AWS HADOOP

Big Data Analytics Project Phase 1

YASH JIVANI

PARTH DODIA

DATA ANALYSIS TASKS

Case 1

Determine the three **airports** with the highest delay time (in hours) for assigned year

Case 2

Determine the three **carriers** with the highest delay time (in hours) for assigned year

Case 3

Determine overall which type of delay (**arrivals** or **departures**) is the largest for your carriers

AIRPORT DELAY

1999

YASH JIVANI

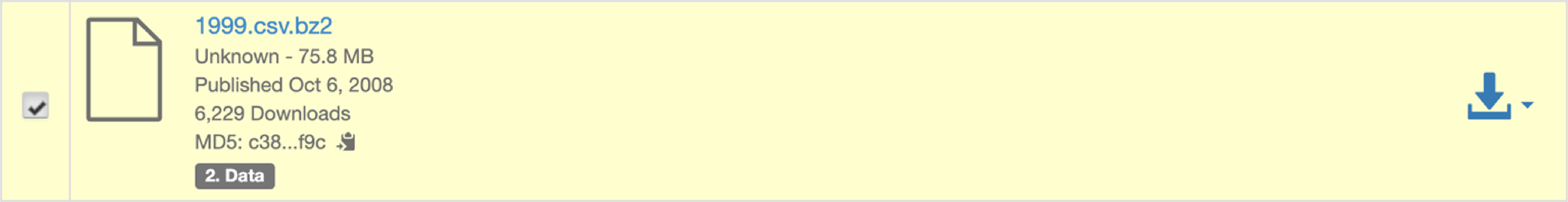
DATA SOURCE

Data Expo 2009: Airline on Time Data

The data represents flight arrival and departure details for all commercial flights within the USA for year 1999

Download Link

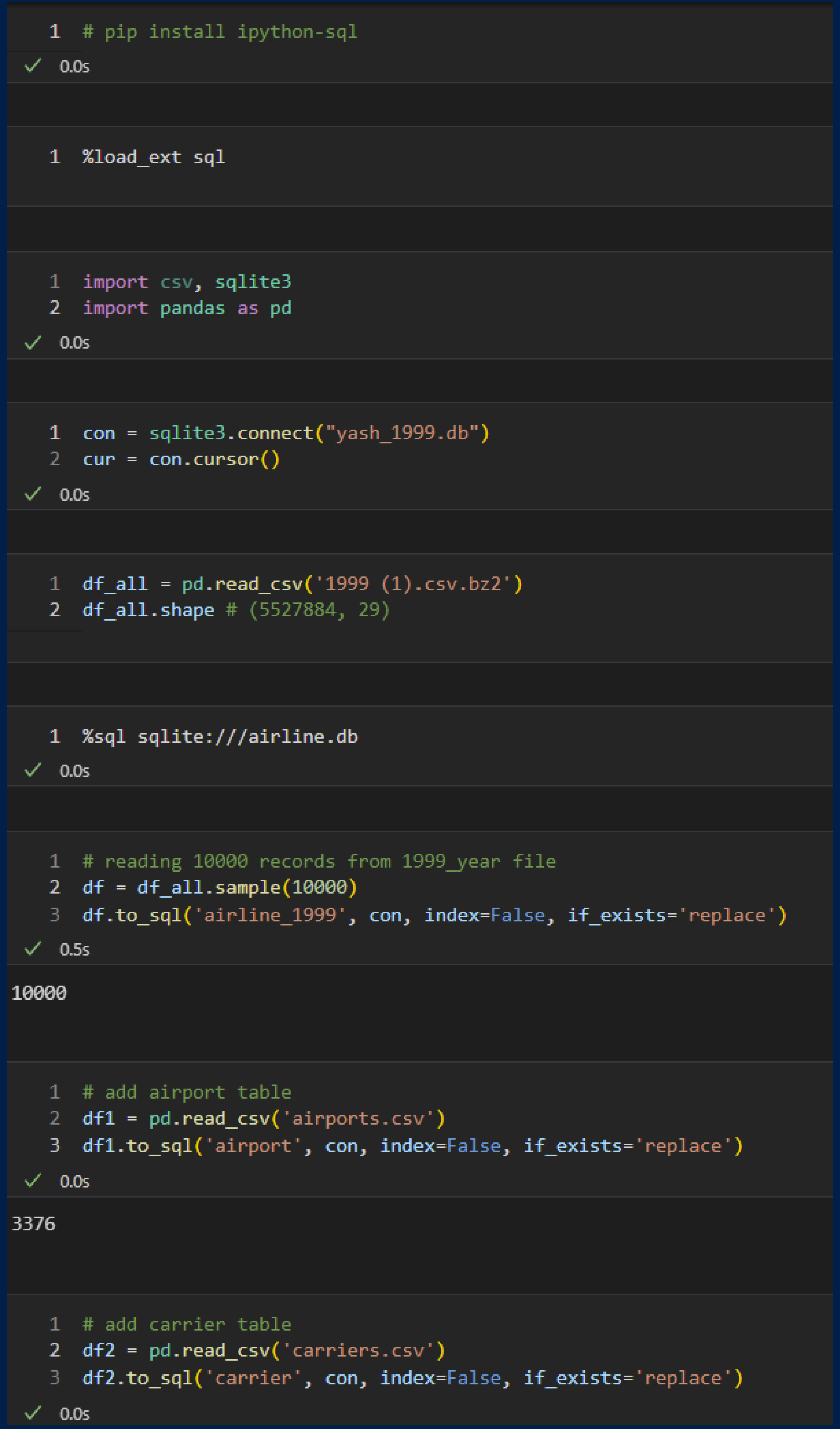
Data Expo 2009: Harward - Airline on Time Data



☑ 1999.csv.bz2
Unknown - 75.8 MB
Published Oct 6, 2008
6,229 Downloads
MD5: c38...f9c
2. Data

Yash Jivani

Sample Data load in Python using Pandas



```
# pip install ipython-sql
```

✓ 0.0s

```
%load_ext sql
```

```
import csv, sqlite3
import pandas as pd
```

✓ 0.0s

```
con = sqlite3.connect("yash_1999.db")
cur = con.cursor()
```

✓ 0.0s

```
df_all = pd.read_csv('1999 (1).csv.bz2')
df_all.shape # (5527884, 29)
```

```
%sql sqlite:///airline.db
```

✓ 0.0s

```
# reading 10000 records from 1999_year file
df = df_all.sample(10000)
df.to_sql('airline_1999', con, index=False, if_exists='replace')
```

✓ 0.5s

```
10000
```

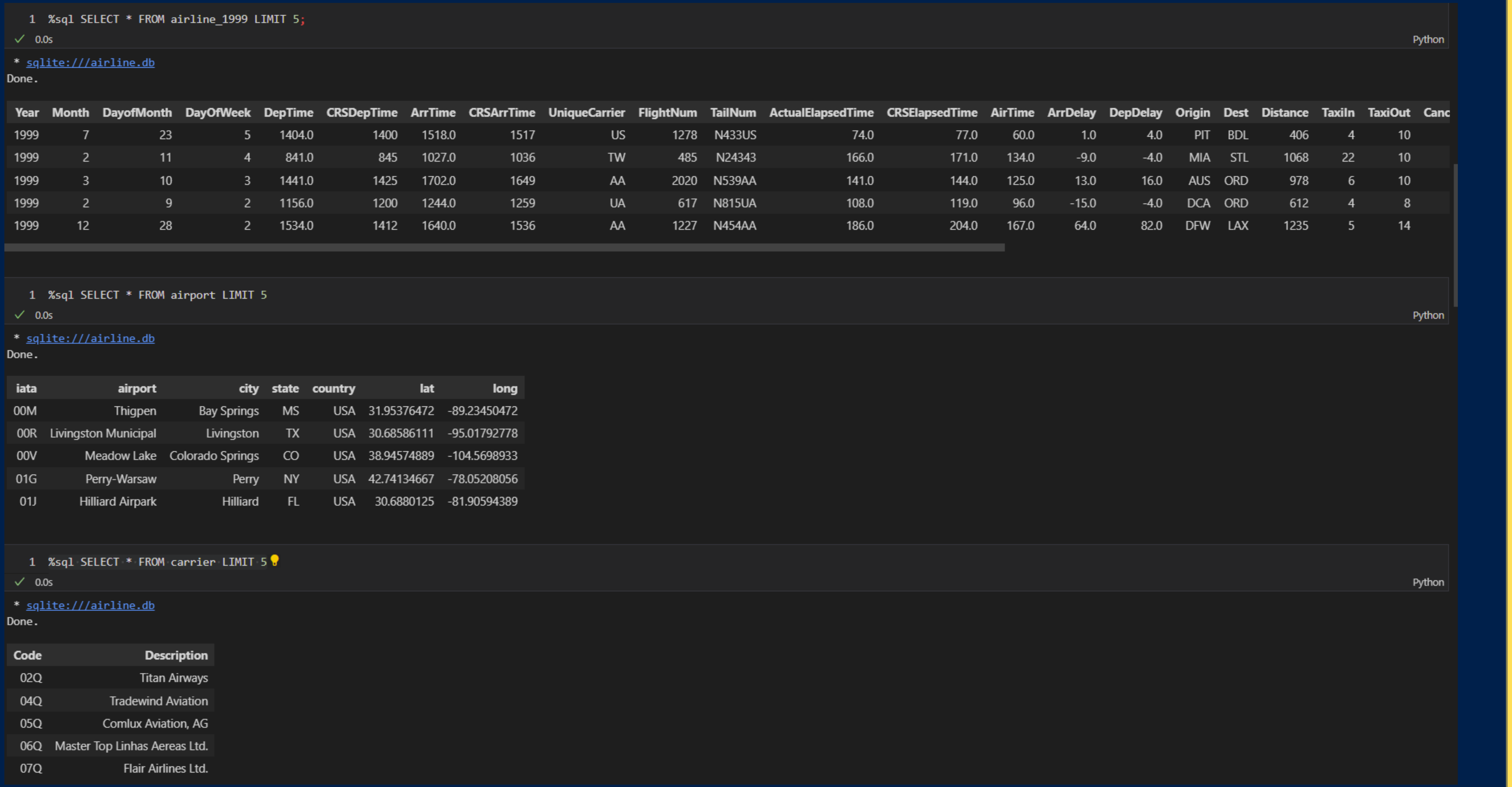
```
# add airport table
df1 = pd.read_csv('airports.csv')
df1.to_sql('airport', con, index=False, if_exists='replace')
```

✓ 0.0s

```
3376
```

```
# add carrier table
df2 = pd.read_csv('carriers.csv')
df2.to_sql('carrier', con, index=False, if_exists='replace')
```

✓ 0.0s



```
%sql SELECT * FROM airline_1999 LIMIT 5;
```

✓ 0.0s

```
 * sqlite:///airline.db
Done.
```

Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Canc
1999	7	23	5	1404.0	1400	1518.0	1517	US	1278	N433US	74.0	77.0	60.0	1.0	4.0	PIT	BDL	406	4	10	
1999	2	11	4	841.0	845	1027.0	1036	TW	485	N24343	166.0	171.0	134.0	-9.0	-4.0	MIA	STL	1068	22	10	
1999	3	10	3	1441.0	1425	1702.0	1649	AA	2020	N539AA	141.0	144.0	125.0	13.0	16.0	AUS	ORD	978	6	10	
1999	2	9	2	1156.0	1200	1244.0	1259	UA	617	N815UA	108.0	119.0	96.0	-15.0	-4.0	DCA	ORD	612	4	8	
1999	12	28	2	1534.0	1412	1640.0	1536	AA	1227	N454AA	186.0	204.0	167.0	64.0	82.0	DFW	LAX	1235	5	14	

```
%sql SELECT * FROM airport LIMIT 5
```

✓ 0.0s

```
 * sqlite:///airline.db
Done.
```

iata	airport	city	state	country	lat	long
00M	Thigpen	Bay Springs	MS	USA	31.95376472	-89.23450472
00R	Livingston Municipal	Livingston	TX	USA	30.68586111	-95.01792778
00V	Meadow Lake	Colorado Springs	CO	USA	38.94574889	-104.5698933
01G	Perry-Warsaw	Perry	NY	USA	42.74134667	-78.05208056
01J	Hilliard Airpark	Hilliard	FL	USA	30.6880125	-81.90594389

```
%sql SELECT * FROM carrier LIMIT 5
```

✓ 0.0s

```
 * sqlite:///airline.db
Done.
```

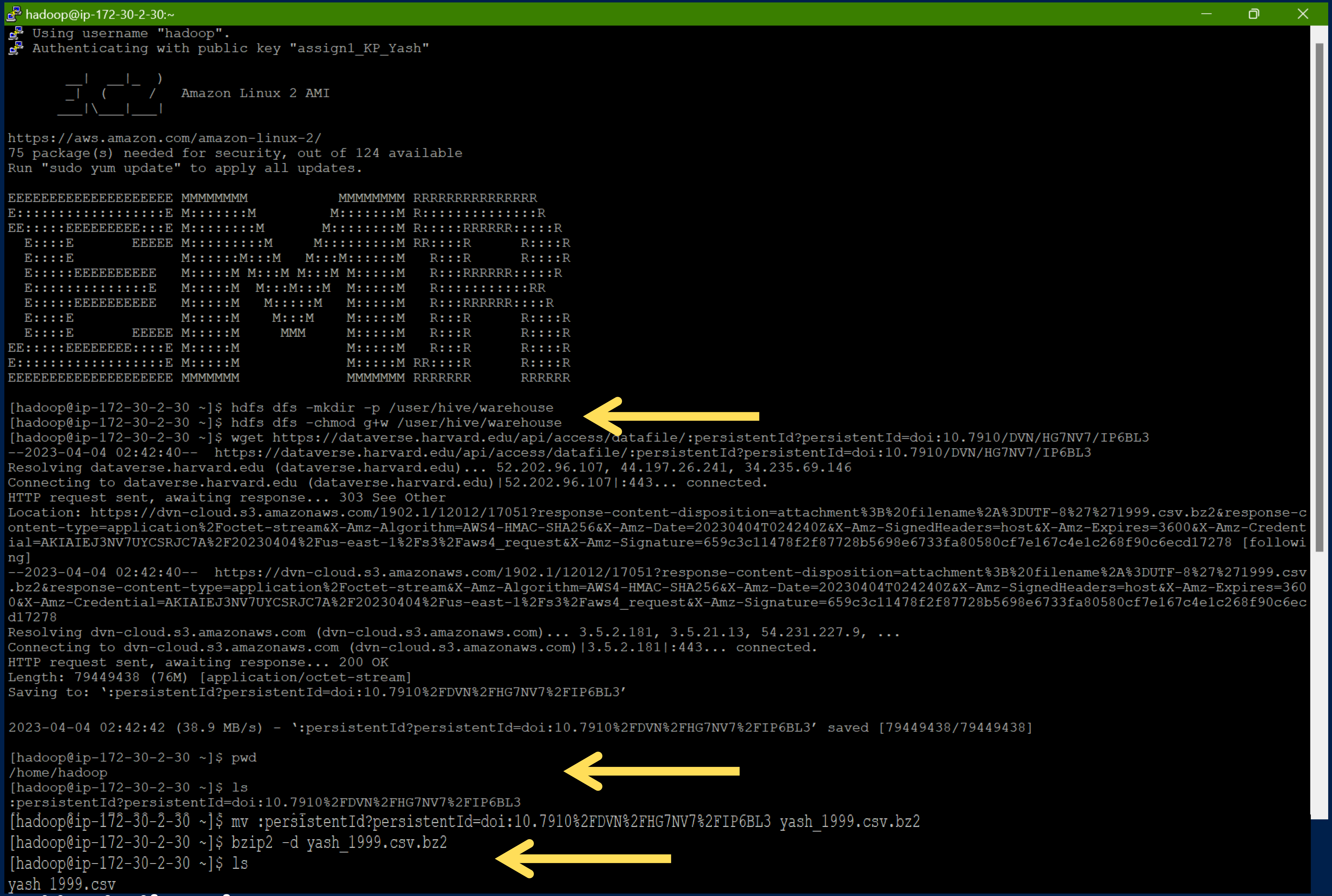
Code	Description
02Q	Titan Airways
04Q	Tradewind Aviation
05Q	Comlux Aviation, AG
06Q	Master Top Linhas Aereas Ltd.
07Q	Flair Airlines Ltd.

Yash Jivani

Steps for working with Data in Hadoop

1. hdfs dfs -mkdir -p /user/hive/warehouse
2. hdfs dfs -chmod g+w /user/hive/warehouse
3. wget https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/IP6BL3
4. mv :persistentId\?persistentId\=doi\:10.7910%2FDVN%2FHG7NV7%2FIP6BL3 yash_1999.csv.bz2
5. bzip2 -d yash_1999.csv.bz2
6. wget https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/XTPZZY
7. mv :persistentId\?persistentId\=doi\:10.7910%2FDVN%2FHG7NV7%2FXTPZZY airports.csv
8. wget https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/3NOQ6Q
9. mv :persistentId\?persistentId\=doi\:10.7910%2FDVN%2FHG7NV7%2F3NOQ6Q carriers.csv
10. pwd
11. ls
12. Used hive command to navigate to hive shell for performing SQL query for creating and getting the csv data into tables and perform analysis.

