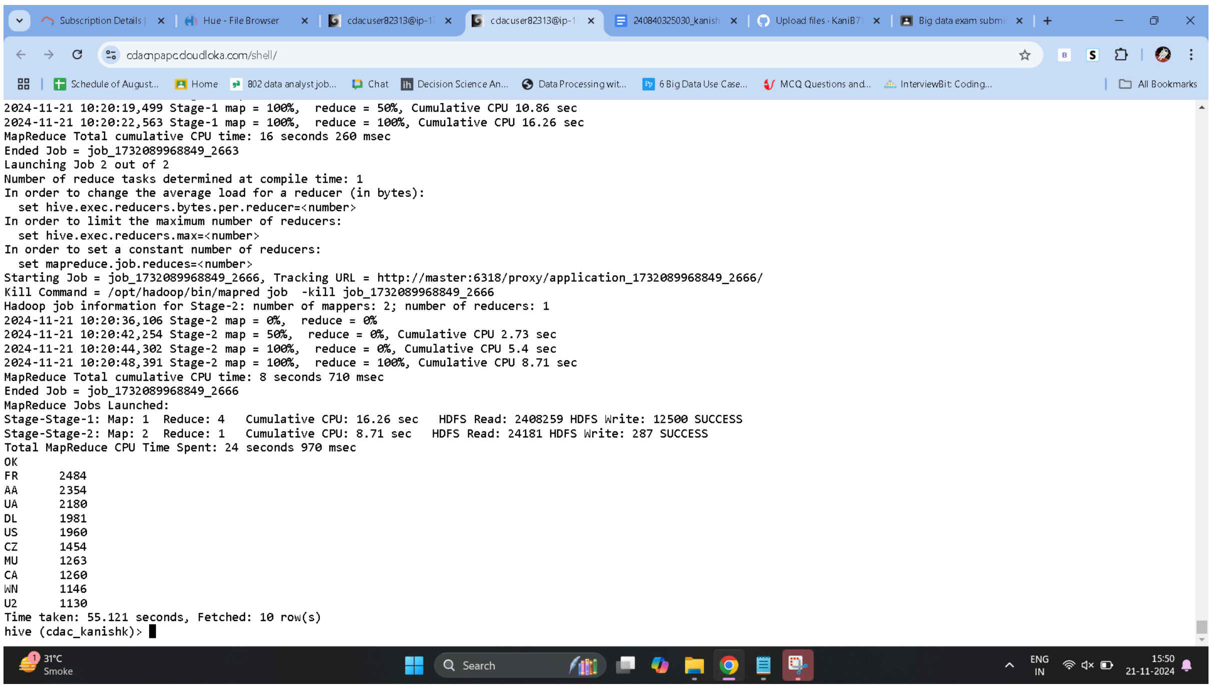
2.select equipment , count(*) as total from routes group by equipment order by total desc limit 10;



```
2024-11-21 09:08:37,263 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 12.88 sec
2024-11-21 09:08:39,312 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.12 sec
MapReduce Total cumulative CPU time: 20 seconds 120 msec
Ended Job = job_1732089968849_2406
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_2408, Tracking URL = http://master:6318/proxy/application_1732089968849_2408/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2408
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 09:08:52,240 Stage-2 map = 0%,  reduce = 0%
2024-11-21 09:08:59,419 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 3.01 sec
2024-11-21 09:09:00,446 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 6.0 sec
2024-11-21 09:09:05,570 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 9.49 sec
MapReduce Total cumulative CPU time: 9 seconds 490 msec
Ended Job = job_1732089968849_2408
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 20.12 sec   HDFS Read: 2411810 HDFS Write: 120431 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 9.49 sec   HDFS Read: 132366 HDFS Write: 297 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 610 msec
OK
320     9180
738     7124
319     3420
737     2211
73H     2044
CRJ     1238
AT7     1194
73W     1159
321     1146
E90     1049
Time taken: 57.242 seconds, Fetched: 10 row(s)
hive (cdac_kanishk)>
```

3:

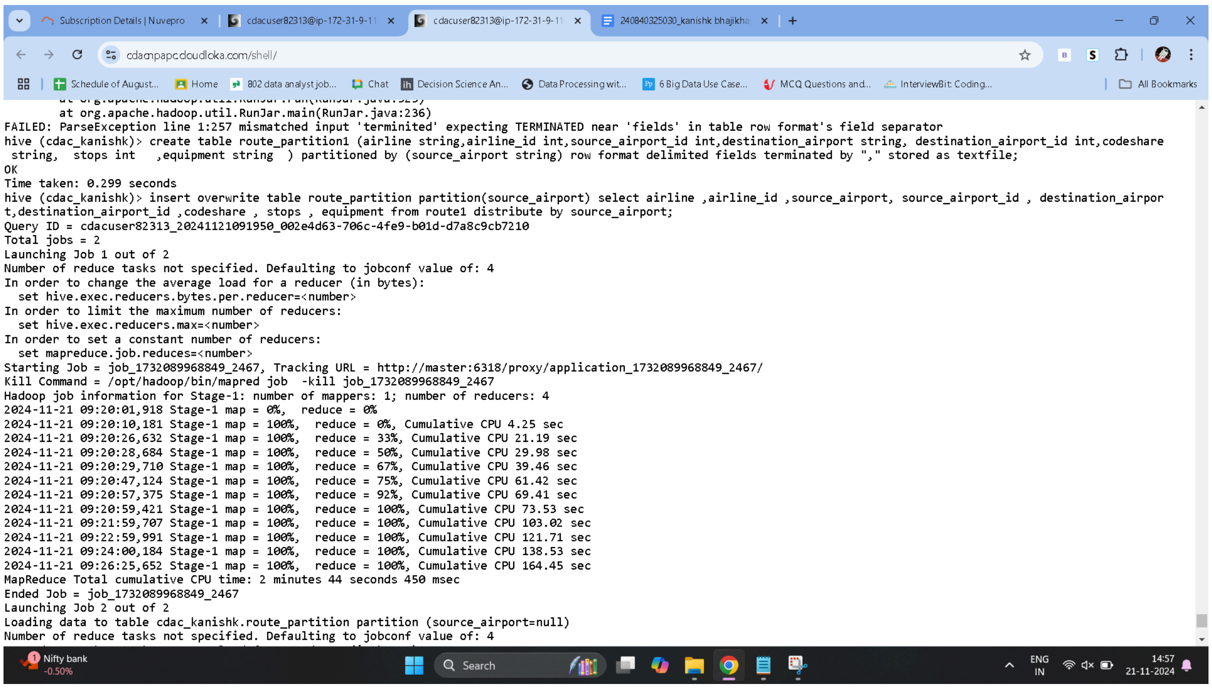
select airline ,count(*)as total from route1 group by airline order by total desc limit 10;



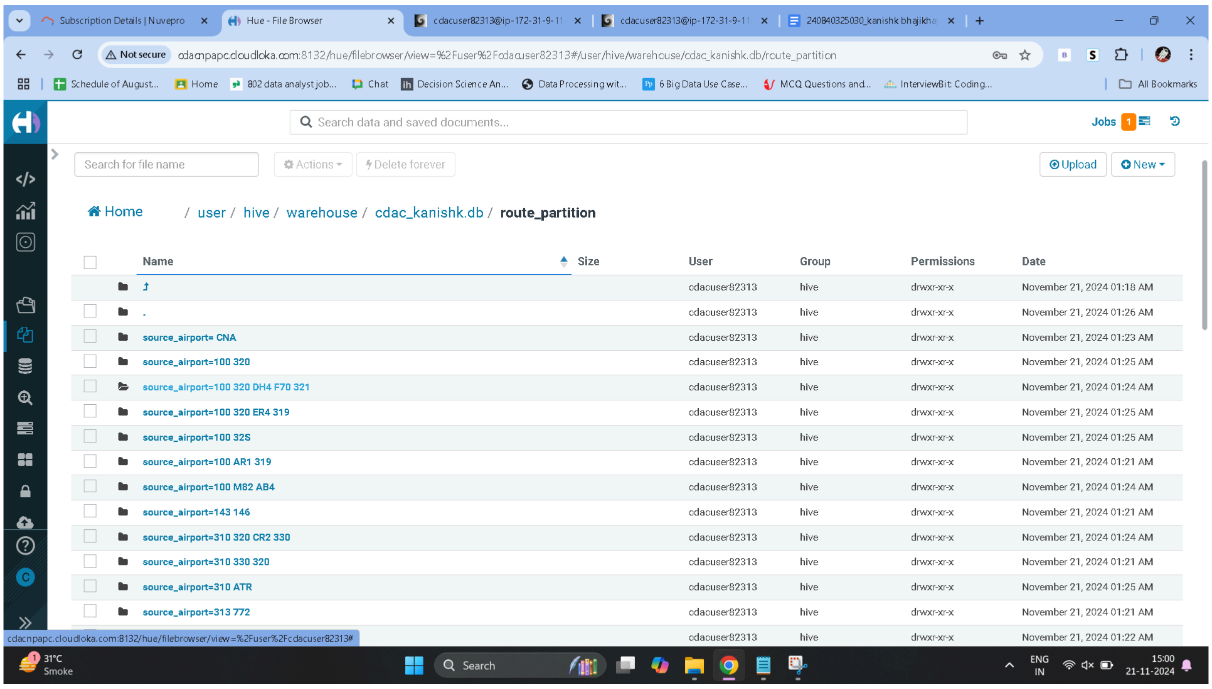
```
2024-11-21 10:20:19,499 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 10.86 sec
2024-11-21 10:20:22,563 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.26 sec
MapReduce Total cumulative CPU time: 16 seconds 260 msec
Ended Job = job_1732089968849_2663
Launching Job 2 out of 2
Number of reduce 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_2666, Tracking URL = http://master:6318/proxy/application_1732089968849_2666/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2666
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 10:20:36,106 Stage-2 map = 0%,  reduce = 0%
2024-11-21 10:20:42,254 Stage-2 map = 50%,  reduce = 0%, Cumulative CPU 2.73 sec
2024-11-21 10:20:44,302 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.4 sec
2024-11-21 10:20:48,391 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.71 sec
MapReduce Total cumulative CPU time: 8 seconds 710 msec
Ended Job = job_1732089968849_2666
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 16.26 sec   HDFS Read: 2408259 HDFS Write: 12500 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 8.71 sec   HDFS Read: 24181 HDFS Write: 287 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 970 msec
OK
FR      2484
AA      2354
UA      2180
DL      1981
US      1960
CZ      1454
MU      1263
CA      1260
WN      1146
U2      1130
Time taken: 55.121 seconds, Fetched: 10 row(s)
hive (cdac_kanishk)>
```