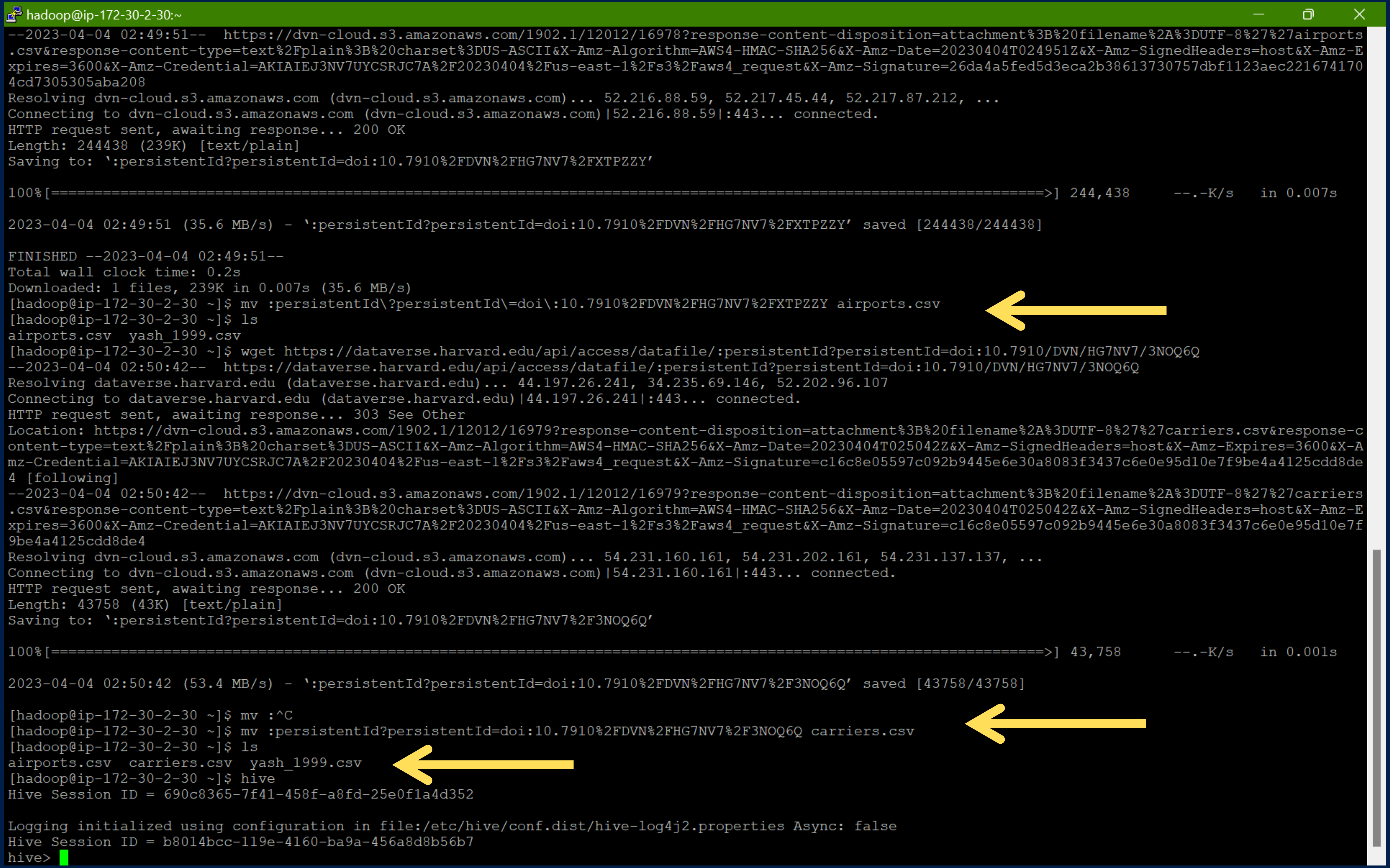
Steps for working with Data in Hadoop



```
Using username "hadoop".
Authenticating with public key "assign1_KP_Yash"

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
75 package(s) needed for security, out of 124 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-30-2-30 ~]$ hdfs dfs -mkdir -p /user/hive/warehouse
[hadoop@ip-172-30-2-30 ~]$ hdfs dfs -chmod g+w /user/hive/warehouse
[hadoop@ip-172-30-2-30 ~]$ wget https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/IP6BL3
--2023-04-04 02:42:40--  https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/IP6BL3
Resolving dataverse.harvard.edu (dataverse.harvard.edu)... 52.202.96.107, 44.197.26.241, 34.235.69.146
Connecting to dataverse.harvard.edu (dataverse.harvard.edu)|52.202.96.107|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://dvn-cloud.s3.amazonaws.com/1902.1/12012/17051?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%271999.csv.bz2&response-c
ontent-type=application%2Foctet-stream&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230404T024240Z&X-Amz-SignedHeaders=host&X-Amz-Expires=3600&X-Amz-Credent
ial=AKIAIEJ3NV7UYCSRJC7A%2F20230404%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=659c3c11478f2f87728b5698e6733fa80580cf7e167c4e1c268f90c6ecd17278 [followi
ng]
--2023-04-04 02:42:40--  https://dvn-cloud.s3.amazonaws.com/1902.1/12012/17051?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%271999.csv
.bz2&response-content-type=application%2Foctet-stream&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230404T024240Z&X-Amz-SignedHeaders=host&X-Amz-Expires=360
0&X-Amz-Credential=AKIAIEJ3NV7UYCSRJC7A%2F20230404%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=659c3c11478f2f87728b5698e6733fa80580cf7e167c4e1c268f90c6ec
d17278
Resolving dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)... 3.5.2.181, 3.5.21.13, 54.231.227.9, ...
Connecting to dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)|3.5.2.181|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 79449438 (76M) [application/octet-stream]
Saving to: ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FIP6BL3’


2023-04-04 02:42:42 (38.9 MB/s) - ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FIP6BL3’ saved [79449438/79449438]

[hadoop@ip-172-30-2-30 ~]$ pwd
/home/hadoop
[hadoop@ip-172-30-2-30 ~]$ ls
:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FIP6BL3
[hadoop@ip-172-30-2-30 ~]$ mv :persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FIP6BL3 yash_1999.csv.bz2
[hadoop@ip-172-30-2-30 ~]$ bzip2 -d yash_1999.csv.bz2
[hadoop@ip-172-30-2-30 ~]$ ls
yash 1999.csv
```

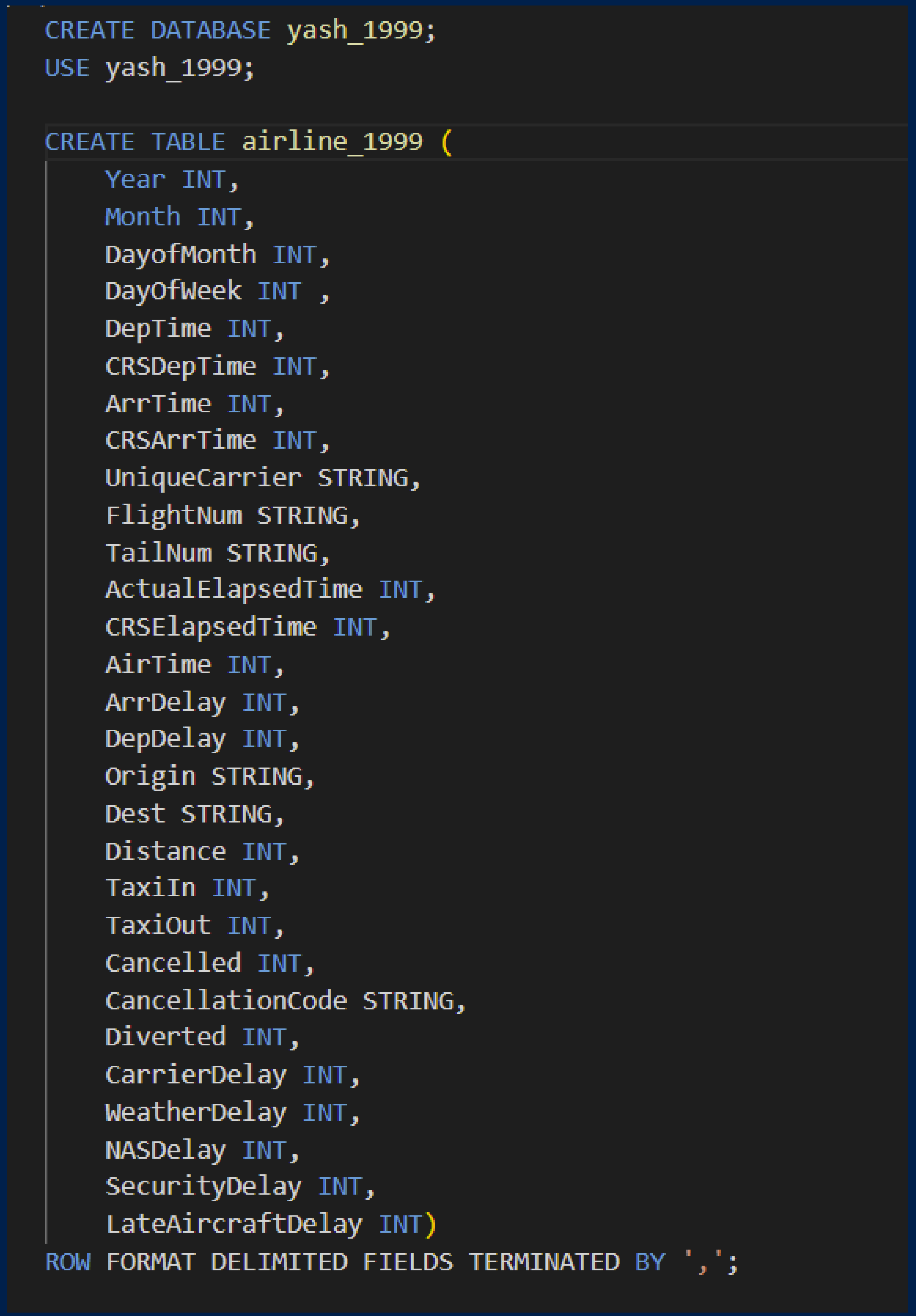
Yash Jivani

Steps for working with Data in Hadoop

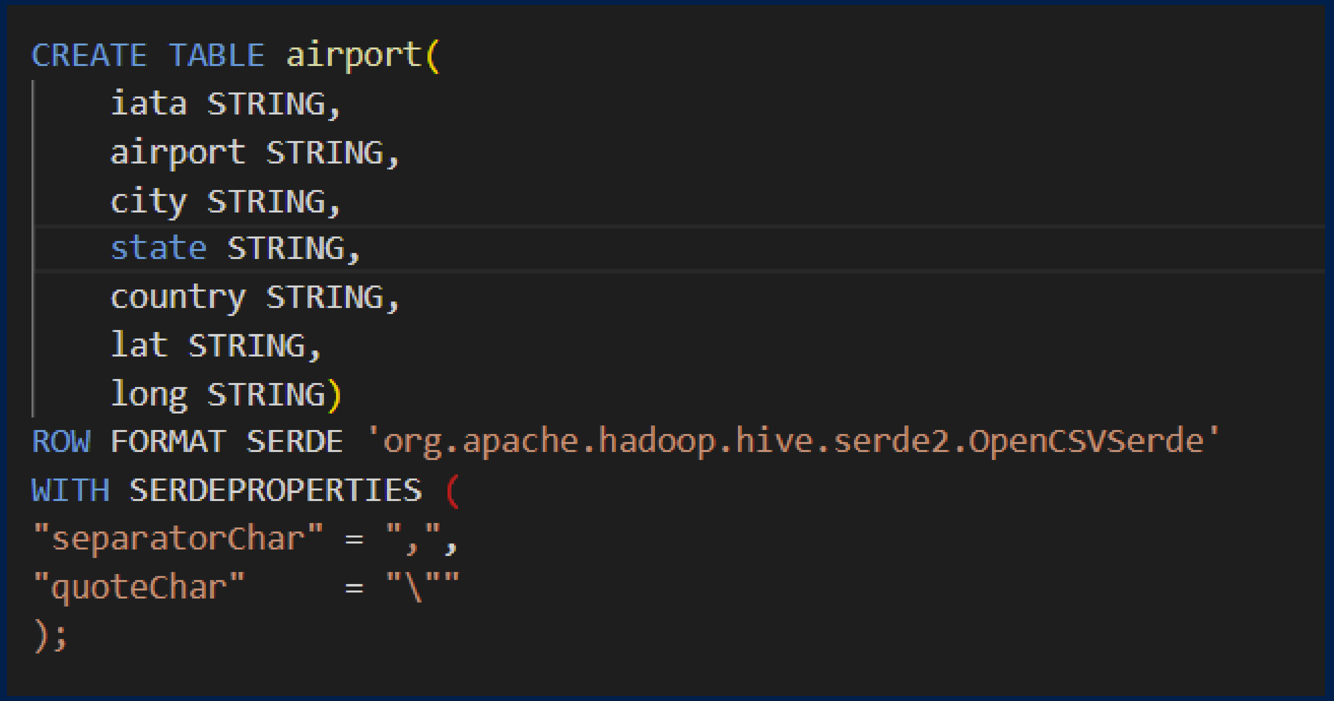


```
--2023-04-04 02:49:51--  https://dvn-cloud.s3.amazonaws.com/1902.1/12012/16978?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%27airports
.csv&response-content-type=text%2Fplain%3B%20charset%3DUS-ASCII&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230404T024951Z&X-Amz-SignedHeaders=host&X-Amz-E
xpires=3600&X-Amz-Credential=AKIAIEJ3NV7UYCSRJC7A%2F20230404%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=26da4a5fed5d3eca2b38613730757dbf1123aec221674170
4cd7305305aba208
Resolving dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)... 52.216.88.59, 52.217.45.44, 52.217.87.212, ...
Connecting to dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)|52.216.88.59|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 244438 (239K) [text/plain]
Saving to: ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FXTPZZY’

100%[=============================================================================================================>] 244,438     --.-K/s   in 0.007s

2023-04-04 02:49:51 (35.6 MB/s) - ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2FXTPZZY’ saved [244438/244438]

FINISHED --2023-04-04 02:49:51--
Total wall clock time: 0.2s
Downloaded: 1 files, 239K in 0.007s (35.6 MB/s)
[hadoop@ip-172-30-2-30 ~]$ mv :persistentId\?persistentId\=doi\:10.7910%2FDVN%2FHG7NV7%2FXTPZZY airports.csv
[hadoop@ip-172-30-2-30 ~]$ ls
airports.csv  yash_1999.csv
[hadoop@ip-172-30-2-30 ~]$ wget https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/3NOQ6Q
--2023-04-04 02:50:42--  https://dataverse.harvard.edu/api/access/datafile/:persistentId?persistentId=doi:10.7910/DVN/HG7NV7/3NOQ6Q
Resolving dataverse.harvard.edu (dataverse.harvard.edu)... 44.197.26.241, 34.235.69.146, 52.202.96.107
Connecting to dataverse.harvard.edu (dataverse.harvard.edu)|44.197.26.241|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://dvn-cloud.s3.amazonaws.com/1902.1/12012/16979?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%27carriers.csv&response-c
ontent-type=text%2Fplain%3B%20charset%3DUS-ASCII&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230404T025042Z&X-Amz-SignedHeaders=host&X-Amz-Expires=3600&X-A
mz-Credential=AKIAIEJ3NV7UYCSRJC7A%2F20230404%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=c16c8e05597c092b9445e6e30a8083f3437c6e0e95d10e7f9be4a4125cdd8de
4 [following]
--2023-04-04 02:50:42--  https://dvn-cloud.s3.amazonaws.com/1902.1/12012/16979?response-content-disposition=attachment%3B%20filename%2A%3DUTF-8%27%27carriers
.csv&response-content-type=text%2Fplain%3B%20charset%3DUS-ASCII&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20230404T025042Z&X-Amz-SignedHeaders=host&X-Amz-E
xpires=3600&X-Amz-Credential=AKIAIEJ3NV7UYCSRJC7A%2F20230404%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=c16c8e05597c092b9445e6e30a8083f3437c6e0e95d10e7f
9be4a4125cdd8de4
Resolving dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)... 54.231.160.161, 54.231.202.161, 54.231.137.137, ...
Connecting to dvn-cloud.s3.amazonaws.com (dvn-cloud.s3.amazonaws.com)|54.231.160.161|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 43758 (43K) [text/plain]
Saving to: ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2F3NOQ6Q’

100%[=============================================================================================================>] 43,758      --.-K/s   in 0.001s

2023-04-04 02:50:42 (53.4 MB/s) - ‘:persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2F3NOQ6Q’ saved [43758/43758]

[hadoop@ip-172-30-2-30 ~]$ mv :^C
[hadoop@ip-172-30-2-30 ~]$ mv :persistentId?persistentId=doi:10.7910%2FDVN%2FHG7NV7%2F3NOQ6Q carriers.csv
[hadoop@ip-172-30-2-30 ~]$ ls
airports.csv  carriers.csv  yash_1999.csv
[hadoop@ip-172-30-2-30 ~]$ hive
Hive Session ID = 690c8365-7f41-458f-a8fd-25e0f1a4d352

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
Hive Session ID = b8014bcc-119e-4160-ba9a-456a8d8b56b7
hive>
```

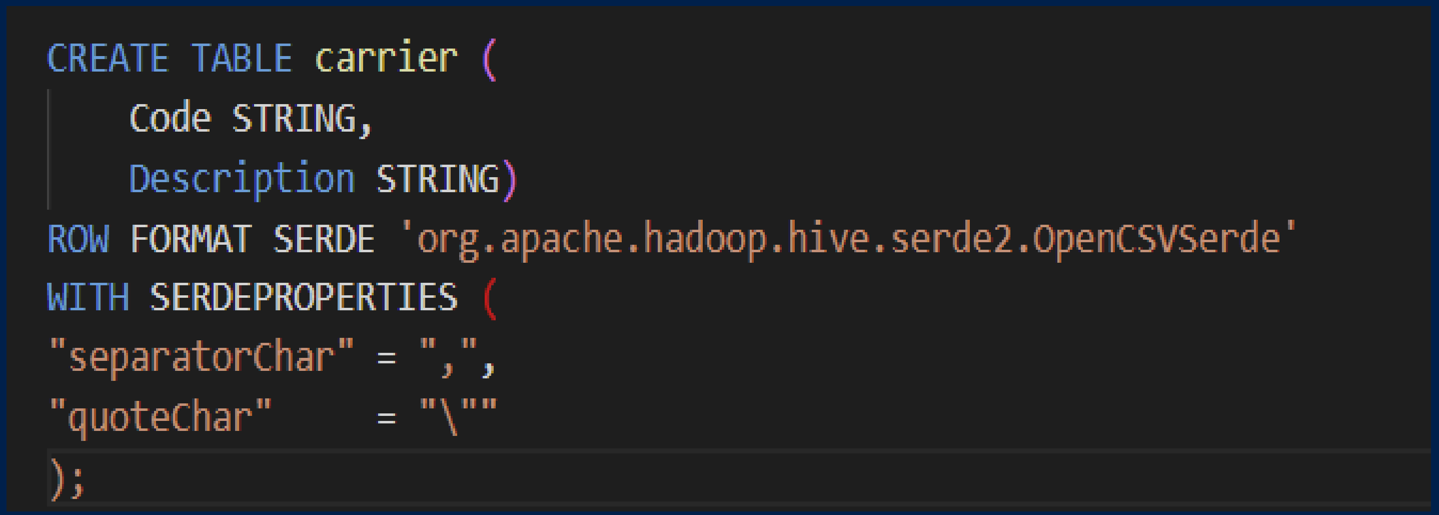
Database and Table structures



```sql
CREATE DATABASE yash_1999;
USE yash_1999;

CREATE TABLE airline_1999 (
    Year INT,
    Month INT,
    DayofMonth INT,
    DayOfWeek INT ,
    DepTime INT,
    CRSDepTime INT,
    ArrTime INT,
    CRSArrTime INT,
    UniqueCarrier STRING,
    FlightNum STRING,
    TailNum STRING,
    ActualElapsedTime INT,
    CRSElapsedTime INT,
    AirTime INT,
    ArrDelay INT,
    DepDelay INT,
    Origin STRING,
    Dest STRING,
    Distance INT,
    TaxiIn INT,
    TaxiOut INT,
    Cancelled INT,
    CancellationCode STRING,
    Diverted INT,
    CarrierDelay INT,
    WeatherDelay INT,
    NASDelay INT,
    SecurityDelay INT,
    LateAircraftDelay INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

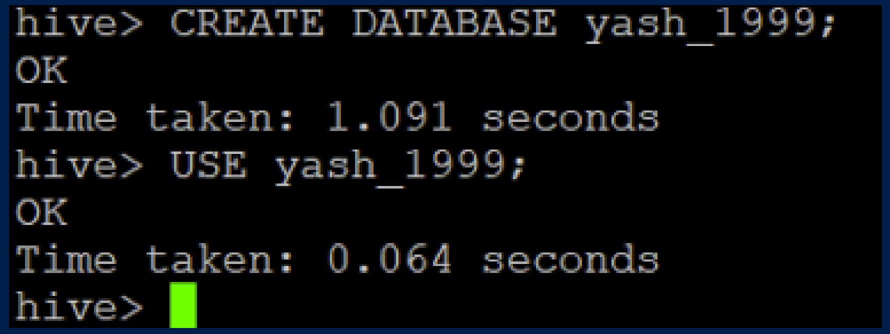


```sql
CREATE TABLE airport(
    iata STRING,
    airport STRING,
    city STRING,
    state STRING,
    country STRING,
    lat STRING,
    long STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = ",",
"quoteChar"     = "\""
);
```

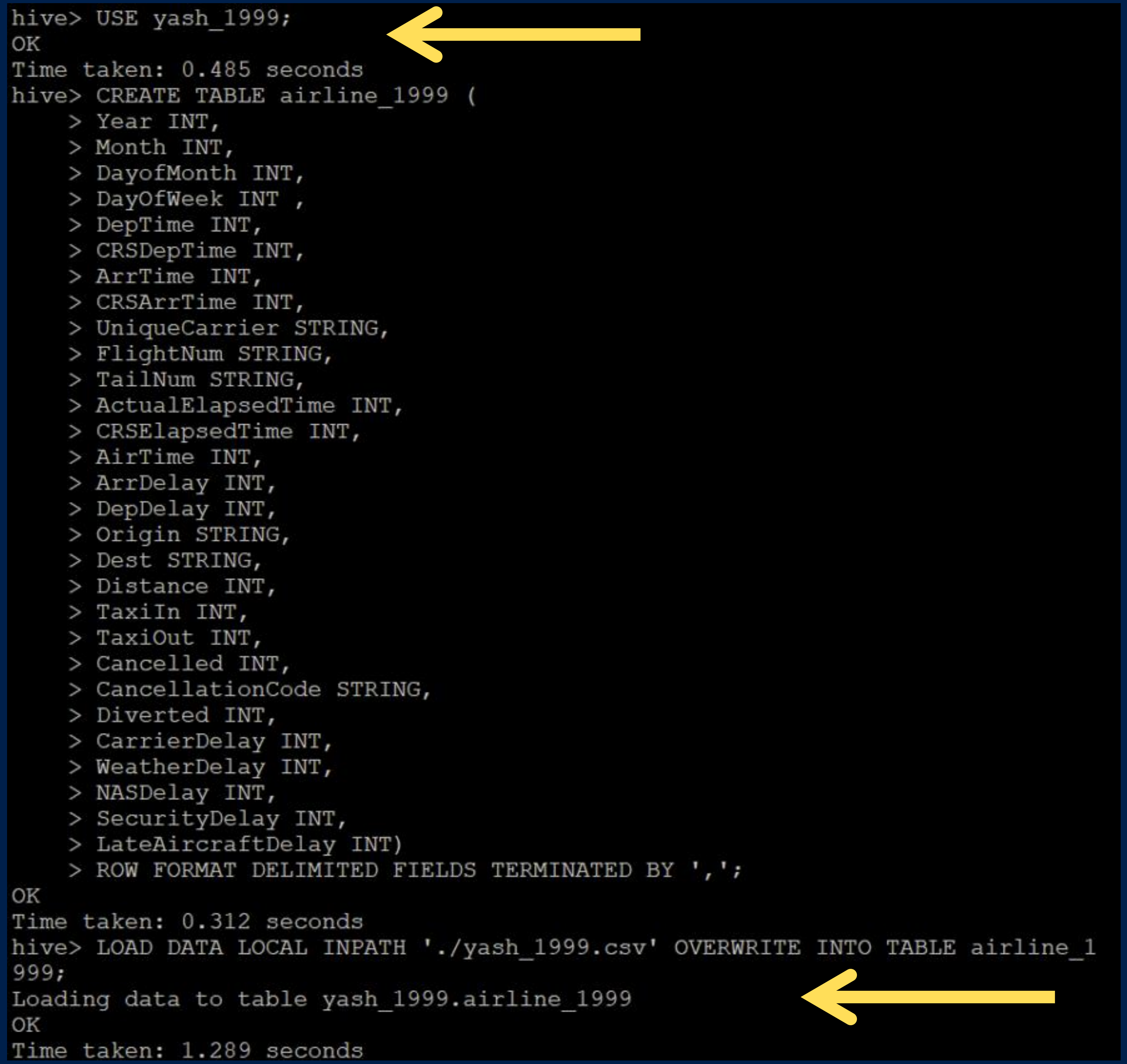


```sql
CREATE TABLE carrier (
    Code STRING,
    Description STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = ",",
"quoteChar"     = "\""
);
```

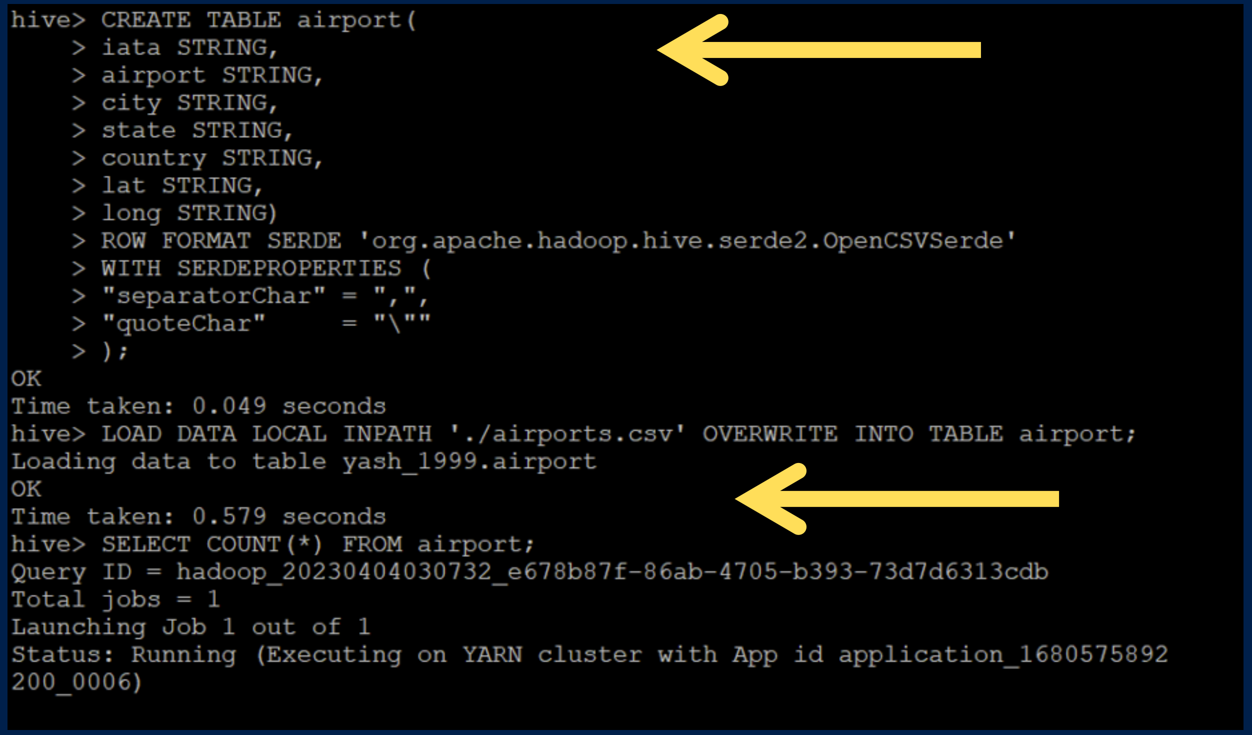
Load CSV Data Into Hadoop Tables



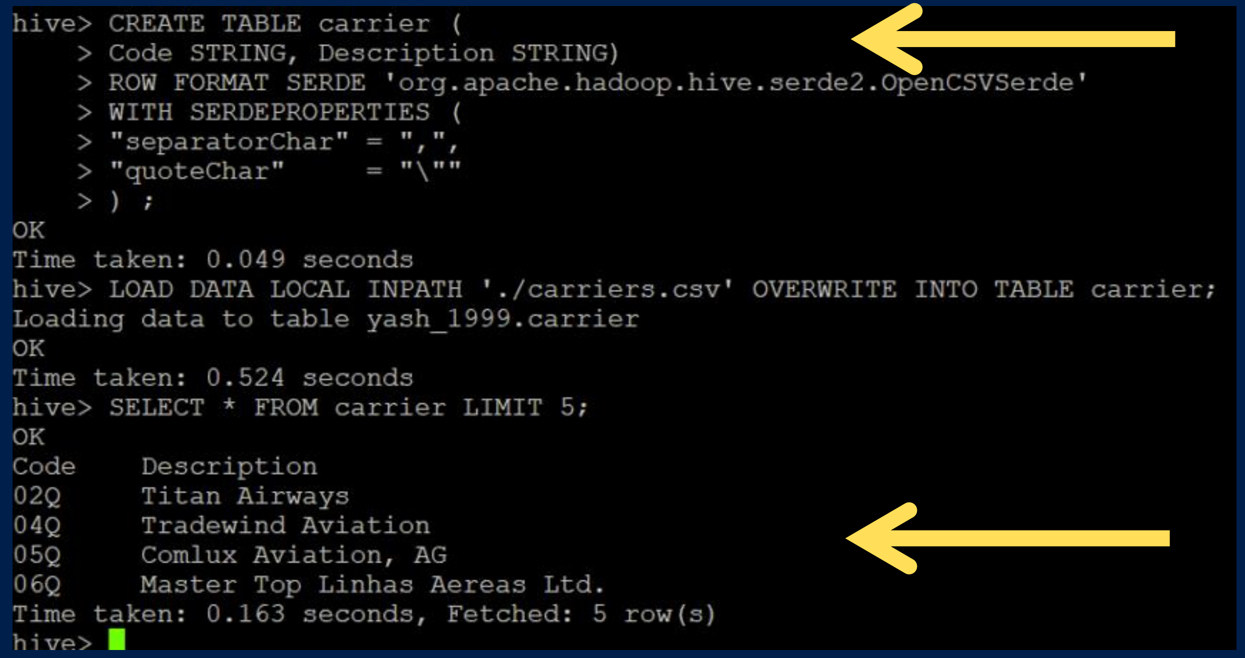
```
hive> CREATE DATABASE yash_1999;
OK
Time taken: 1.091 seconds
hive> USE yash_1999;
OK
Time taken: 0.064 seconds
hive>
```



```
hive> USE yash_1999;
OK
Time taken: 0.485 seconds
hive> CREATE TABLE airline_1999 (
    > Year INT,
    > Month INT,
    > DayofMonth INT,
    > DayOfWeek INT ,
    > DepTime INT,
    > CRSDepTime INT,
    > ArrTime INT,
    > CRSArrTime INT,
    > UniqueCarrier STRING,
    > FlightNum STRING,
    > TailNum STRING,
    > ActualElapsedTime INT,
    > CRSElapsedTime INT,
    > AirTime INT,
    > ArrDelay INT,
    > DepDelay INT,
    > Origin STRING,
    > Dest STRING,
    > Distance INT,
    > TaxiIn INT,
    > TaxiOut INT,
    > Cancelled INT,
    > CancellationCode STRING,
    > Diverted INT,
    > CarrierDelay INT,
    > WeatherDelay INT,
    > NASDelay INT,
    > SecurityDelay INT,
    > LateAircraftDelay INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.312 seconds
hive> LOAD DATA LOCAL INPATH './yash_1999.csv' OVERWRITE INTO TABLE airline_1
999;
Loading data to table yash_1999.airline_1999
OK
Time taken: 1.289 seconds
```

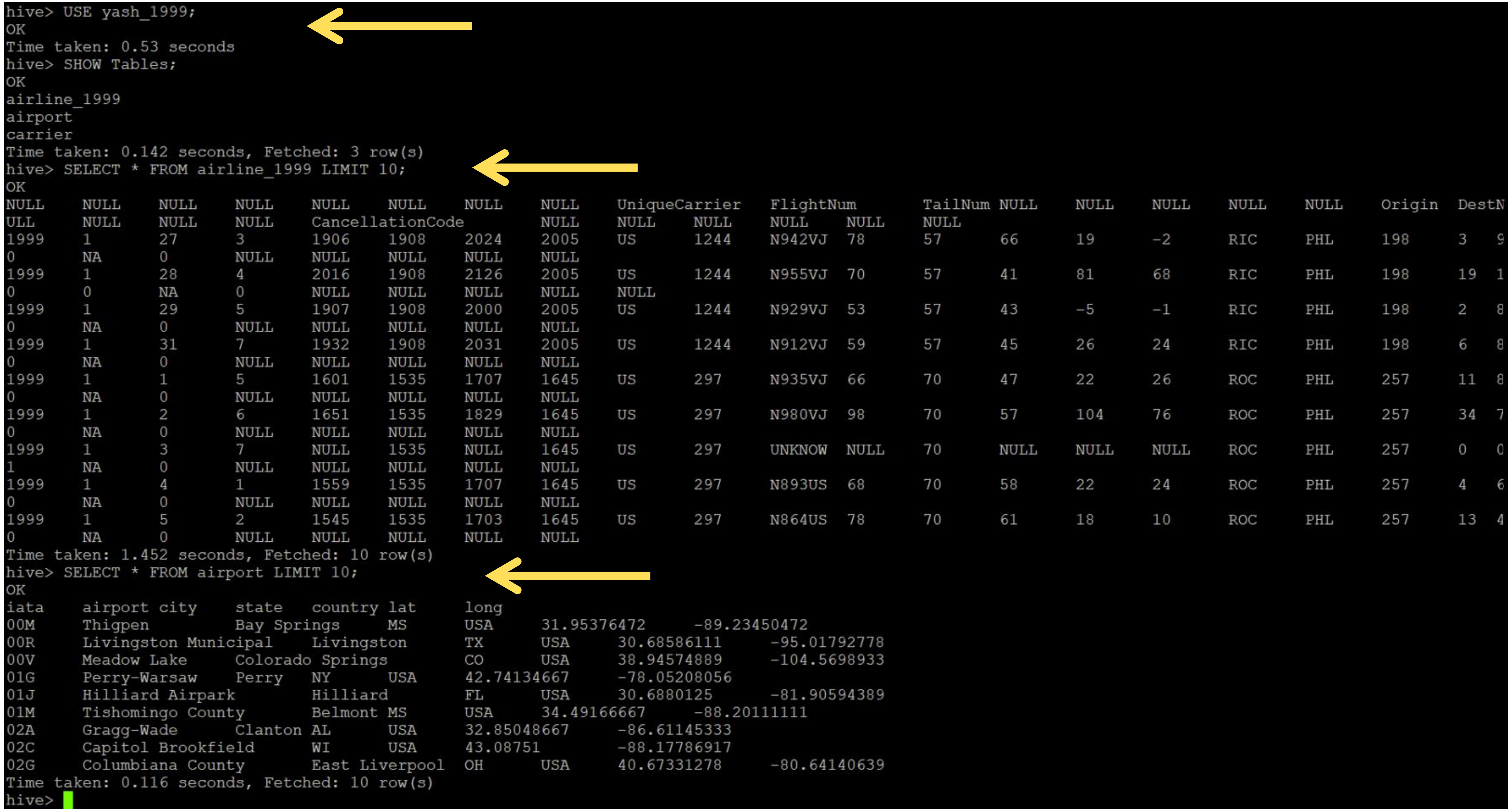


```
hive> CREATE TABLE airport(
    > iata STRING,
    > airport STRING,
    > city STRING,
    > state STRING,
    > country STRING,
    > lat STRING,
    > long STRING)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    > "separatorChar" = ",",
    > "quoteChar"     = "\""
    > );
OK
Time taken: 0.049 seconds
hive> LOAD DATA LOCAL INPATH './airports.csv' OVERWRITE INTO TABLE airport;
Loading data to table yash_1999.airport
OK
Time taken: 0.579 seconds
hive> SELECT COUNT(*) FROM airport;
Query ID = hadoop_20230404030732_e678b87f-86ab-4705-b393-73d7d6313cdb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680575892
200_0006)
```



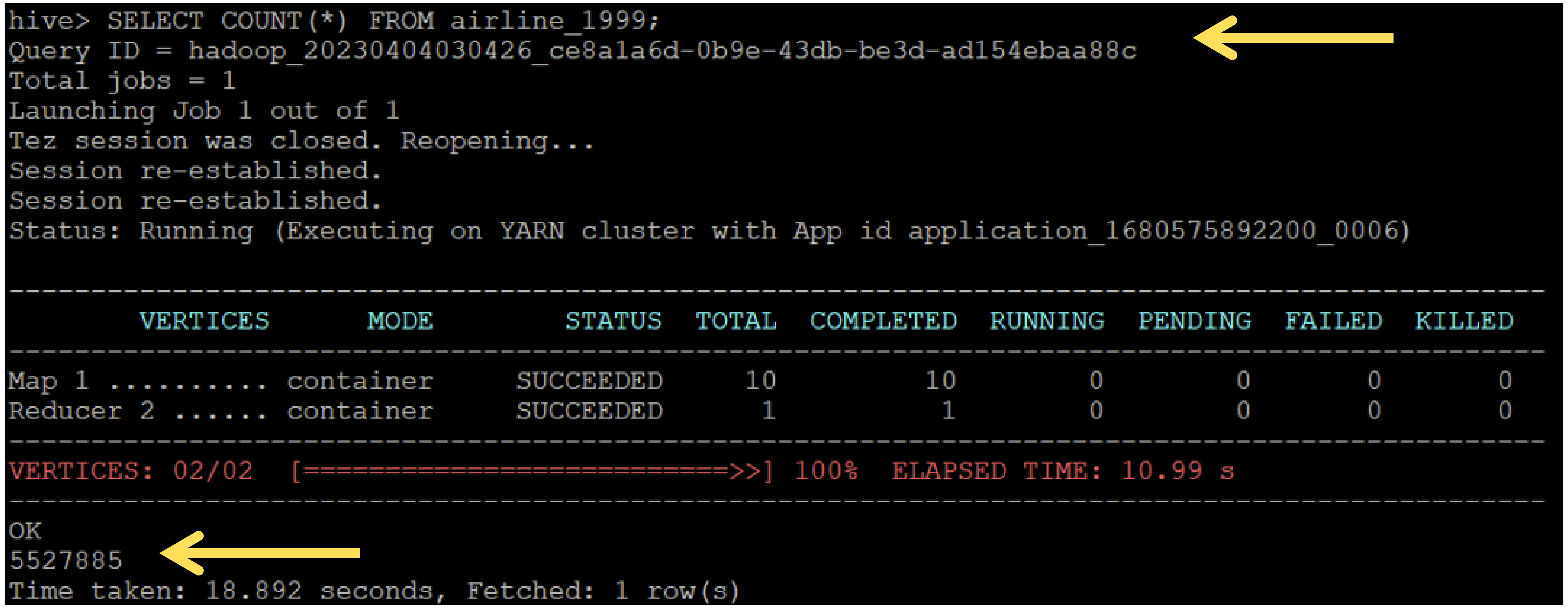
```
hive> CREATE TABLE carrier (
    > Code STRING, Description STRING)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    > "separatorChar" = ",",
    > "quoteChar"     = "\""
    > ) ;
OK
Time taken: 0.049 seconds
hive> LOAD DATA LOCAL INPATH './carriers.csv' OVERWRITE INTO TABLE carrier;
Loading data to table yash_1999.carrier
OK
Time taken: 0.524 seconds
hive> SELECT * FROM carrier LIMIT 5;
OK
Code    Description
02Q     Titan Airways
04Q     Tradewind Aviation
05Q     Comlux Aviation, AG
06Q     Master Top Linhas Aereas Ltd.
Time taken: 0.163 seconds, Fetched: 5 row(s)
hive>
```

Sample data of Tables