Question 2:

```
1: create table route_partition (airline string,airline_id
int,source_airport_id int,destination_airport string,
destination_airport_id int,codeshare
string,  stops int   ,equipment string  ) partitioned by
(source_airport string) row format delimited fields terminated by
"," stored as textfile;

insert overwrite table route_partition partition(source_airport)
select airline ,airline_id ,source_airport, source_airport_id ,
destination_airpor
t,destination_airport_id ,codeshare , stops , equipment from
route1 distribute by source_airport;
```



```
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:257 mismatched input 'terminited' expecting TERMINATED near 'fields' in table row format's field separator
hive (cdac_kanishk)> create table route_partition1 (airline string,airline_id int,source_airport_id int,destination_airport string, destination_airport_id int,codeshare
 string,  stops int   ,equipment string  ) partitioned by (source_airport string) row format delimited fields terminated by "," stored as textfile;
OK
Time taken: 0.299 seconds
hive (cdac_kanishk)> insert overwrite table route_partition partition(source_airport) select airline ,airline_id ,source_airport, source_airport_id , destination_airpor
t,destination_airport_id ,codeshare , stops , equipment from route1 distribute by source_airport;
Query ID = cdacuser82313_20241121091950_002e4d63-706c-4fe9-b01d-d7a8c9cb7210
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_2467, Tracking URL = http://master:6318/proxy/application_1732089968849_2467/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2467
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 09:20:01,918 Stage-1 map = 0%,  reduce = 0%
2024-11-21 09:20:10,181 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.25 sec
2024-11-21 09:20:26,632 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 21.19 sec
2024-11-21 09:20:28,684 Stage-1 map = 100%,  reduce = 50%, Cumulative CPU 29.98 sec
2024-11-21 09:20:29,710 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 39.46 sec
2024-11-21 09:20:47,124 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 61.42 sec
2024-11-21 09:20:57,375 Stage-1 map = 100%,  reduce = 92%, Cumulative CPU 69.41 sec
2024-11-21 09:20:59,421 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 73.53 sec
2024-11-21 09:21:59,707 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 103.02 sec
2024-11-21 09:22:59,991 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 121.71 sec
2024-11-21 09:24:00,184 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 138.53 sec
2024-11-21 09:26:25,652 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 164.45 sec
MapReduce Total cumulative CPU time: 2 minutes 44 seconds 450 msec
Ended Job = job_1732089968849_2467
Launching Job 2 out of 2
Loading data to table cdac_kanishk.route_partition partition (source_airport=null)
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
```