```
hive> USE yash_1999;
OK
Time taken: 0.53 seconds
hive> SHOW Tables;
OK
airline_1999
airport
carrier
Time taken: 0.142 seconds, Fetched: 3 row(s)
hive> SELECT * FROM airline_1999 LIMIT 10;
OK
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    UniqueCarrier   FlightNum       TailNum NULL    NULL    NULL    NULL    NULL    Origin  DestN
ULL     NULL    NULL    NULL    CancellationCode        NULL    NULL    NULL    NULL    NULL    NULL
1999    1       27      3       1906    1908    2024    2005    US      1244    N942VJ  78      57      66      19      -2      RIC     PHL     198     3   9
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       28      4       2016    1908    2126    2005    US      1244    N955VJ  70      57      41      81      68      RIC     PHL     198     19  1
0       0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       29      5       1907    1908    2000    2005    US      1244    N929VJ  53      57      43      -5      -1      RIC     PHL     198     2   8
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       31      7       1932    1908    2031    2005    US      1244    N912VJ  59      57      45      26      24      RIC     PHL     198     6   8
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       1       5       1601    1535    1707    1645    US      297     N935VJ  66      70      47      22      26      ROC     PHL     257     11  8
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       2       6       1651    1535    1829    1645    US      297     N980VJ  98      70      57      104     76      ROC     PHL     257     34  7
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       3       7       NULL    1535    NULL    1645    US      297     UNKNOW  NULL    70      NULL    NULL    NULL    ROC     PHL     257     0   0
1       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       4       1       1559    1535    1707    1645    US      297     N893US  68      70      58      22      24      ROC     PHL     257     4   6
0       NA      0       NULL    NULL    NULL    NULL    NULL
1999    1       5       2       1545    1535    1703    1645    US      297     N864US  78      70      61      18      10      ROC     PHL     257     13  4
0       NA      0       NULL    NULL    NULL    NULL    NULL
Time taken: 1.452 seconds, Fetched: 10 row(s)
hive> SELECT * FROM airport LIMIT 10;
OK
iata    airport city    state   country lat     long
00M     Thigpen         Bay Springs     MS      USA     31.95376472     -89.23450472
00R     Livingston Municipal    Livingston      TX      USA     30.68586111     -95.01792778
00V     Meadow Lake     Colorado Springs        CO      USA     38.94574889     -104.5698933
01G     Perry-Warsaw    Perry   NY      USA     42.74134667     -78.05208056
01J     Hilliard Airpark        Hilliard        FL      USA     30.6880125      -81.90594389
01M     Tishomingo County       Belmont MS      USA     34.49166667     -88.20111111
02A     Gragg-Wade      Clanton AL      USA     32.85048667     -86.61145333
02C     Capitol Brookfield      WI      USA     43.08751        -88.17786917
02G     Columbiana County       East Liverpool  OH      USA     40.67331278     -80.64140639
Time taken: 0.116 seconds, Fetched: 10 row(s)
hive>
```

Sample data of Tables

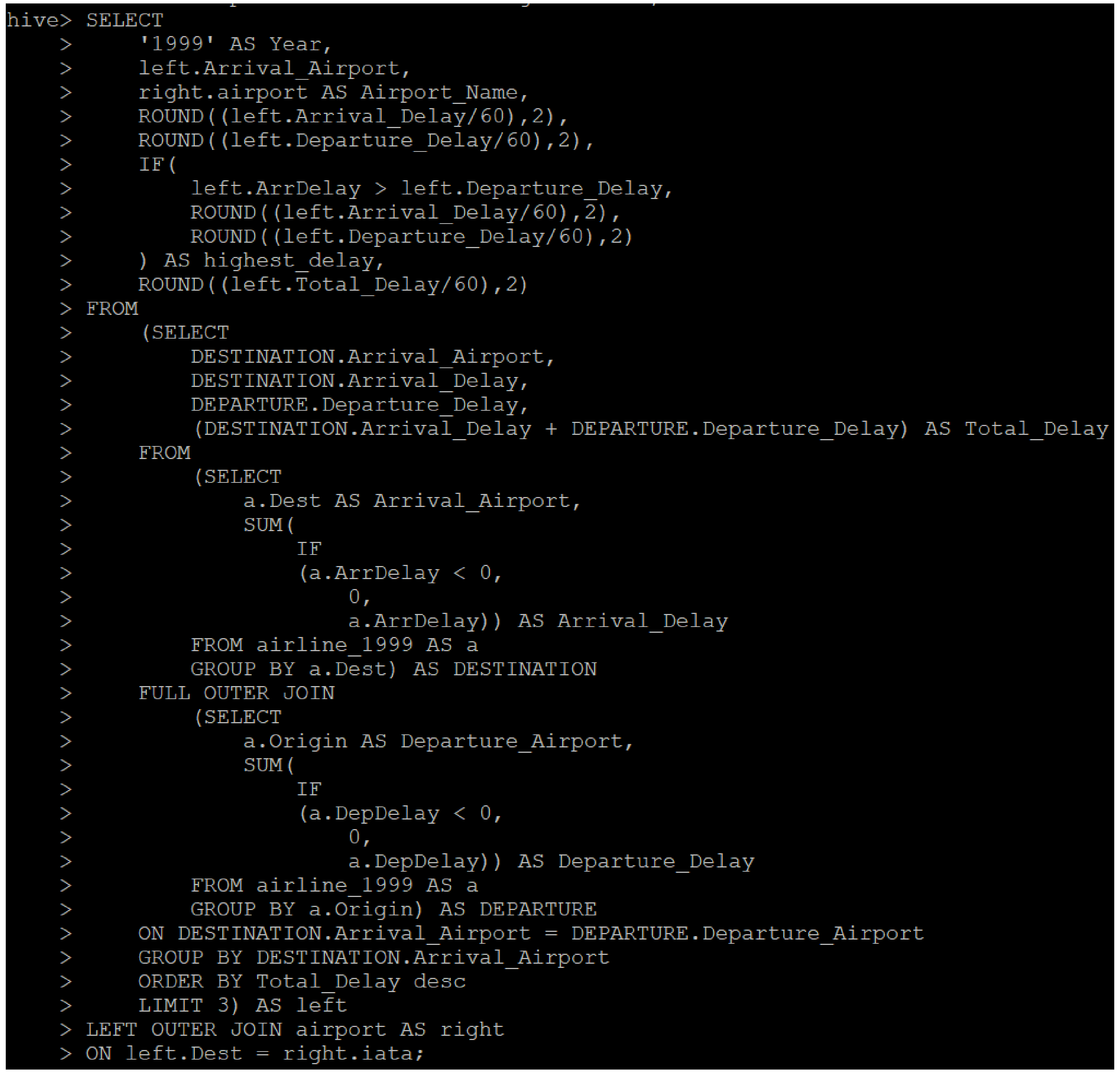


```
hive> SELECT COUNT(*) FROM airline_1999;
Query ID = hadoop_20230404030426_ce8a1a6d-0b9e-43db-be3d-ad154ebaa88c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1680575892200_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.99 s
----------------------------------------------------------------------------------------------
OK
5527885
Time taken: 18.892 seconds, Fetched: 1 row(s)
```

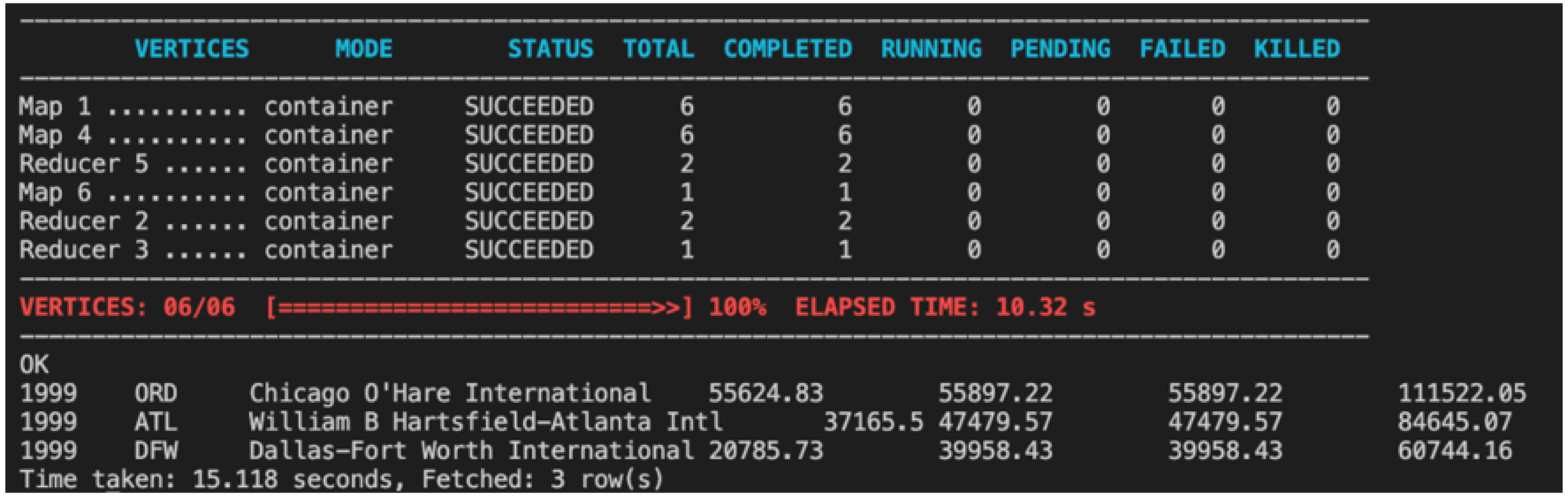
CASE 1

Determined the three **airports** with the highest delay time (in hours) for 1999

(Displayed in hours)



```
hive> SELECT
    >     '1999' AS Year,
    >     left.Arrival_Airport,
    >     right.airport AS Airport_Name,
    >     ROUND((left.Arrival_Delay/60),2),
    >     ROUND((left.Departure_Delay/60),2),
    >     IF(
    >         left.ArrDelay > left.Departure_Delay,
    >         ROUND((left.Arrival_Delay/60),2),
    >         ROUND((left.Departure_Delay/60),2)
    >     ) AS highest_delay,
    >     ROUND((left.Total_Delay/60),2)
    > FROM
    >     (SELECT
    >         DESTINATION.Arrival_Airport,
    >         DESTINATION.Arrival_Delay,
    >         DEPARTURE.Departure_Delay,
    >         (DESTINATION.Arrival_Delay + DEPARTURE.Departure_Delay) AS Total_Delay
    >     FROM
    >         (SELECT
    >             a.Dest AS Arrival_Airport,
    >             SUM(
    >                 IF
    >                 (a.ArrDelay < 0,
    >                     0,
    >                     a.ArrDelay)) AS Arrival_Delay
    >         FROM airline_1999 AS a
    >         GROUP BY a.Dest) AS DESTINATION
    >     FULL OUTER JOIN
    >         (SELECT
    >             a.Origin AS Departure_Airport,
    >             SUM(
    >                 IF
    >                 (a.DepDelay < 0,
    >                     0,
    >                     a.DepDelay)) AS Departure_Delay
    >         FROM airline_1999 AS a
    >         GROUP BY a.Origin) AS DEPARTURE
    >     ON DESTINATION.Arrival_Airport = DEPARTURE.Departure_Airport
    >     GROUP BY DESTINATION.Arrival_Airport
    >     ORDER BY Total_Delay desc
    >     LIMIT 3) AS left
    > LEFT OUTER JOIN airport AS right
    > ON left.Dest = right.iata;
```



```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 4 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      2          2        0        0       0       0
Map 6 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 10.32 s
----------------------------------------------------------------------------------------------
OK
1999    ORD     Chicago O'Hare International    55624.83        55897.22        55897.22        111522.05
1999    ATL     William B Hartsfield-Atlanta Intl       37165.5 47479.57        47479.57        84645.07
1999    DFW     Dallas-Fort Worth International 20785.73        39958.43        39958.43        60744.16
Time taken: 15.118 seconds, Fetched: 3 row(s)
```

Yash Jivani

CASE 1: RESULT

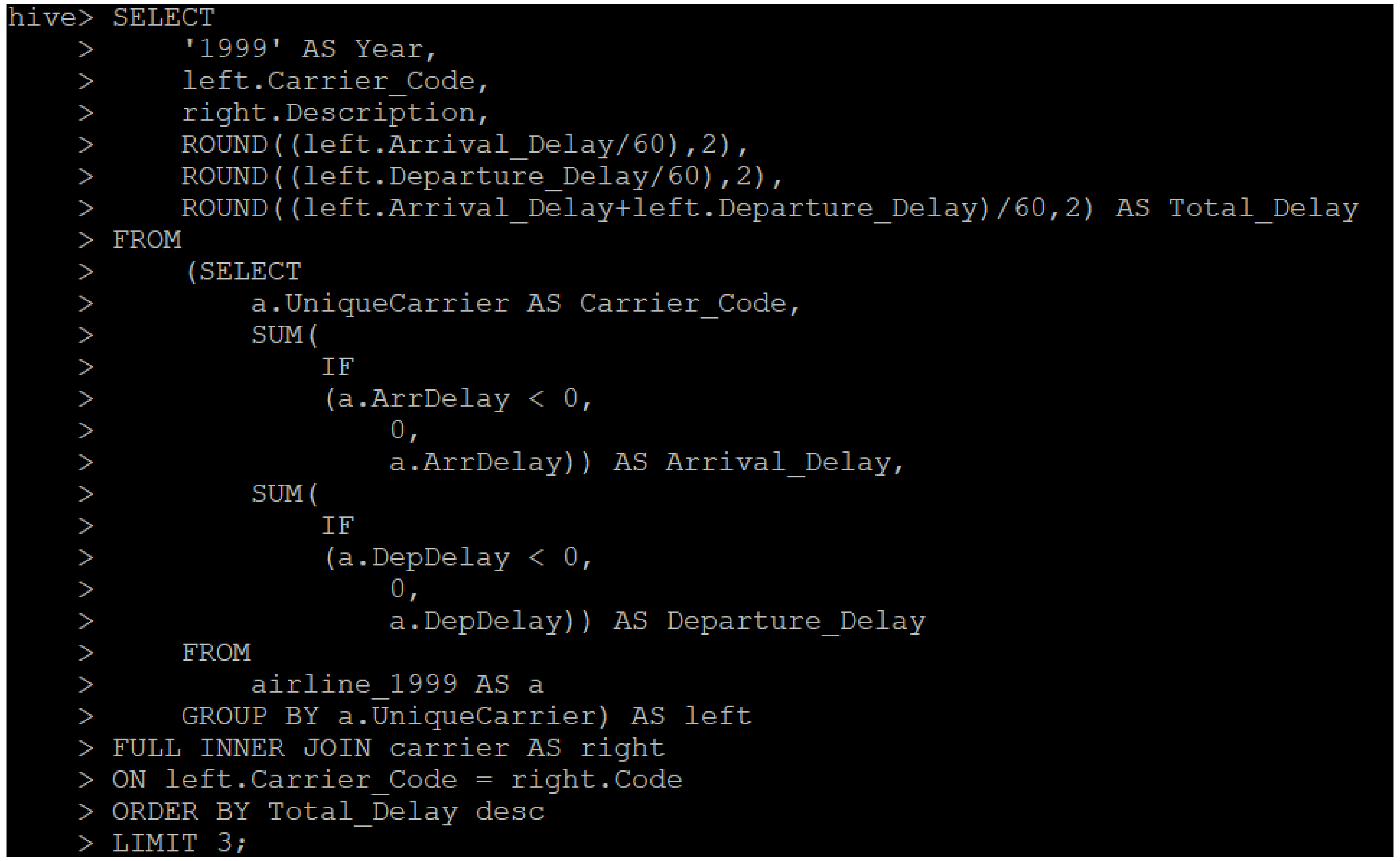
The three **airports** with the highest delay time (in hours)

Year	Airports	Arrival Time Delay (In Hours)	Departure time Delay (In Hours)	Total Delay (In Hours)
1999	Chicago O'Hare International	55624.83	**55897.22**	111522.05
1999	William B Hartsfield-Atlanta Intl	37165.5	**47479.57**	84645.07
1999	Dallas-Fort Worth International	20785.73	**39958.43**	60744.16

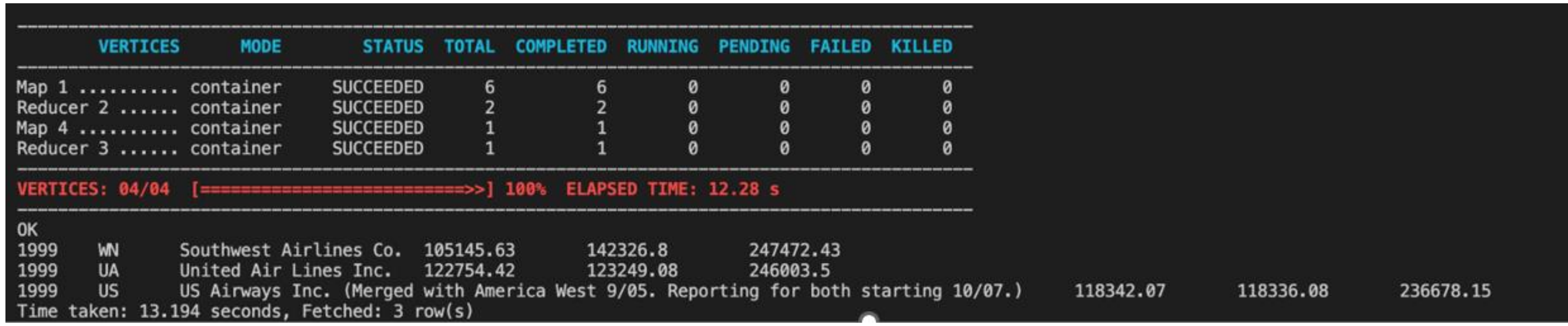
CASE 2

Determined the three **carriers** with the highest delay time (in hours) for 1999

(Displayed in hours)



```
hive> SELECT
    >      '1999' AS Year,
    >      left.Carrier_Code,
    >      right.Description,
    >      ROUND((left.Arrival_Delay/60),2),
    >      ROUND((left.Departure_Delay/60),2),
    >      ROUND((left.Arrival_Delay+left.Departure_Delay)/60,2) AS Total_Delay
    > FROM
    >      (SELECT
    >          a.UniqueCarrier AS Carrier_Code,
    >          SUM(
    >              IF
    >              (a.ArrDelay < 0,
    >                  0,
    >                  a.ArrDelay)) AS Arrival_Delay,
    >          SUM(
    >              IF
    >              (a.DepDelay < 0,
    >                  0,
    >                  a.DepDelay)) AS Departure_Delay
    >      FROM
    >          airline_1999 AS a
    >      GROUP BY a.UniqueCarrier) AS left
    > FULL INNER JOIN carrier AS right
    > ON left.Carrier_Code = right.Code
    > ORDER BY Total_Delay desc
    > LIMIT 3;
```



```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 12.28 s
----------------------------------------------------------------------------------------------
OK
1999    WN      Southwest Airlines Co.  105145.63       142326.8        247472.43
1999    UA      United Air Lines Inc.   122754.42       123249.08       246003.5
1999    US      US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)     118342.07       118336.08       236678.15
Time taken: 13.194 seconds, Fetched: 3 row(s)
```

Yash Jivani

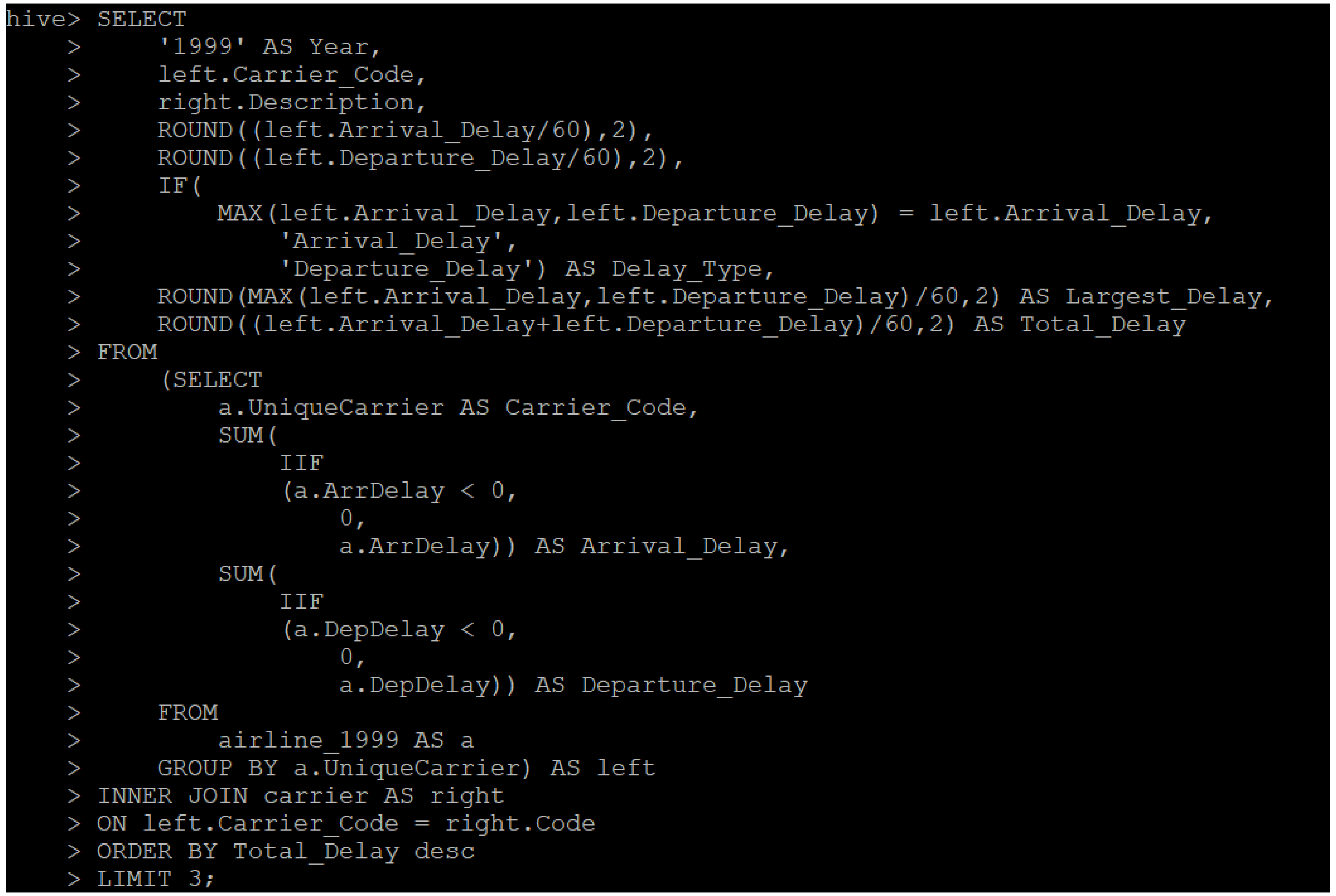
CASE 2: RESULT

The three **carriers** with the highest delay time (in hours)

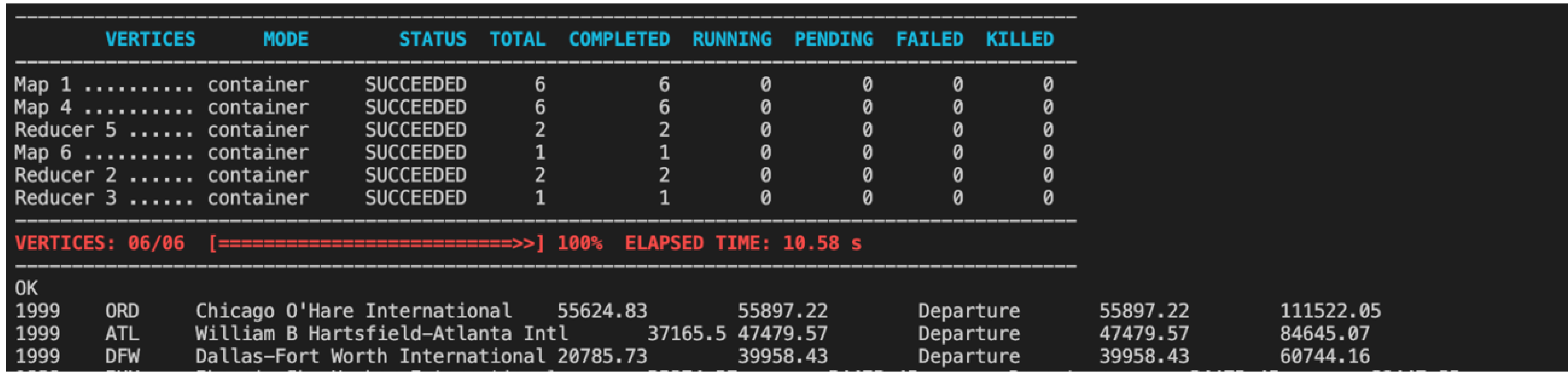
Year	Carriers	Arrival Time Delay (In Hours)	Departure time Delay (In Hours)	Total Delay (In Hours)
1999	Southwest Airlines Co.	105145.63	**142326.80**	247472.43
1999	United Air Lines Inc.	122754.42	**123249.08**	246003.5
1999	US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)	**118342.07**	118336.08	236678.15

CASE 3

(Total Delays(in Hours) for year **1999**)



```
hive> SELECT
    >     '1999' AS Year,
    >     left.Carrier_Code,
    >     right.Description,
    >     ROUND((left.Arrival_Delay/60),2),
    >     ROUND((left.Departure_Delay/60),2),
    >     IF(
    >         MAX(left.Arrival_Delay,left.Departure_Delay) = left.Arrival_Delay,
    >             'Arrival_Delay',
    >             'Departure_Delay') AS Delay_Type,
    >     ROUND(MAX(left.Arrival_Delay,left.Departure_Delay)/60,2) AS Largest_Delay,
    >     ROUND((left.Arrival_Delay+left.Departure_Delay)/60,2) AS Total_Delay
    > FROM
    >     (SELECT
    >         a.UniqueCarrier AS Carrier_Code,
    >         SUM(
    >             IIF
    >             (a.ArrDelay < 0,
    >                 0,
    >                 a.ArrDelay)) AS Arrival_Delay,
    >         SUM(
    >             IIF
    >             (a.DepDelay < 0,
    >                 0,
    >                 a.DepDelay)) AS Departure_Delay
    >     FROM
    >         airline_1999 AS a
    >     GROUP BY a.UniqueCarrier) AS left
    > INNER JOIN carrier AS right
    > ON left.Carrier_Code = right.Code
    > ORDER BY Total_Delay desc
    > LIMIT 3;
```