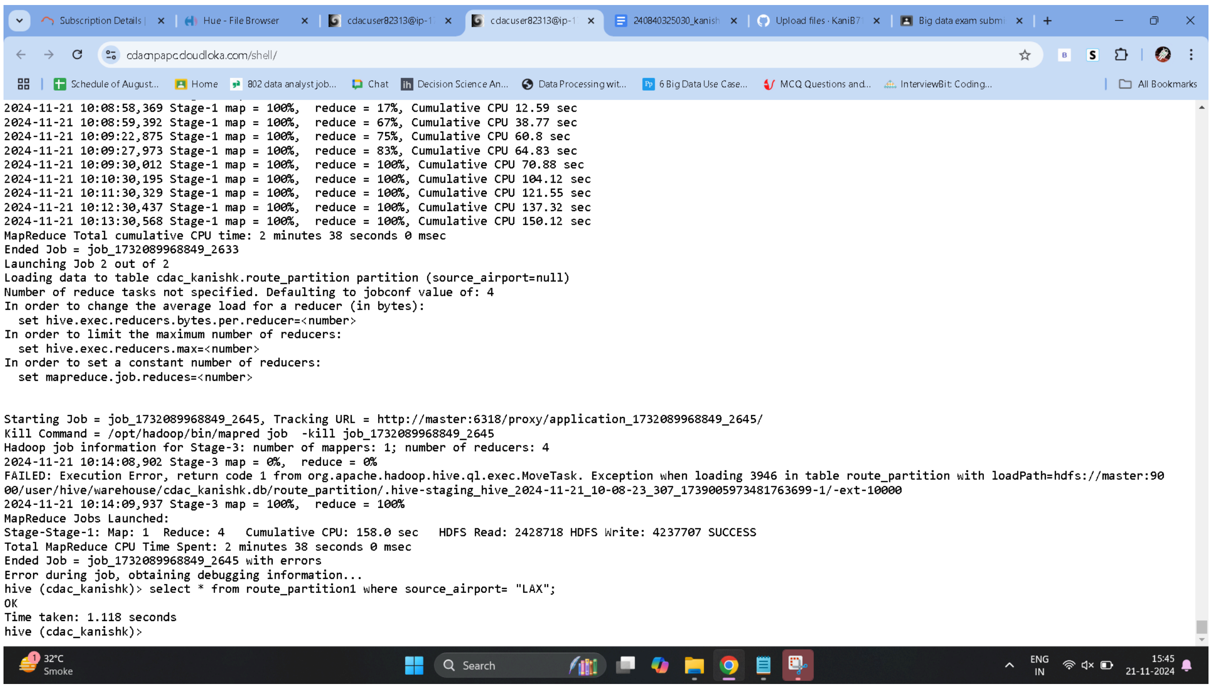


Home / user / hive / warehouse / cdac_kanishk.db / route_partition

Name	Size	User	Group	Permissions	Date
↥		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:18 AM
.		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:26 AM
source_airport= CNA		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:23 AM
source_airport=100 320		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:25 AM
source_airport=100 320 DH4 F70 321		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:24 AM
source_airport=100 320 ER4 319		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:25 AM
source_airport=100 32S		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:25 AM
source_airport=100 AR1 319		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:21 AM
source_airport=100 M82 AB4		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:24 AM
source_airport=143 146		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:21 AM
source_airport=310 320 CR2 330		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:24 AM
source_airport=310 330 320		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:21 AM
source_airport=310 ATR		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:25 AM
source_airport=313 772		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:21 AM
		cdacuser82313	hive	drwxr-xr-x	November 21, 2024 01:22 AM

3.

Select * from routes_partition where source_airport = 'LAX'



```
2024-11-21 10:08:58,369 Stage-1 map = 100%,  reduce = 17%, Cumulative CPU 12.59 sec
2024-11-21 10:08:59,392 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 38.77 sec
2024-11-21 10:09:22,875 Stage-1 map = 100%,  reduce = 75%, Cumulative CPU 60.8 sec
2024-11-21 10:09:27,973 Stage-1 map = 100%,  reduce = 83%, Cumulative CPU 64.83 sec
2024-11-21 10:09:30,012 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 70.88 sec
2024-11-21 10:10:30,195 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 104.12 sec
2024-11-21 10:11:30,329 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 121.55 sec
2024-11-21 10:12:30,437 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 137.32 sec
2024-11-21 10:13:30,568 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 150.12 sec
MapReduce Total cumulative CPU time: 2 minutes 38 seconds 0 msec
Ended Job = job_1732089968849_2633
Launching Job 2 out of 2
Loading data to table cdac_kanishk.route_partition partition (source_airport=null)
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_2645, Tracking URL = http://master:6318/proxy/application_1732089968849_2645/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2645
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 4
2024-11-21 10:14:08,902 Stage-3 map = 0%,  reduce = 0%
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask. Exception when loading 3946 in table route_partition with loadPath=hdfs://master:90
00/user/hive/warehouse/cdac_kanishk.db/route_partition/.hive-staging_hive_2024-11-21_10-08-23_307_1739005973481763699-1/-ext-10000
2024-11-21 10:14:09,937 Stage-3 map = 100%,  reduce = 100%
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 158.0 sec   HDFS Read: 2428718 HDFS Write: 4237707 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 38 seconds 0 msec
Ended Job = job_1732089968849_2645 with errors
Error during job, obtaining debugging information...
hive (cdac_kanishk)> select * from route_partition1 where source_airport= "LAX";
OK
Time taken: 1.118 seconds
hive (cdac_kanishk)>
```

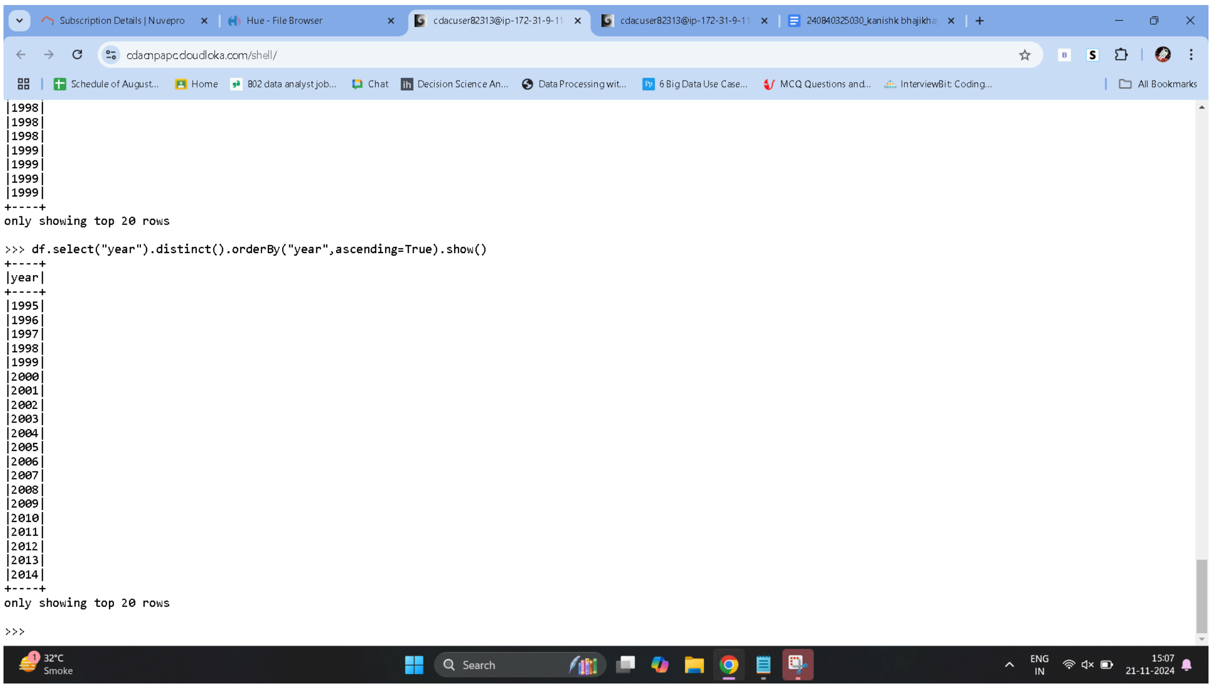
Spark:

Question 1:

```
df=spark.read.format('csv').option("header",True).option("inferSchema",True).load("/user/cdacuser82313/airlinedata.csv")
>>> df.count()
```

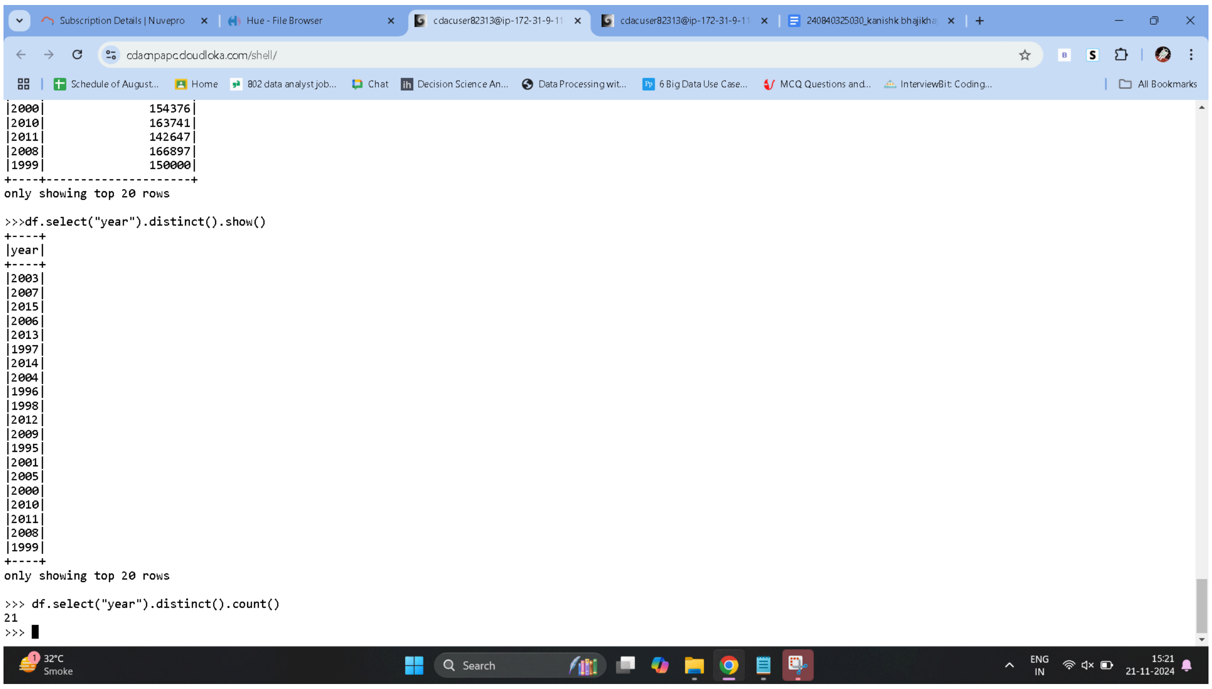
```
84
```

```
df.select("year").distinct().orderBy("year",ascending=True).show()
```



```
|1998|
|1998|
|1998|
|1999|
|1999|
|1999|
|1999|
+----+
only showing top 20 rows

>>> df.select("year").distinct().orderBy("year",ascending=True).show()
+----+
|year|
+----+
|1995|
|1996|
|1997|
|1998|
|1999|
|2000|
|2001|
|2002|
|2003|
|2004|
|2005|
|2006|
|2007|
|2008|
|2009|
|2010|
|2011|
|2012|
|2013|
|2014|
+----+
only showing top 20 rows

>>>
```

```
df.select("year").distinct().count()
```