```
----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Map 4 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      2          2        0        0       0       0
Map 6 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 06/06  [==========================>>] 100%  ELAPSED TIME: 10.58 s
----------------------------------------------------------------------------------------------
OK
1999    ORD     Chicago O'Hare International    55624.83        55897.22        Departure       55897.22        111522.05
1999    ATL     William B Hartsfield-Atlanta Intl       37165.5 47479.57        Departure       47479.57        84645.07
1999    DFW     Dallas-Fort Worth International 20785.73        39958.43        Departure       39958.43        60744.16
```

Yash Jivani

CASE 3

Determined overall which type of delay (**arrivals** or **departures**) is the largest for airports

(Overall Delays for year **1999**)
For each airports, got the largest delay type

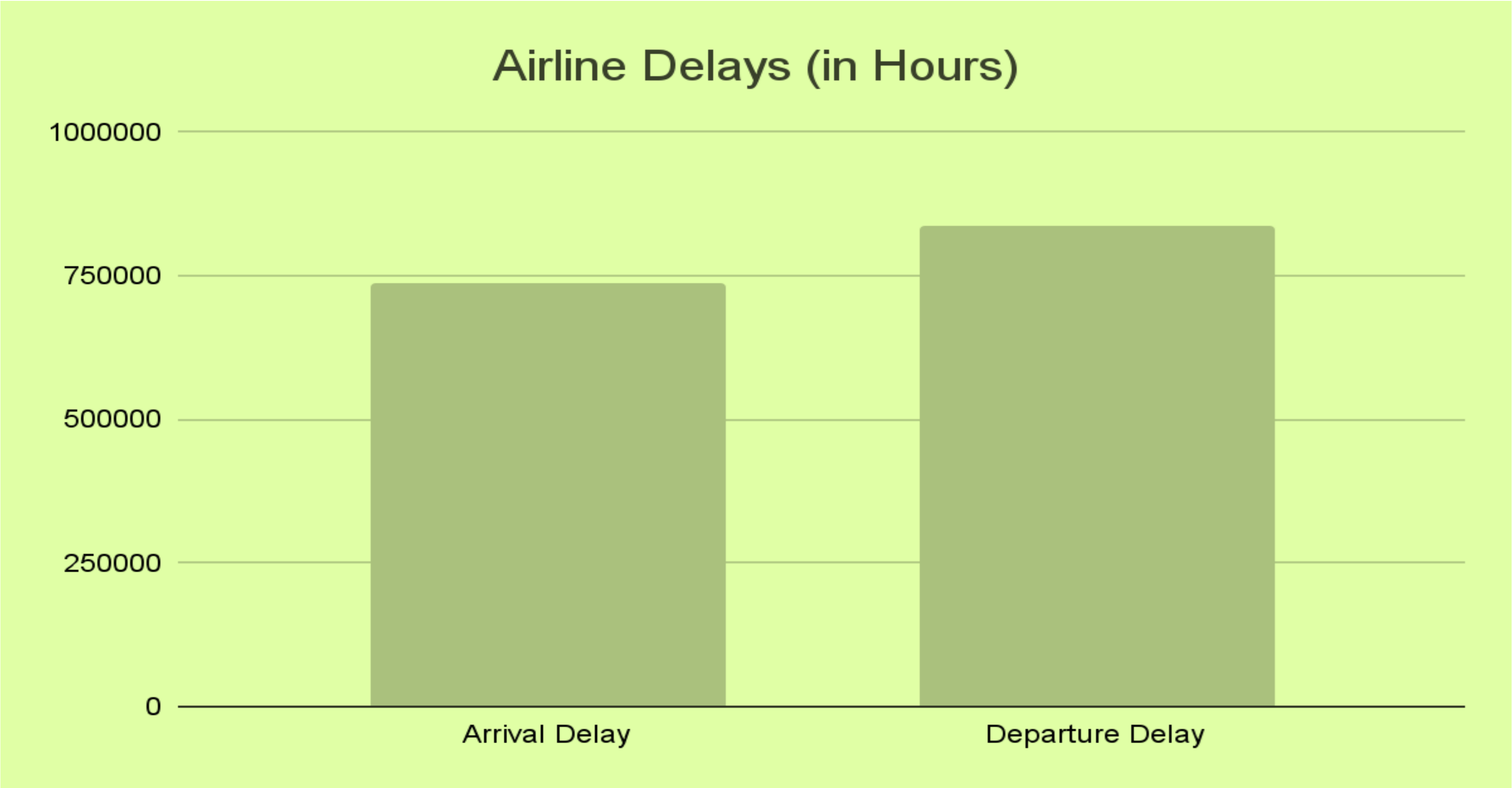
Year	Airports	Arrival Delay (in Hours)	Departure Delay (In Hours)	Largest Delay Type	Total Delay (in hours)
1999	Chicago O'Hare International	55624.83	**55897.22**	Departure	111522.05
1999	William B Hartsfield-Atlanta Intl	37165.5	**47479.57**	Departure	84645.07
1999	Dallas-Fort Worth International	20785.73	**39958.43**	Departure	60744.16

Analysis of Data 1999

Total Arrival and Departure delay in Hours

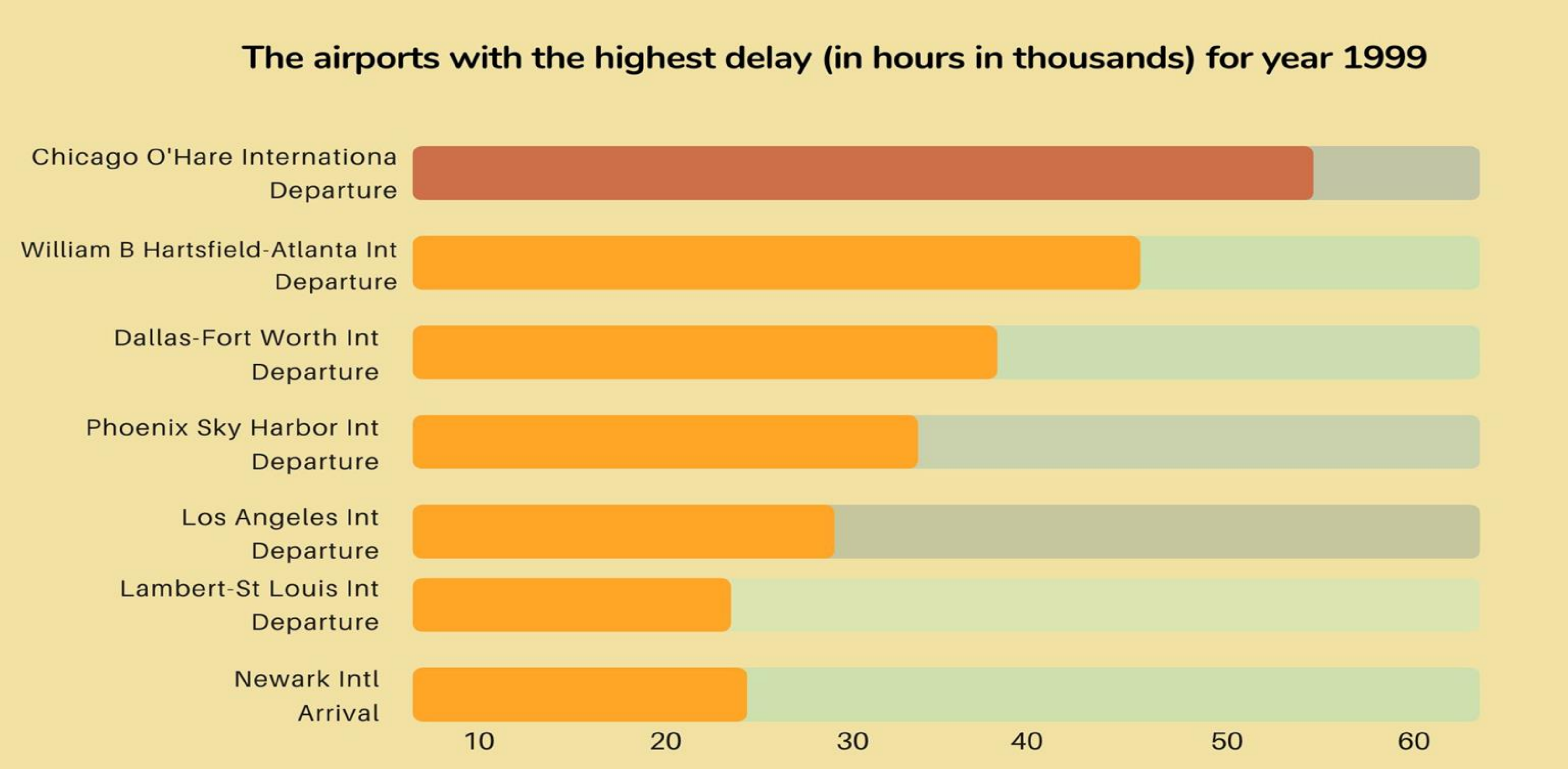
Year	Total Arrival delay	Total Departure delay
1999	736698.83	836226.55

Overall Departure Delay is Greater Than Arrival Delay



Yash Jivani

Analysis of Data 1999


The airports with the highest delay (in hours in thousands) for year 1999
Chicago O'Hare Internationa Departure
William B Hartsfield-Atlanta Int Departure
Dallas-Fort Worth Int Departure
Phoenix Sky Harbor Int Departure
Los Angeles Int Departure
Lambert-St Louis Int Departure
Newark Intl Arrival
10
20
30
40
50
60

Yash Jivani

AIRPORT DELAY

2000

Parth Dodia

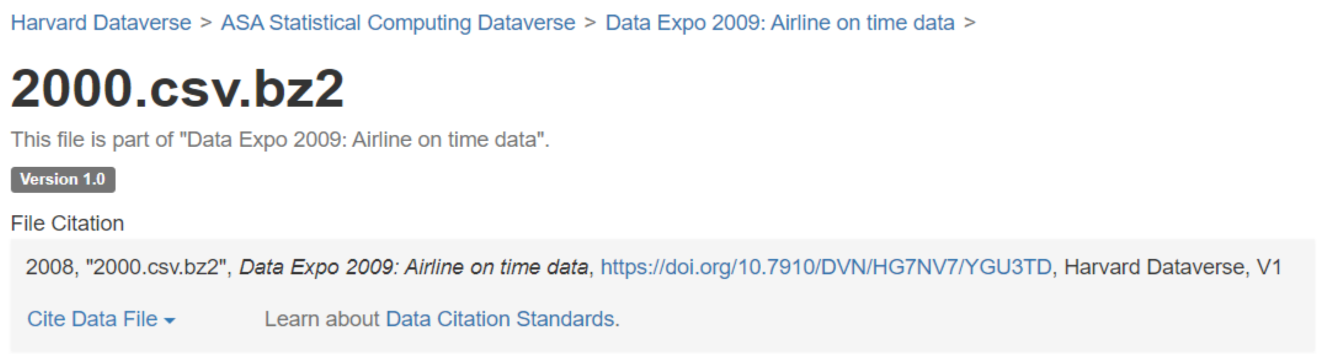
DATA SOURCE

Data Expo 2009: Airline on Time Data

The data represents flight arrival and departure details for all commercial flights within the USA for year 2000

Download Link

Data Expo 2009: Harward - Airline on Time Data



Harvard Dataverse > ASA Statistical Computing Dataverse > Data Expo 2009: Airline on time data >

2000.csv.bz2

This file is part of "Data Expo 2009: Airline on time data".

Version 1.0

File Citation

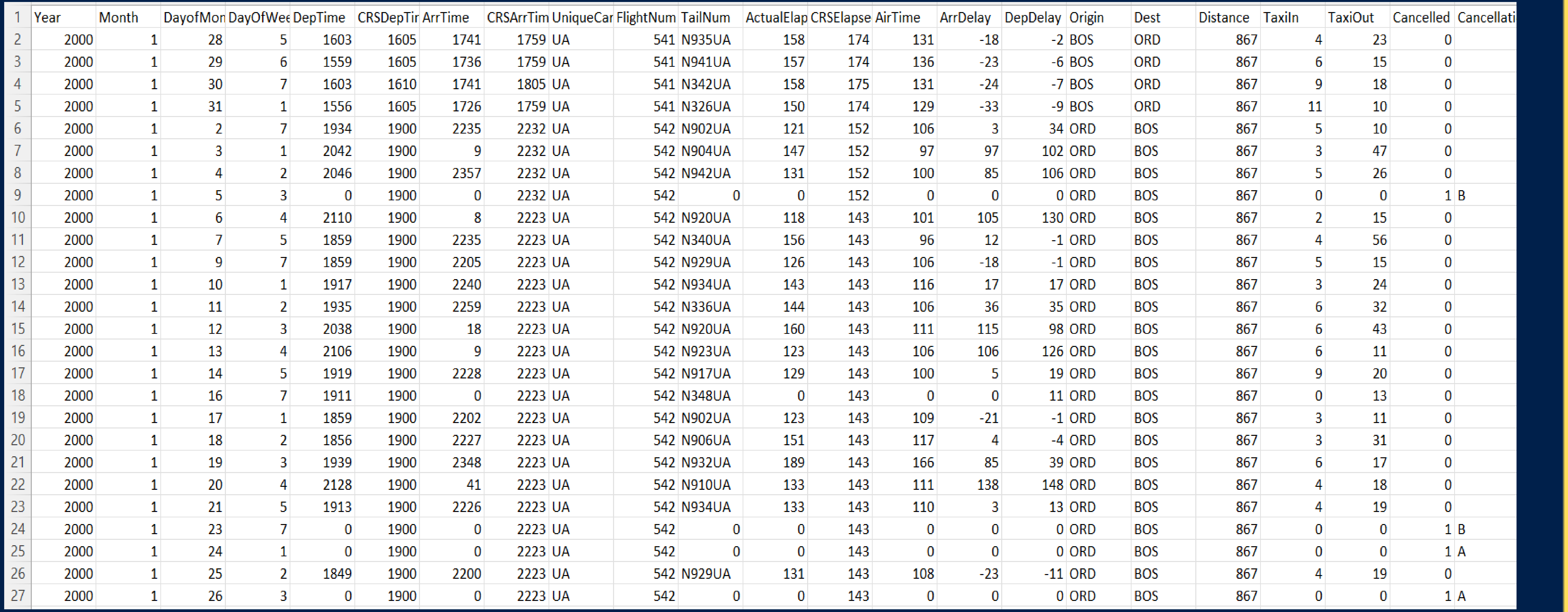
2008, "2000.csv.bz2", *Data Expo 2009: Airline on time data*, https://doi.org/10.7910/DVN/HG7NV7/YGU3TD, Harvard Dataverse, V1

Learn about Data Citation Standards.

Steps for working with Data

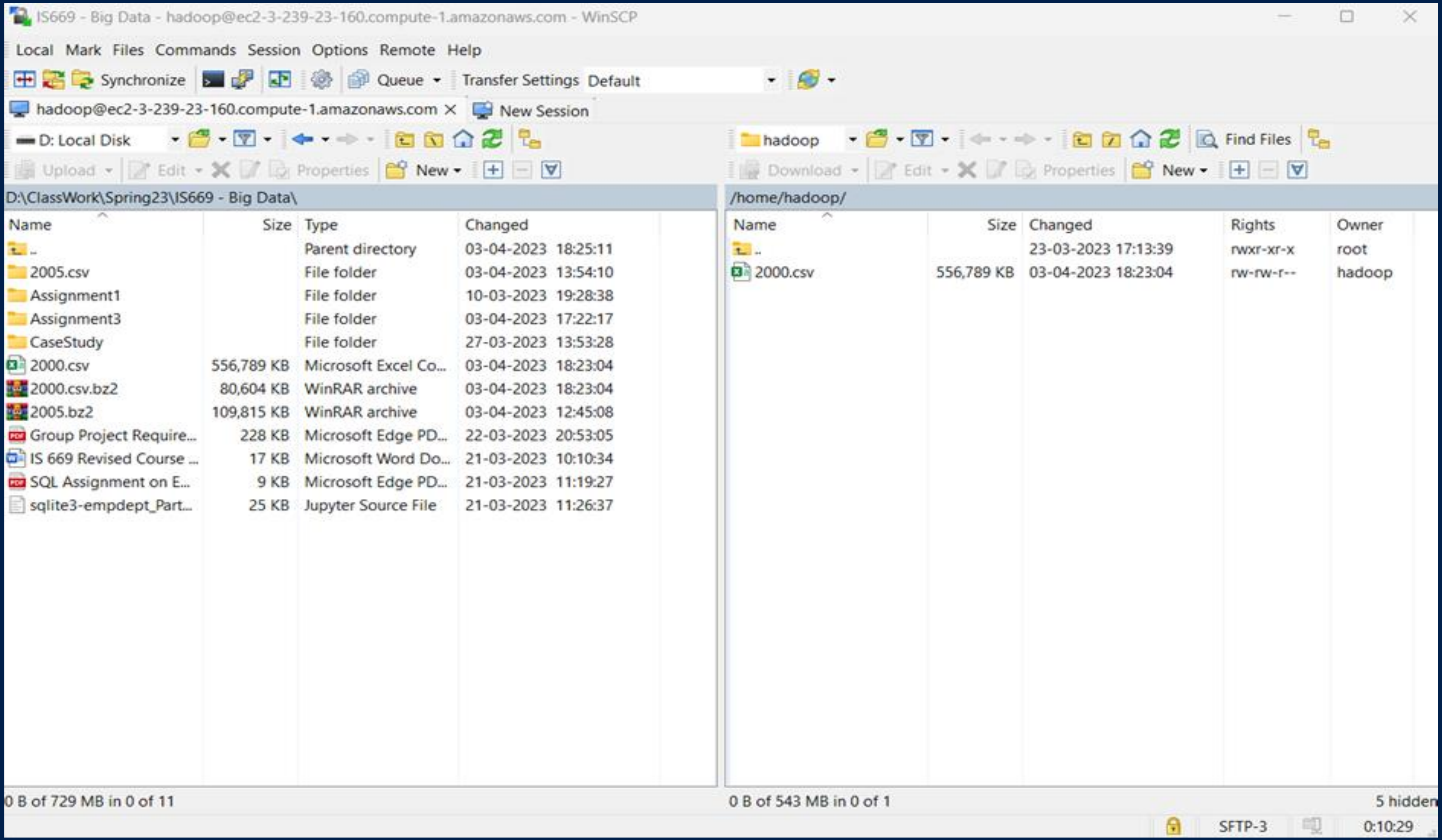
1. Install WinSCP to transfer local files to the server.
2. Start the EMR cluster.
3. Copy the hostname and connect to the server in WinSCP, proving the private key.
4. Transfer the required files to the server.
5. Connect to the cluster using putty.
6. Use 'ls' to see the directory of files.
7. Display the head of the file to show the content
8. 'hive' to go into hive.
9. Create a table 'parth_flight' for the data.
10. Load the data in the table using :
    'LOAD DATA LOCAL INPATH '2000.csv' OVERWRITE INTO TABLE parth_flight;
11. Start using SQL queries to get the required output.

Excel Spreadsheet



1	Year	Month	DayofMon	DayOfWee	DepTime	CRSDepTir	ArrTime	CRSArrTim	UniqueCar	FlightNum	TailNum	ActualElap	CRSElapse	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellati
2	2000	1	28	5	1603	1605	1741	1759	UA	541	N935UA	158	174	131	-18	-2	BOS	ORD	867	4	23	0	
3	2000	1	29	6	1559	1605	1736	1759	UA	541	N941UA	157	174	136	-23	-6	BOS	ORD	867	6	15	0	
4	2000	1	30	7	1603	1610	1741	1805	UA	541	N342UA	158	175	131	-24	-7	BOS	ORD	867	9	18	0	
5	2000	1	31	1	1556	1605	1726	1759	UA	541	N326UA	150	174	129	-33	-9	BOS	ORD	867	11	10	0	
6	2000	1	2	7	1934	1900	2235	2232	UA	542	N902UA	121	152	106	3	34	ORD	BOS	867	5	10	0	
7	2000	1	3	1	2042	1900	9	2232	UA	542	N904UA	147	152	97	97	102	ORD	BOS	867	3	47	0	
8	2000	1	4	2	2046	1900	2357	2232	UA	542	N942UA	131	152	100	85	106	ORD	BOS	867	5	26	0	
9	2000	1	5	3	0	1900	0	2232	UA	542	0	0	152	0	0	0	ORD	BOS	867	0	0	1	B
10	2000	1	6	4	2110	1900	8	2223	UA	542	N920UA	118	143	101	105	130	ORD	BOS	867	2	15	0	
11	2000	1	7	5	1859	1900	2235	2223	UA	542	N340UA	156	143	96	12	-1	ORD	BOS	867	4	56	0	
12	2000	1	9	7	1859	1900	2205	2223	UA	542	N929UA	126	143	106	-18	-1	ORD	BOS	867	5	15	0	
13	2000	1	10	1	1917	1900	2240	2223	UA	542	N934UA	143	143	116	17	17	ORD	BOS	867	3	24	0	
14	2000	1	11	2	1935	1900	2259	2223	UA	542	N336UA	144	143	106	36	35	ORD	BOS	867	6	32	0	
15	2000	1	12	3	2038	1900	18	2223	UA	542	N920UA	160	143	111	115	98	ORD	BOS	867	6	43	0	
16	2000	1	13	4	2106	1900	9	2223	UA	542	N923UA	123	143	106	106	126	ORD	BOS	867	6	11	0	
17	2000	1	14	5	1919	1900	2228	2223	UA	542	N917UA	129	143	100	5	19	ORD	BOS	867	9	20	0	
18	2000	1	16	7	1911	1900	0	2223	UA	542	N348UA	0	143	0	0	11	ORD	BOS	867	0	13	0	
19	2000	1	17	1	1859	1900	2202	2223	UA	542	N902UA	123	143	109	-21	-1	ORD	BOS	867	3	11	0	
20	2000	1	18	2	1856	1900	2227	2223	UA	542	N906UA	151	143	117	4	-4	ORD	BOS	867	3	31	0	
21	2000	1	19	3	1939	1900	2348	2223	UA	542	N932UA	189	143	166	85	39	ORD	BOS	867	6	17	0	
22	2000	1	20	4	2128	1900	41	2223	UA	542	N910UA	133	143	111	138	148	ORD	BOS	867	4	18	0	
23	2000	1	21	5	1913	1900	2226	2223	UA	542	N934UA	133	143	110	3	13	ORD	BOS	867	4	19	0	
24	2000	1	23	7	0	1900	0	2223	UA	542	0	0	143	0	0	0	ORD	BOS	867	0	0	1	B
25	2000	1	24	1	0	1900	0	2223	UA	542	0	0	143	0	0	0	ORD	BOS	867	0	0	1	A
26	2000	1	25	2	1849	1900	2200	2223	UA	542	N929UA	131	143	108	-23	-11	ORD	BOS	867	4	19	0	
27	2000	1	26	3	0	1900	0	2223	UA	542	0	0	143	0	0	0	ORD	BOS	867	0	0	1	A

Transferring files to server



IS669 - Big Data - hadoop@ec2-3-239-23-160.compute-1.amazonaws.com - WinSCP

Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
hadoop@ec2-3-239-23-160.compute-1.amazonaws.com
New Session
D: Local Disk
Upload
Edit
Properties
New
hadoop
Find Files
Download
Edit
Properties
New

D:\ClassWork\Spring23\IS669 - Big Data\

Name	Size	Type	Changed
..		Parent directory	03-04-2023 18:25:11
2005.csv		File folder	03-04-2023 13:54:10
Assignment1		File folder	10-03-2023 19:28:38
Assignment3		File folder	03-04-2023 17:22:17
CaseStudy		File folder	27-03-2023 13:53:28
2000.csv	556,789 KB	Microsoft Excel Co...	03-04-2023 18:23:04
2000.csv.bz2	80,604 KB	WinRAR archive	03-04-2023 18:23:04
2005.bz2	109,815 KB	WinRAR archive	03-04-2023 12:45:08
Group Project Require...	228 KB	Microsoft Edge PD...	22-03-2023 20:53:05
IS 669 Revised Course ...	17 KB	Microsoft Word Do...	21-03-2023 10:10:34
SQL Assignment on E...	9 KB	Microsoft Edge PD...	21-03-2023 11:19:27
sqlite3-empdept_Part...	25 KB	Jupyter Source File	21-03-2023 11:26:37

0 B of 729 MB in 0 of 11

/home/hadoop/

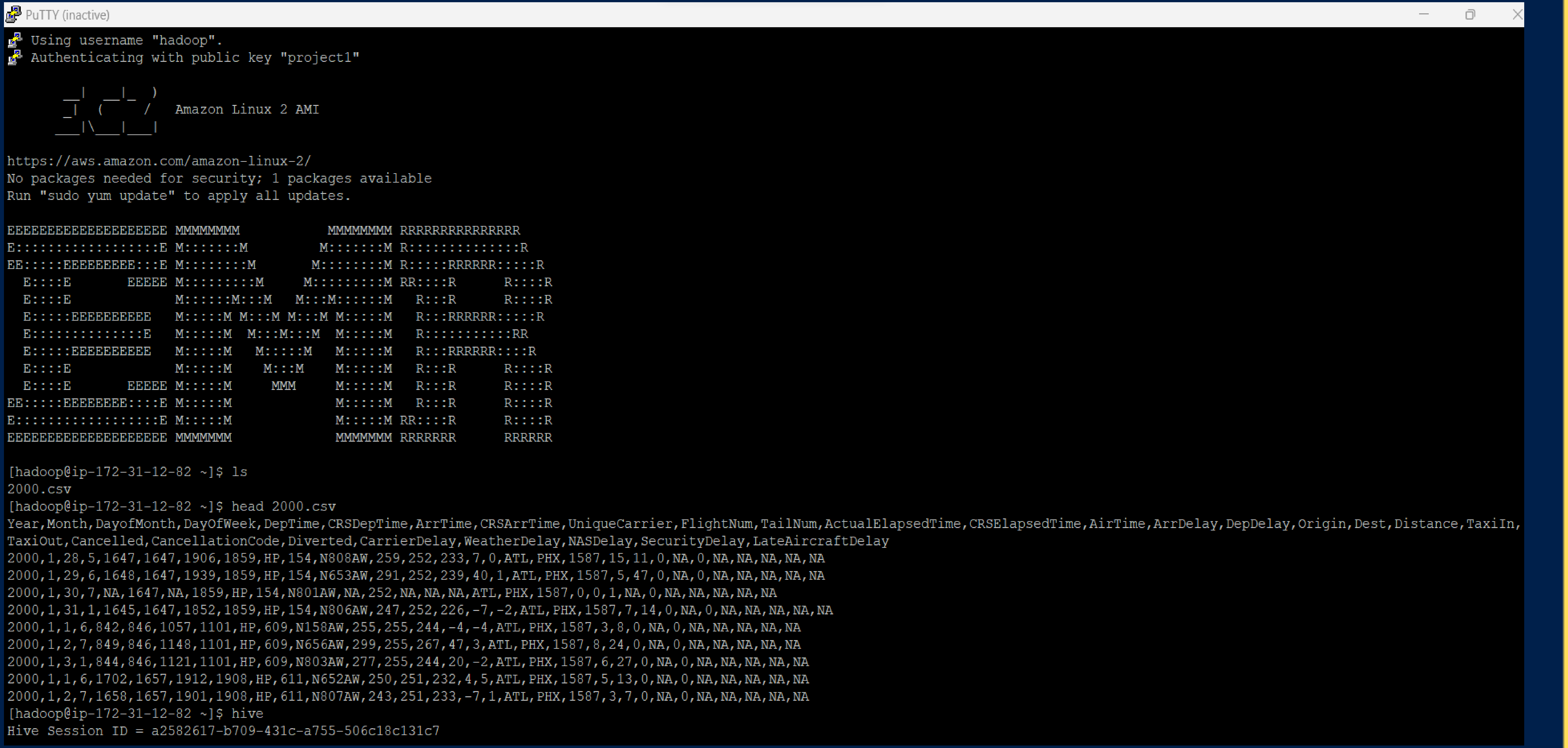
Name	Size	Changed	Rights	Owner
..		23-03-2023 17:13:39	rwxr-xr-x	root
2000.csv	556,789 KB	03-04-2023 18:23:04	rw-rw-r--	hadoop

0 B of 543 MB in 0 of 1

5 hidden

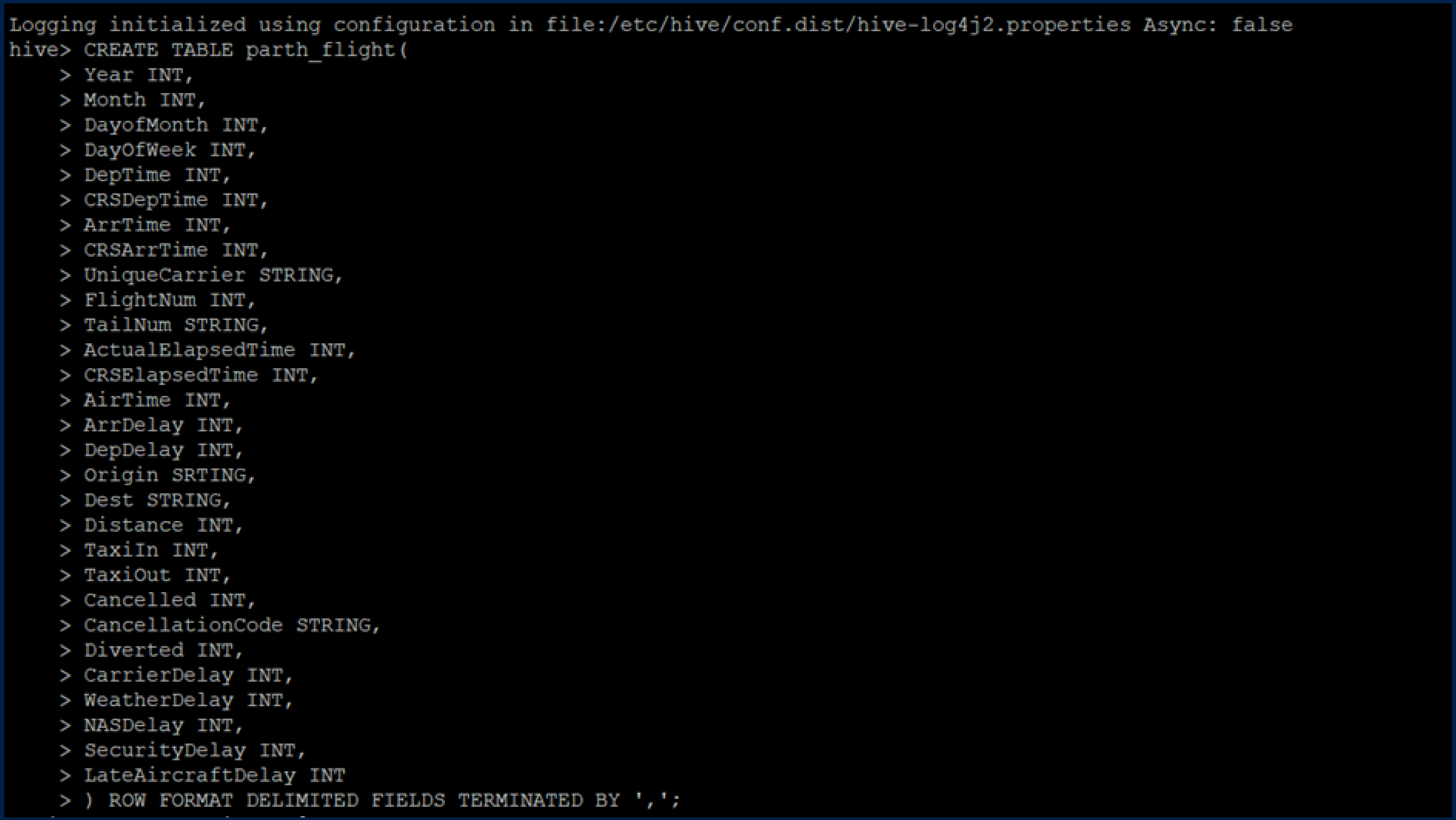
SFTP-3 0:10:29

Steps for working with Data

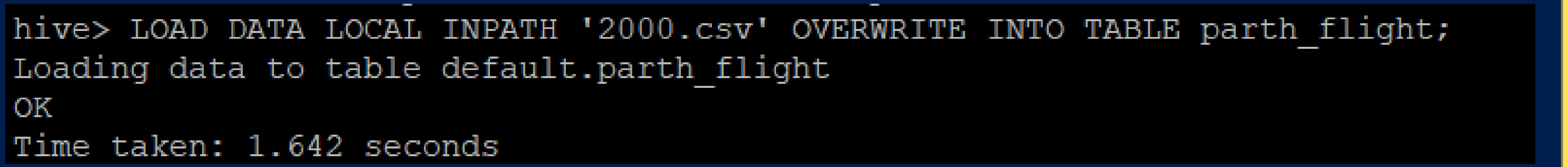