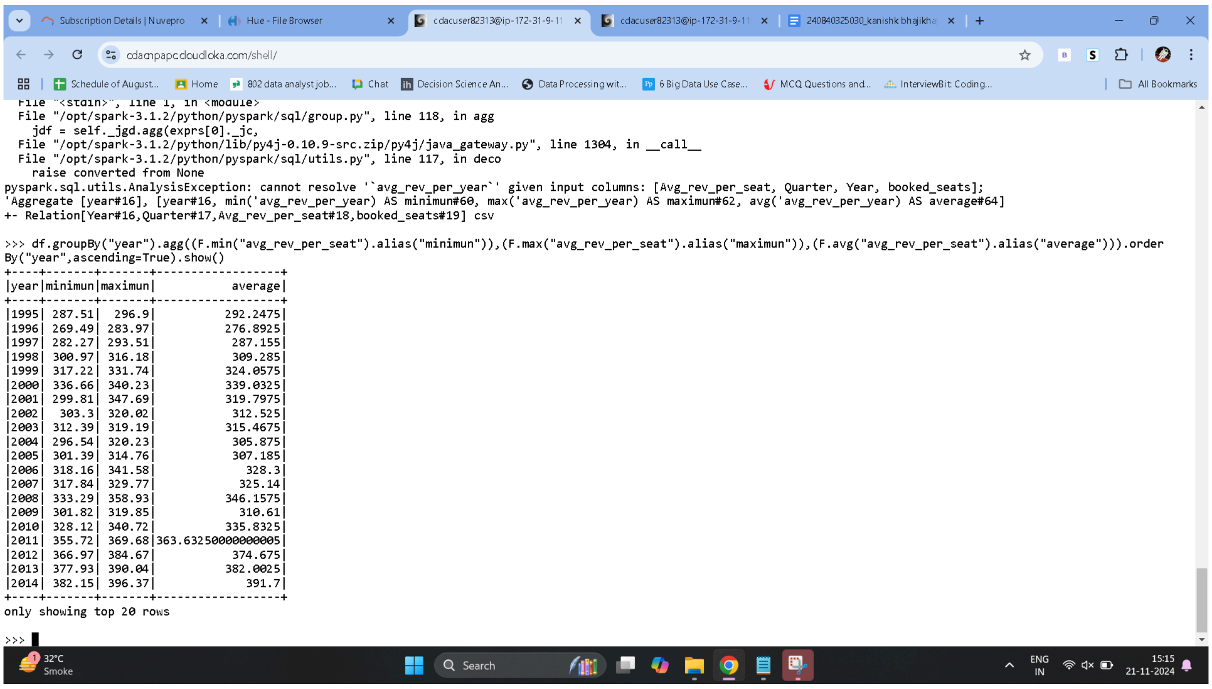


```
|2000|              154376|
|2010|              163741|
|2011|              142647|
|2008|              166897|
|1999|              150000|
+----+--------------------+
only showing top 20 rows

>>>df.select("year").distinct().show()
+----+
|year|
+----+
|2003|
|2007|
|2015|
|2006|
|2013|
|1997|
|2014|
|2004|
|1996|
|1998|
|2012|
|2009|
|1995|
|2001|
|2005|
|2000|
|2010|
|2011|
|2008|
|1999|
+----+
only showing top 20 rows

>>> df.select("year").distinct().count()
21
>>> 
```

Question 2:

1

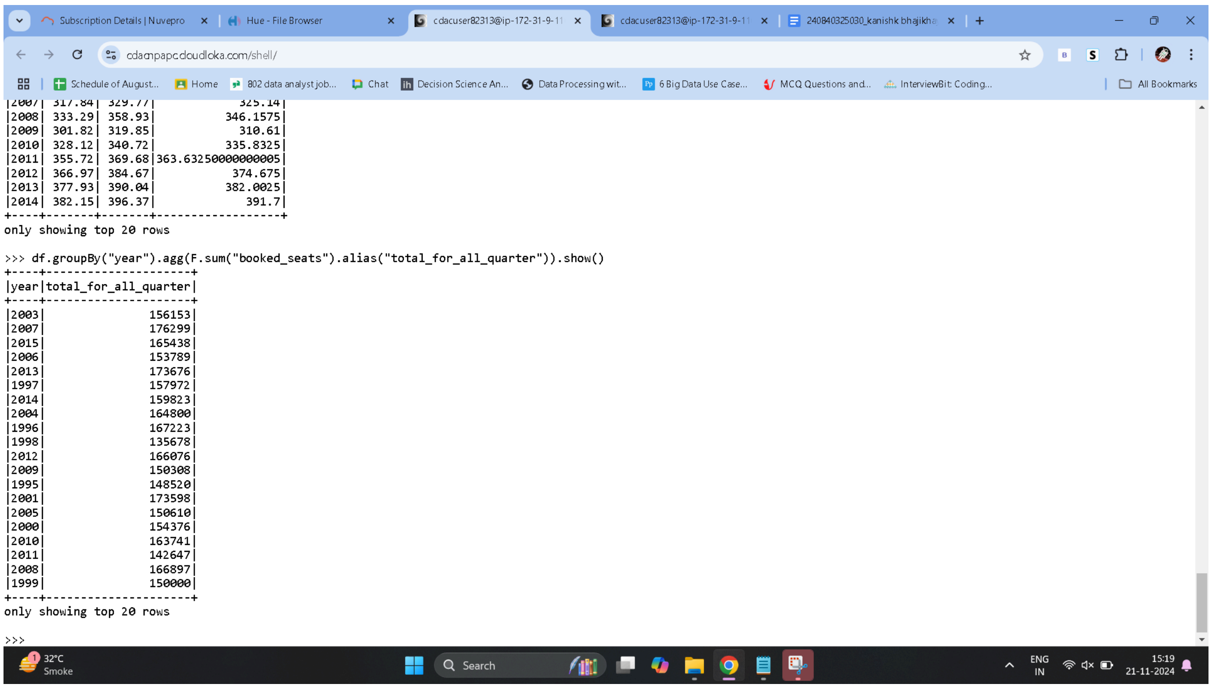
```
df.groupBy("year").agg((F.min("avg_rev_per_seat").alias("minimun")),(F.max("avg_rev_per_seat").alias("maximun")),(F.avg("avg_rev_per_seat").alias("average"))).order
By("year",ascending=True).show()
```



```
  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 '`avg_rev_per_year`' given input columns: [Avg_rev_per_seat, Quarter, Year, booked_seats];
'Aggregate [year#16], [year#16, min('avg_rev_per_year) AS minimun#60, max('avg_rev_per_year) AS maximun#62, avg('avg_rev_per_year) AS average#64]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupBy("year").agg((F.min("avg_rev_per_seat").alias("minimun")),(F.max("avg_rev_per_seat").alias("maximun")),(F.avg("avg_rev_per_seat").alias("average"))).order
By("year",ascending=True).show()
+----+-------+-------+-----------------+
|year|minimun|maximun|          average|
+----+-------+-------+-----------------+
|1995| 287.51|  296.9|         292.2475|
|1996| 269.49| 283.97|         276.8925|
|1997| 282.27| 293.51|          287.155|
|1998| 300.97| 316.18|          309.285|
|1999| 317.22| 331.74|         324.0575|
|2000| 336.66| 340.23|         339.0325|
|2001| 299.81| 347.69|         319.7975|
|2002|  303.3| 320.02|          312.525|
|2003| 312.39| 319.19|         315.4675|
|2004| 296.54| 320.23|          305.875|
|2005| 301.39| 314.76|          307.185|
|2006| 318.16| 341.58|            328.3|
|2007| 317.84| 329.77|           325.14|
|2008| 333.29| 358.93|         346.1575|
|2009| 301.82| 319.85|           310.61|
|2010| 328.12| 340.72|         335.8325|
|2011| 355.72| 369.68|363.6325000000005|
|2012| 366.97| 384.67|          374.675|
|2013| 377.93| 390.04|         382.0025|
|2014| 382.15| 396.37|            391.7|
+----+-------+-------+-----------------+
only showing top 20 rows

>>> 
```