```
Using username "hadoop".
Authenticating with public key "project1"

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
No packages needed for security; 1 packages available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-12-82 ~]$ ls
2000.csv
[hadoop@ip-172-31-12-82 ~]$ head 2000.csv
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,Origin,Dest,Distance,TaxiIn,
TaxiOut,Cancelled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
2000,1,28,5,1647,1647,1906,1859,HP,154,N808AW,259,252,233,7,0,ATL,PHX,1587,15,11,0,NA,0,NA,NA,NA,NA,NA
2000,1,29,6,1648,1647,1939,1859,HP,154,N653AW,291,252,239,40,1,ATL,PHX,1587,5,47,0,NA,0,NA,NA,NA,NA,NA
2000,1,30,7,NA,1647,NA,1859,HP,154,N801AW,NA,252,NA,NA,NA,ATL,PHX,1587,0,0,1,NA,0,NA,NA,NA,NA,NA
2000,1,31,1,1645,1647,1852,1859,HP,154,N806AW,247,252,226,-7,-2,ATL,PHX,1587,7,14,0,NA,0,NA,NA,NA,NA,NA
2000,1,1,6,842,846,1057,1101,HP,609,N158AW,255,255,244,-4,-4,ATL,PHX,1587,3,8,0,NA,0,NA,NA,NA,NA,NA
2000,1,2,7,849,846,1148,1101,HP,609,N656AW,299,255,267,47,3,ATL,PHX,1587,8,24,0,NA,0,NA,NA,NA,NA,NA
2000,1,3,1,844,846,1121,1101,HP,609,N803AW,277,255,244,20,-2,ATL,PHX,1587,6,27,0,NA,0,NA,NA,NA,NA,NA
2000,1,1,6,1702,1657,1912,1908,HP,611,N652AW,250,251,232,4,5,ATL,PHX,1587,5,13,0,NA,0,NA,NA,NA,NA,NA
2000,1,2,7,1658,1657,1901,1908,HP,611,N807AW,243,251,233,-7,1,ATL,PHX,1587,3,7,0,NA,0,NA,NA,NA,NA,NA
[hadoop@ip-172-31-12-82 ~]$ hive
Hive Session ID = a2582617-b709-431c-a755-506c18c131c7
```

Creating Table and loading data

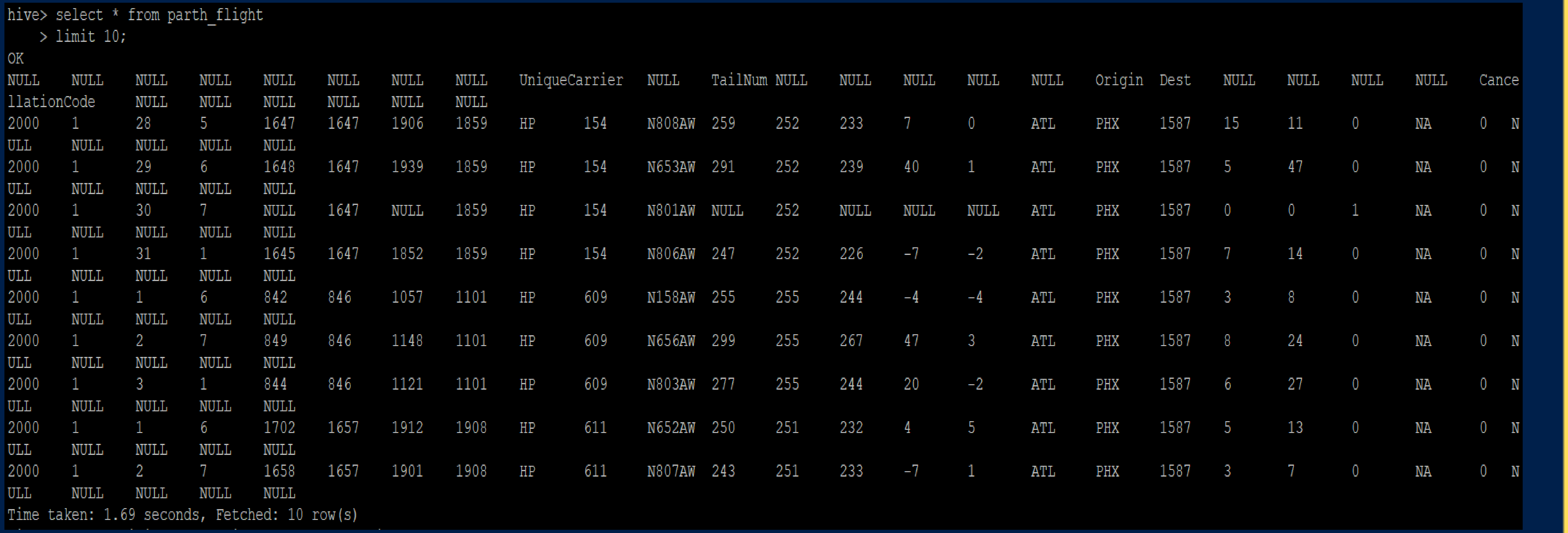


```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> CREATE TABLE parth_flight(
    > Year INT,
    > Month INT,
    > DayofMonth INT,
    > DayOfWeek INT,
    > DepTime INT,
    > CRSDepTime INT,
    > ArrTime INT,
    > CRSArrTime INT,
    > UniqueCarrier STRING,
    > FlightNum INT,
    > TailNum STRING,
    > ActualElapsedTime INT,
    > CRSElapsedTime INT,
    > AirTime INT,
    > ArrDelay INT,
    > DepDelay INT,
    > Origin SRTING,
    > Dest STRING,
    > Distance INT,
    > TaxiIn INT,
    > TaxiOut INT,
    > Cancelled INT,
    > CancellationCode STRING,
    > Diverted INT,
    > CarrierDelay INT,
    > WeatherDelay INT,
    > NASDelay INT,
    > SecurityDelay INT,
    > LateAircraftDelay INT
    > ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```



```
hive> LOAD DATA LOCAL INPATH '2000.csv' OVERWRITE INTO TABLE parth_flight;
Loading data to table default.parth_flight
OK
Time taken: 1.642 seconds
```

Display of the data

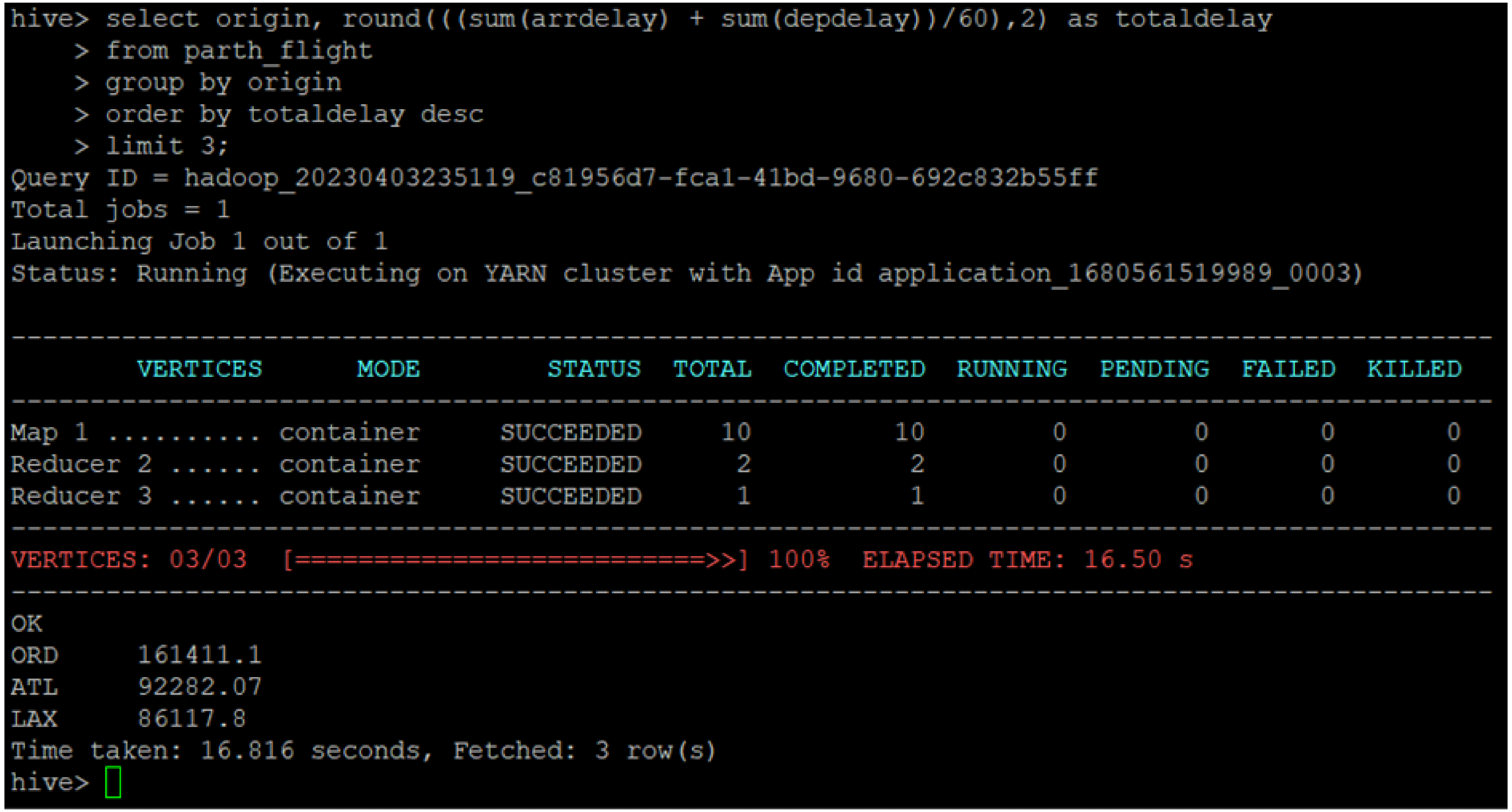


```
hive> select * from parth_flight
    > limit 10;
OK
NULL    NULL    NULL    NULL    NULL    NULL    NULL    NULL    UniqueCarrier   NULL    TailNum NULL    NULL    NULL    NULL    NULL    Origin  Dest    NULL    NULL    NULL    NULL    Cance
llationCode     NULL    NULL    NULL    NULL    NULL    NULL
2000    1       28      5       1647    1647    1906    1859    HP      154     N808AW  259     252     233     7       0       ATL     PHX     1587    15      11      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       29      6       1648    1647    1939    1859    HP      154     N653AW  291     252     239     40      1       ATL     PHX     1587    5       47      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       30      7       NULL    1647    NULL    1859    HP      154     N801AW  NULL    252     NULL    NULL    NULL    ATL     PHX     1587    0       0       1       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       31      1       1645    1647    1852    1859    HP      154     N806AW  247     252     226     -7      -2      ATL     PHX     1587    7       14      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       1       6       842     846     1057    1101    HP      609     N158AW  255     255     244     -4      -4      ATL     PHX     1587    3       8       0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       2       7       849     846     1148    1101    HP      609     N656AW  299     255     267     47      3       ATL     PHX     1587    8       24      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       3       1       844     846     1121    1101    HP      609     N803AW  277     255     244     20      -2      ATL     PHX     1587    6       27      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       1       6       1702    1657    1912    1908    HP      611     N652AW  250     251     232     4       5       ATL     PHX     1587    5       13      0       NA      0   N
ULL     NULL    NULL    NULL    NULL
2000    1       2       7       1658    1657    1901    1908    HP      611     N807AW  243     251     233     -7      1       ATL     PHX     1587    3       7       0       NA      0   N
ULL     NULL    NULL    NULL    NULL
Time taken: 1.69 seconds, Fetched: 10 row(s)
```

Parth Dodia

CASE 1

Determined the three **airports** with the highest delay time (in hours) for year 2000



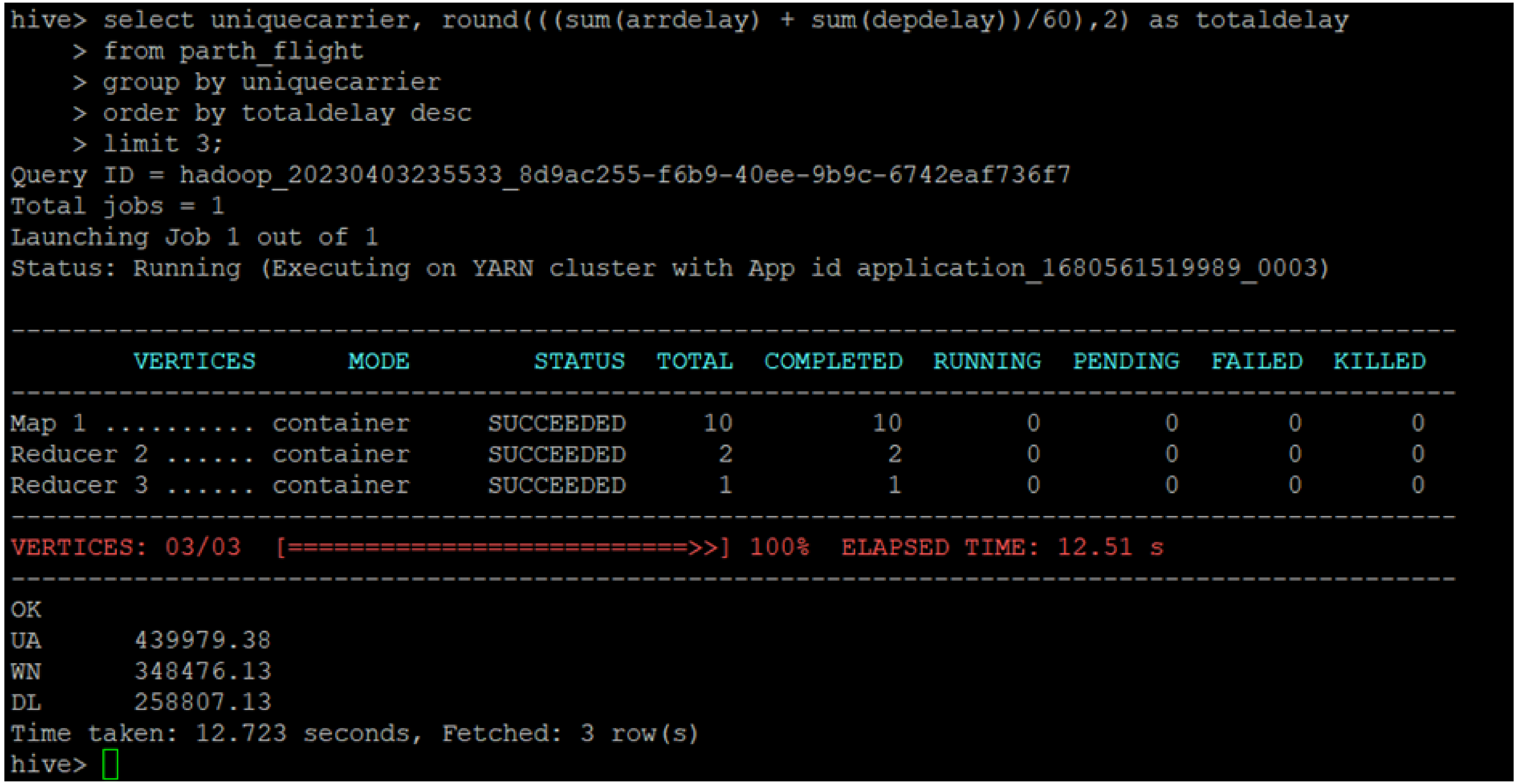
```
hive> select origin, round(((sum(arrdelay) + sum(depdelay))/60),2) as totaldelay
    > from parth_flight
    > group by origin
    > order by totaldelay desc
    > limit 3;
Query ID = hadoop_20230403235119_c81956d7-fca1-41bd-9680-692c832b55ff
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680561519989_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.50 s
----------------------------------------------------------------------------------------------
OK
ORD     161411.1
ATL     92282.07
LAX     86117.8
Time taken: 16.816 seconds, Fetched: 3 row(s)
hive> 
```

The three airports with the highest delay time:

ORD – 161411.1
ATL – 92282.7
LAX – 86117.8

CASE 2

Determined the three **carriers** with the highest delay time (in hours) for year 2000



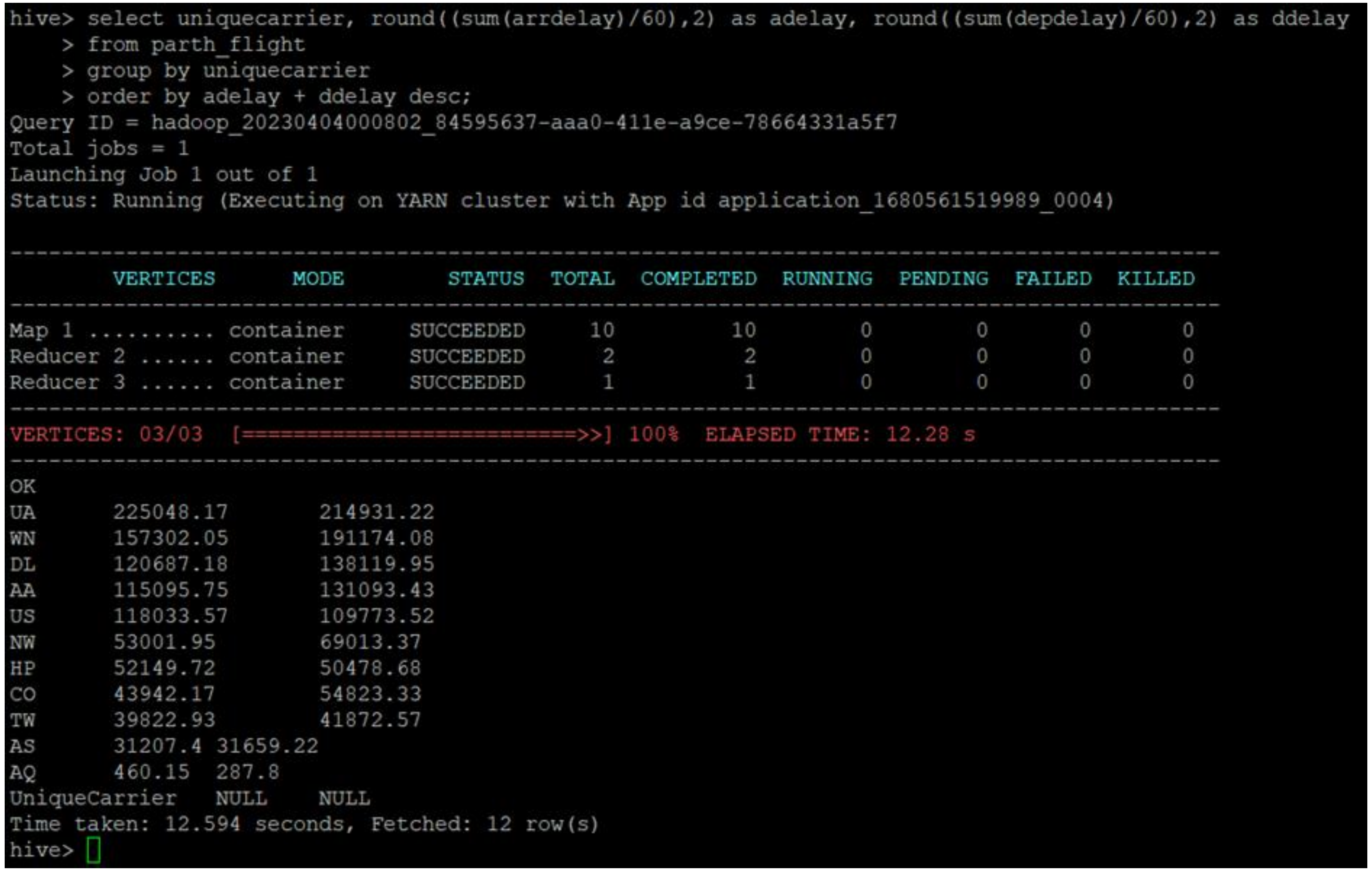
```
hive> select uniquecarrier, round(((sum(arrdelay) + sum(depdelay))/60),2) as totaldelay
    > from parth_flight
    > group by uniquecarrier
    > order by totaldelay desc
    > limit 3;
Query ID = hadoop_20230403235533_8d9ac255-f6b9-40ee-9b9c-6742eaf736f7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680561519989_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.51 s
----------------------------------------------------------------------------------------------
OK
UA      439979.38
WN      348476.13
DL      258807.13
Time taken: 12.723 seconds, Fetched: 3 row(s)
hive>
```

The three carriers with the highest delay time:

UA **-** 439979.38
WN **-** 348476.13
DL **-** 258807.13

CASE 3

Determine overall which type of delay (arrivals or departures) is the largest for your carriers.