3

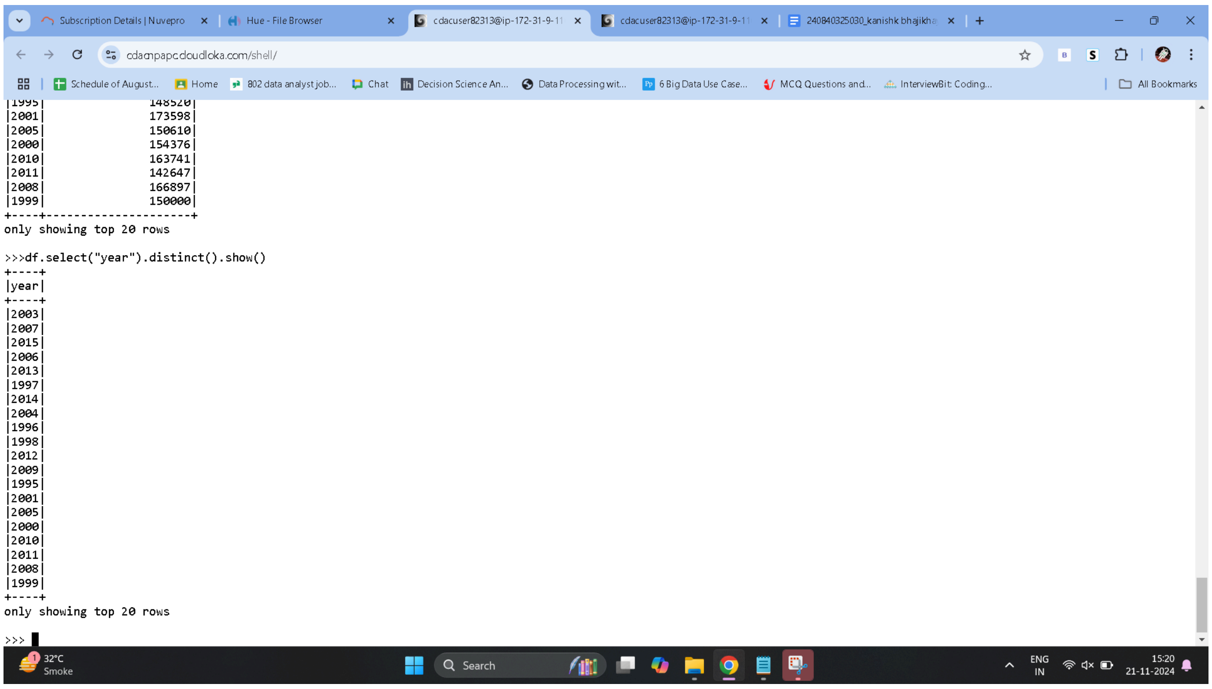
df.groupBy("year").agg(F.sum("booked_seats").alias("total_for_all_quarter")).show()



```
|2007| 317.84| 329.77|            325.14|
|2008| 333.29| 358.93|          346.1575|
|2009| 301.82| 319.85|            310.61|
|2010| 328.12| 340.72|          335.8325|
|2011| 355.72| 369.68|363.63250000000005|
|2012| 366.97| 384.67|           374.675|
|2013| 377.93| 390.04|          382.0025|
|2014| 382.15| 396.37|             391.7|
+----+-------+-------+------------------+
only showing top 20 rows

>>> df.groupBy("year").agg(F.sum("booked_seats").alias("total_for_all_quarter")).show()
+----+---------------------+
|year|total_for_all_quarter|
+----+---------------------+
|2003|               156153|
|2007|               176299|
|2015|               165438|
|2006|               153789|
|2013|               173676|
|1997|               157972|
|2014|               159823|
|2004|               164800|
|1996|               167223|
|1998|               135678|
|2012|               166076|
|2009|               150308|
|1995|               148520|
|2001|               173598|
|2005|               150610|
|2000|               154376|
|2010|               163741|
|2011|               142647|
|2008|               166897|
|1999|               150000|
+----+---------------------+
only showing top 20 rows

>>>
```