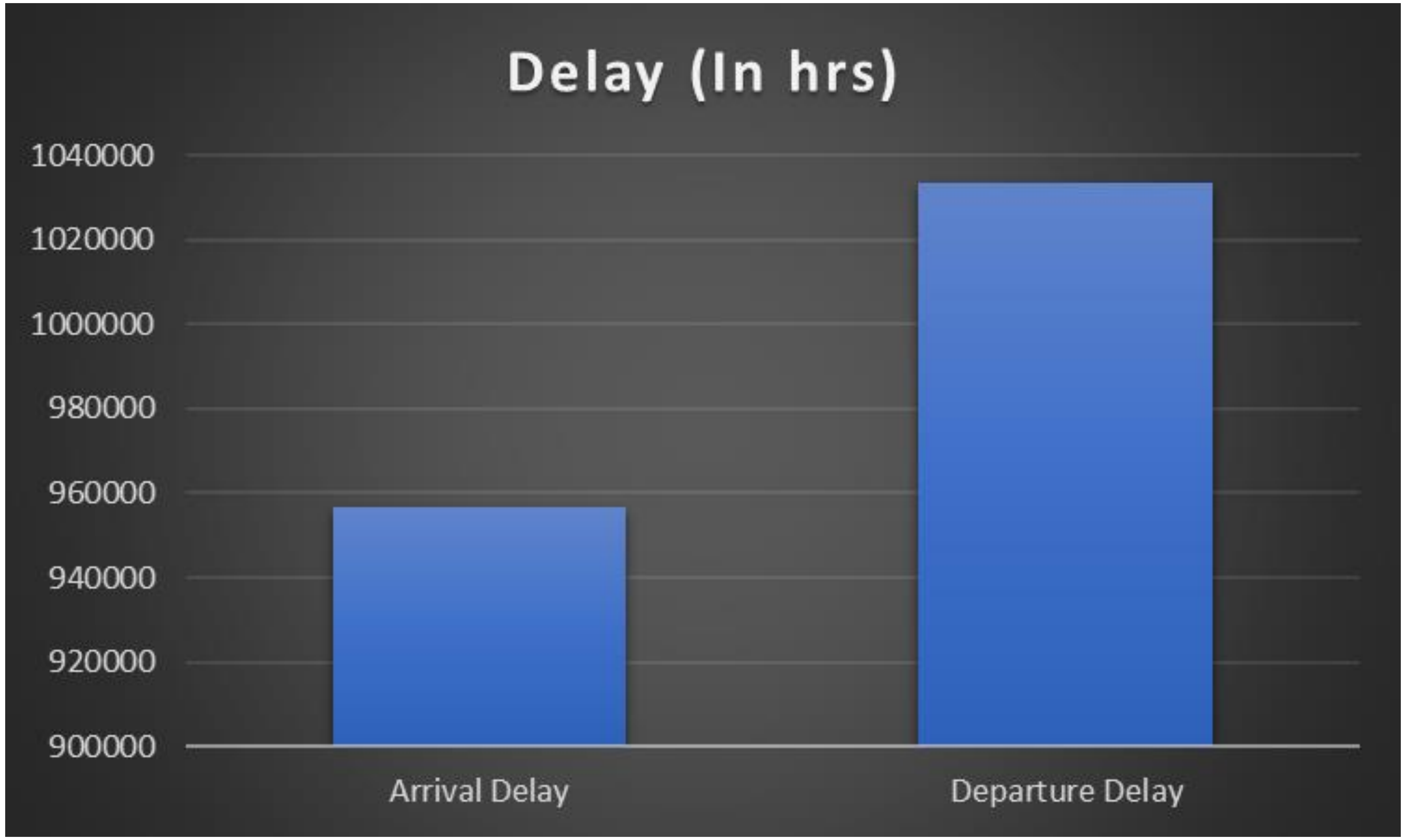
```
hive> select uniquecarrier, round((sum(arrdelay)/60),2) as adelay, round((sum(depdelay)/60),2) as ddelay
    > from parth_flight
    > group by uniquecarrier
    > order by adelay + ddelay desc;
Query ID = hadoop_20230404000802_84595637-aaa0-411e-a9ce-78664331a5f7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1680561519989_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.28 s
----------------------------------------------------------------------------------------------
OK
UA      225048.17       214931.22
WN      157302.05       191174.08
DL      120687.18       138119.95
AA      115095.75       131093.43
US      118033.57       109773.52
NW      53001.95        69013.37
HP      52149.72        50478.68
CO      43942.17        54823.33
TW      39822.93        41872.57
AS      31207.4 31659.22
AQ      460.15  287.8
UniqueCarrier   NULL    NULL
Time taken: 12.594 seconds, Fetched: 12 row(s)
hive> 
```

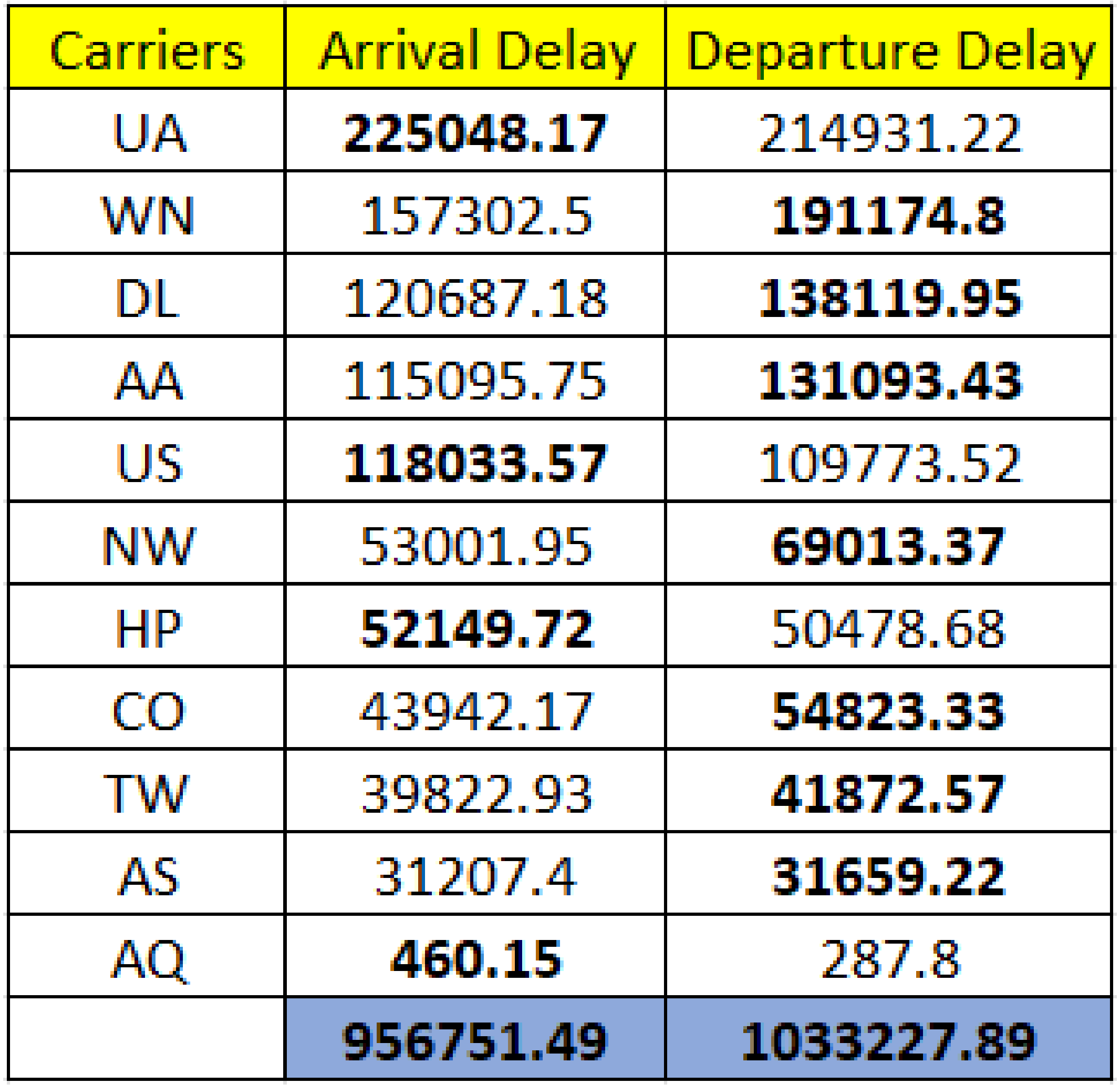
Carriers	Arrival Delay	Departure Delay
UA	**225048.17**	214931.22
WN	157302.5	**191174.8**
DL	120687.18	**138119.95**
AA	115095.75	**131093.43**
US	**118033.57**	109773.52
NW	53001.95	**69013.37**
HP	**52149.72**	50478.68
CO	43942.17	**54823.33**
TW	39822.93	**41872.57**
AS	31207.4	**31659.22**
AQ	**460.15**	287.8

Parth Dodia

The table shows the highlighted value of which delay is greater.

Analysis of Data 2000

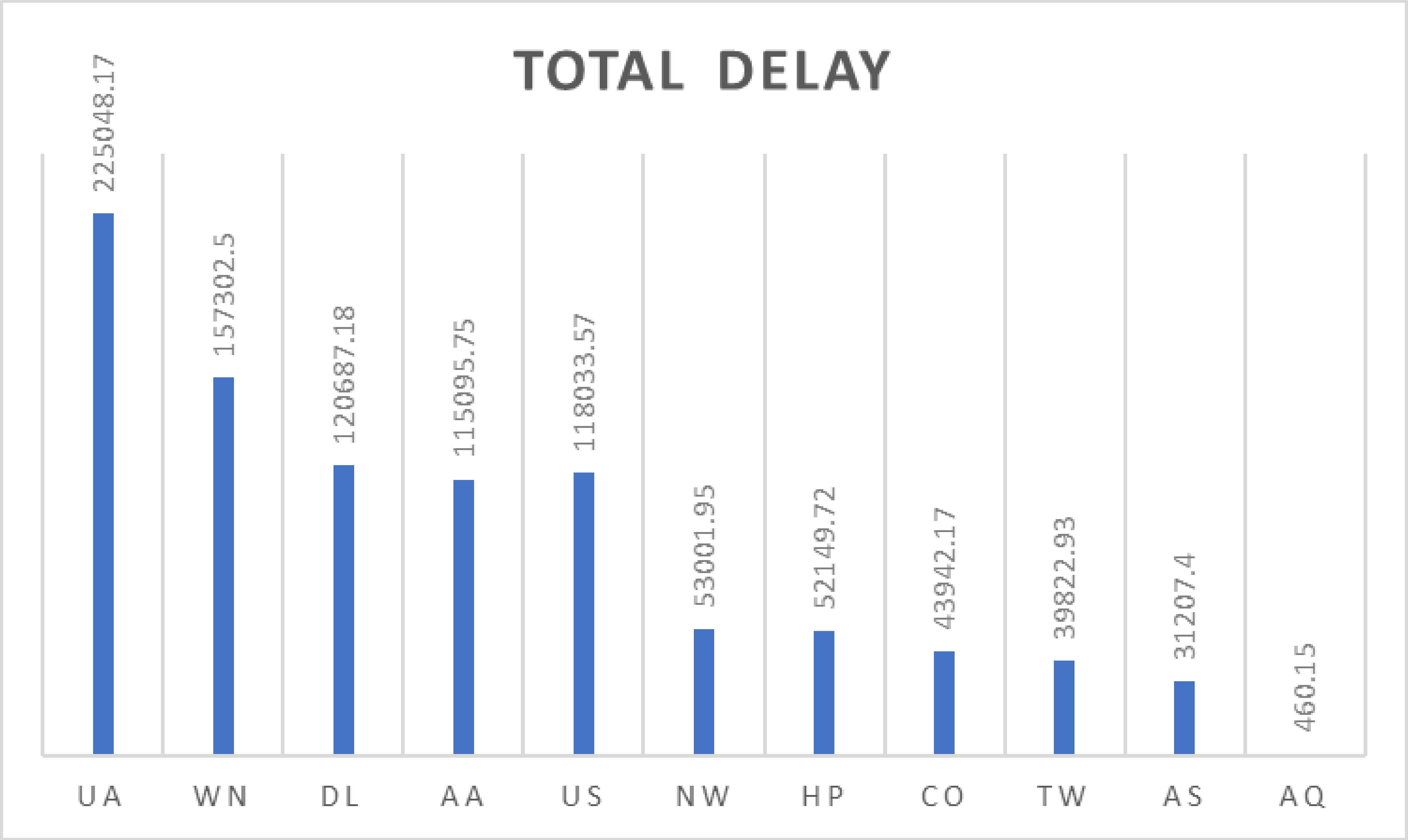




Carriers	Arrival Delay	Departure Delay
UA	**225048.17**	214931.22
WN	157302.5	**191174.8**
DL	120687.18	**138119.95**
AA	115095.75	**131093.43**
US	**118033.57**	109773.52
NW	53001.95	**69013.37**
HP	**52149.72**	50478.68
CO	43942.17	**54823.33**
TW	39822.93	**41872.57**
AS	31207.4	**31659.22**
AQ	**460.15**	287.8
	956751.49	**1033227.89**

Overall Departure Delay is Greater Than Arrival Delay

Analysis of Data 2000



Total Delay (in hrs) for each carrier

Parth Dodia

Final: Analysis

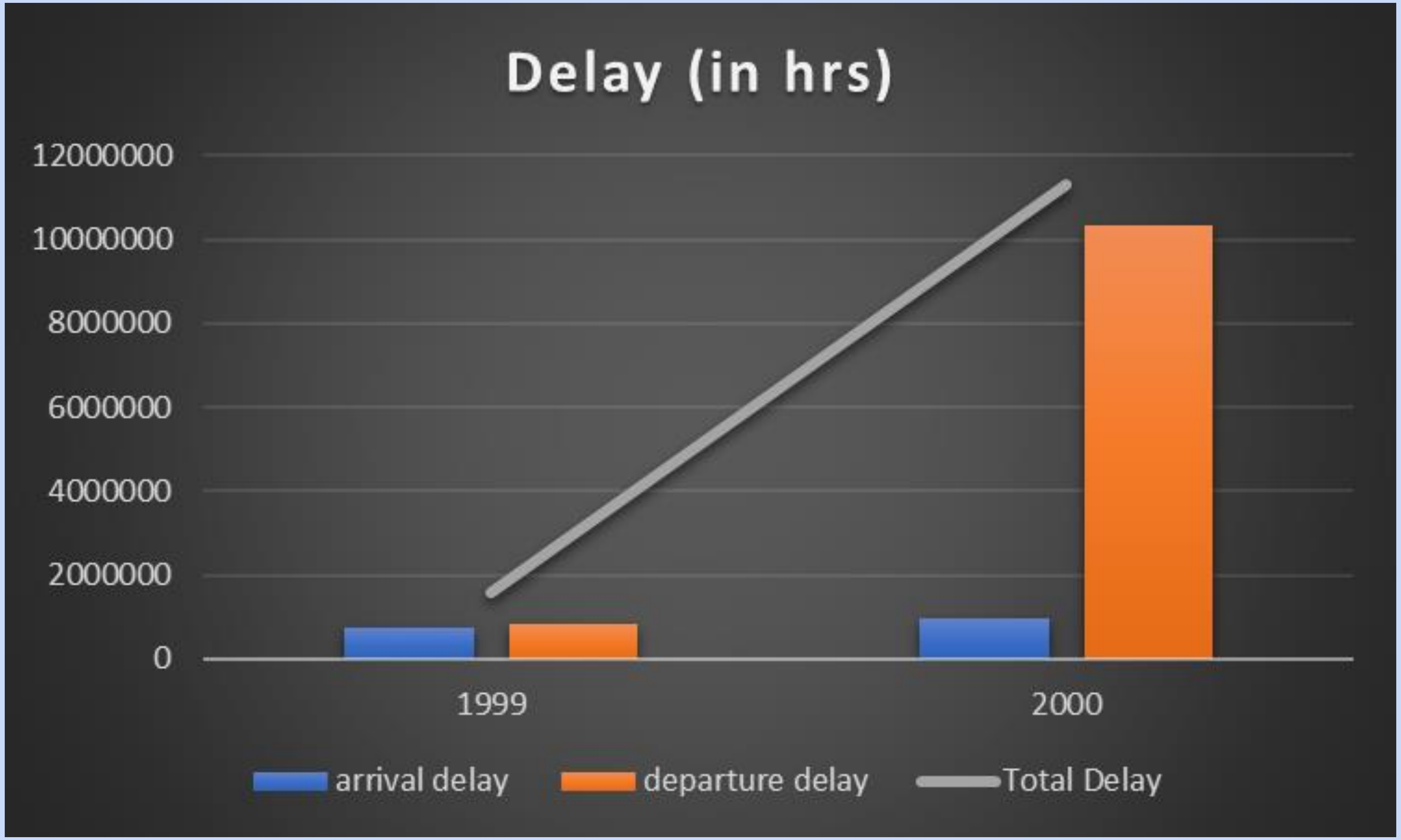
The overall highest delay time (in hours)

Year	Arrival Time Delay (In Hours)	Departure time Delay (In Hours)	Total Delay (In Hours)
1999	736,698.83	836,226.55	1,572,925.38
2000	956,751.49	1,033,227.89	1,989,979.38

The overall delay in the year 2000 is greater than that in the year 1999

Final: Analysis

The overall highest delay time (in hours)



THANK YOU FOR YOUR TIME

AWS Hadoop