4

>>>df.select("year").distinct().show()

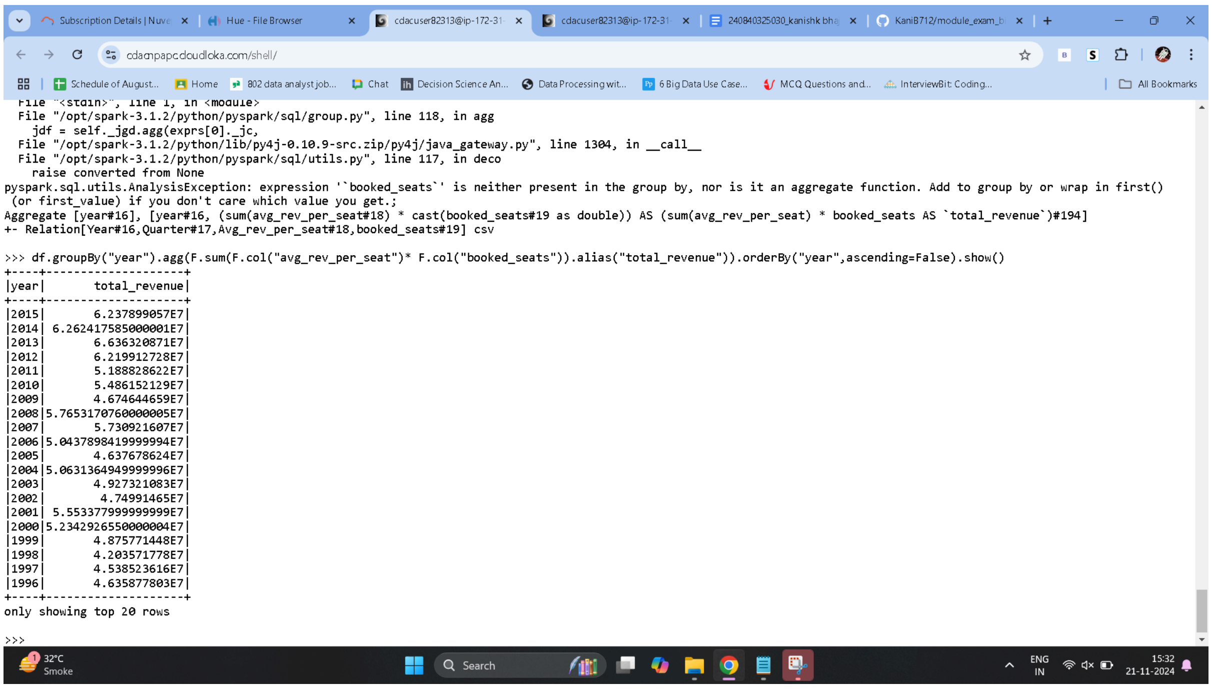


```
|1995|               148520|
|2001|               173598|
|2005|               150610|
|2000|               154376|
|2010|               163741|
|2011|               142647|
|2008|               166897|
|1999|               150000|
+----+---------------------+
only showing top 20 rows

>>>df.select("year").distinct().show()
+----+
|year|
+----+
|2003|
|2007|
|2015|
|2006|
|2013|
|1997|
|2014|
|2004|
|1996|
|1998|
|2012|
|2009|
|1995|
|2001|
|2005|
|2000|
|2010|
|2011|
|2008|
|1999|
+----+
only showing top 20 rows

>>>
```

5

>>> df.groupBy("year").agg(F.sum(F.col("avg_rev_per_seat")* F.col("booked_seats")).alias("total_revenue")).orderBy("year",asce

```
nding=False).show()
```



```
  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: expression '`booked_seats`' is neither present in the group by, nor is it an aggregate function. Add to group by or wrap in first()
 (or first_value) if you don't care which value you get.;
Aggregate [year#16], [year#16, (sum(avg_rev_per_seat#18) * cast(booked_seats#19 as double)) AS (sum(avg_rev_per_seat) * booked_seats AS `total_revenue`)#194]
+- Relation[Year#16,Quarter#17,Avg_rev_per_seat#18,booked_seats#19] csv

>>> df.groupBy("year").agg(F.sum(F.col("avg_rev_per_seat")* F.col("booked_seats")).alias("total_revenue")).orderBy("year",ascending=False).show()
+----+--------------------+
|year|       total_revenue|
+----+--------------------+
|2015|        6.237899057E7|
|2014| 6.262417585000001E7|
|2013|        6.636320871E7|
|2012|        6.219912728E7|
|2011|        5.188828622E7|
|2010|        5.486152129E7|
|2009|        4.674644659E7|
|2008|5.765317076000005E7|
|2007|        5.730921607E7|
|2006|5.043789841999994E7|
|2005|        4.637678624E7|
|2004|5.063136494999996E7|
|2003|        4.927321083E7|
|2002|         4.74991465E7|
|2001| 5.553377999999999E7|
|2000|5.234292655000004E7|
|1999|        4.875771448E7|
|1998|        4.203571778E7|
|1997|        4.538523616E7|
|1996|        4.635877803E7|
+----+--------------------+
only showing top 20 rows

>>>
```