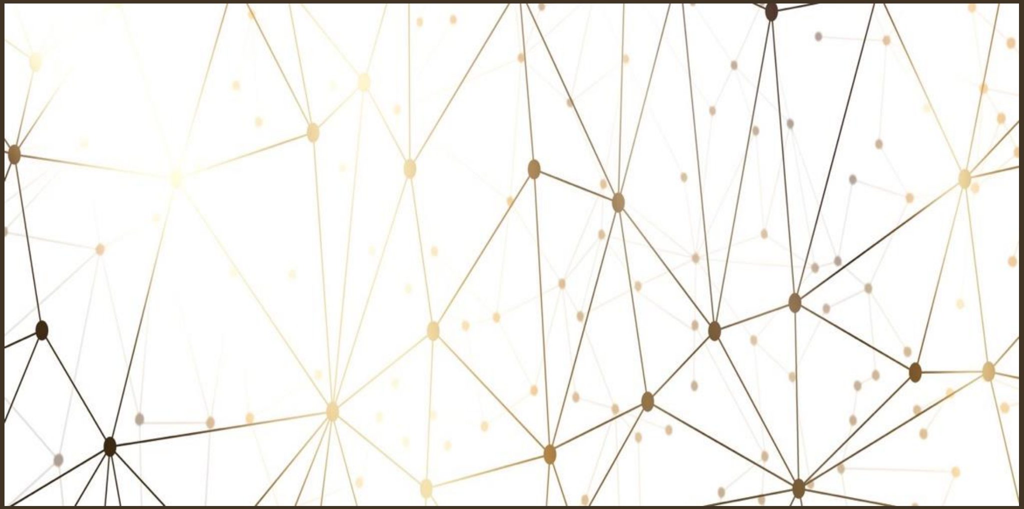
AWS PROJECT

Phase - 1

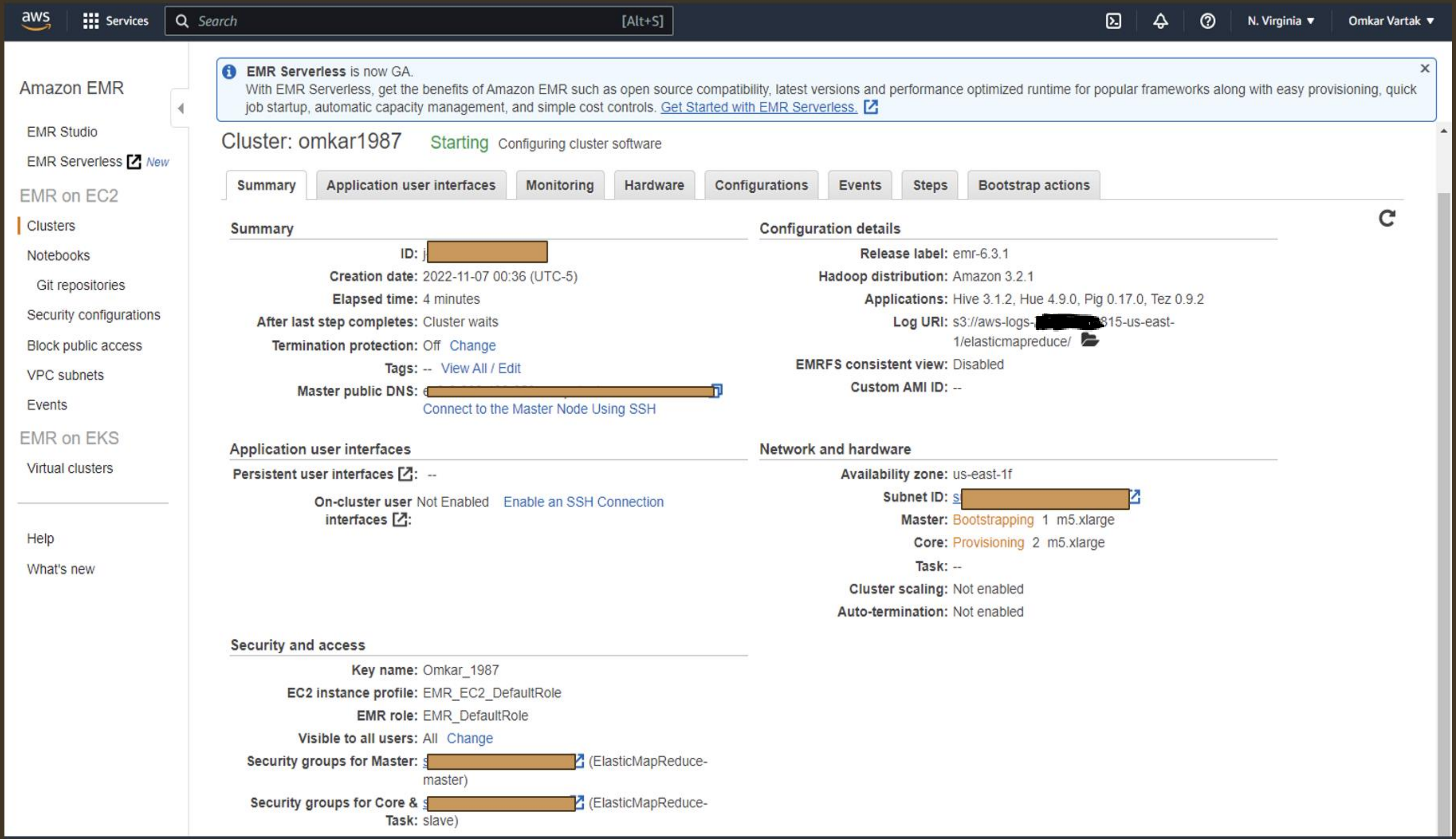
AWS PROJECT PHASE 1



- *Omkar Vartak –Year Selected 1987*

Year Selected 1987

CREATING THE CLUSTER



EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Cluster: omkar1987 Starting Configuring cluster software

Summary

ID: [redacted]
Creation date: 2022-11-07 00:36 (UTC-5)
Elapsed time: 4 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: [redacted]
Connect to the Master Node Using SSH

Configuration details

Release label: emr-6.3.1
Hadoop distribution: Amazon 3.2.1
Applications: Hive 3.1.2, Hue 4.9.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://aws-logs-[redacted]815-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: Not Enabled Enable an SSH Connection

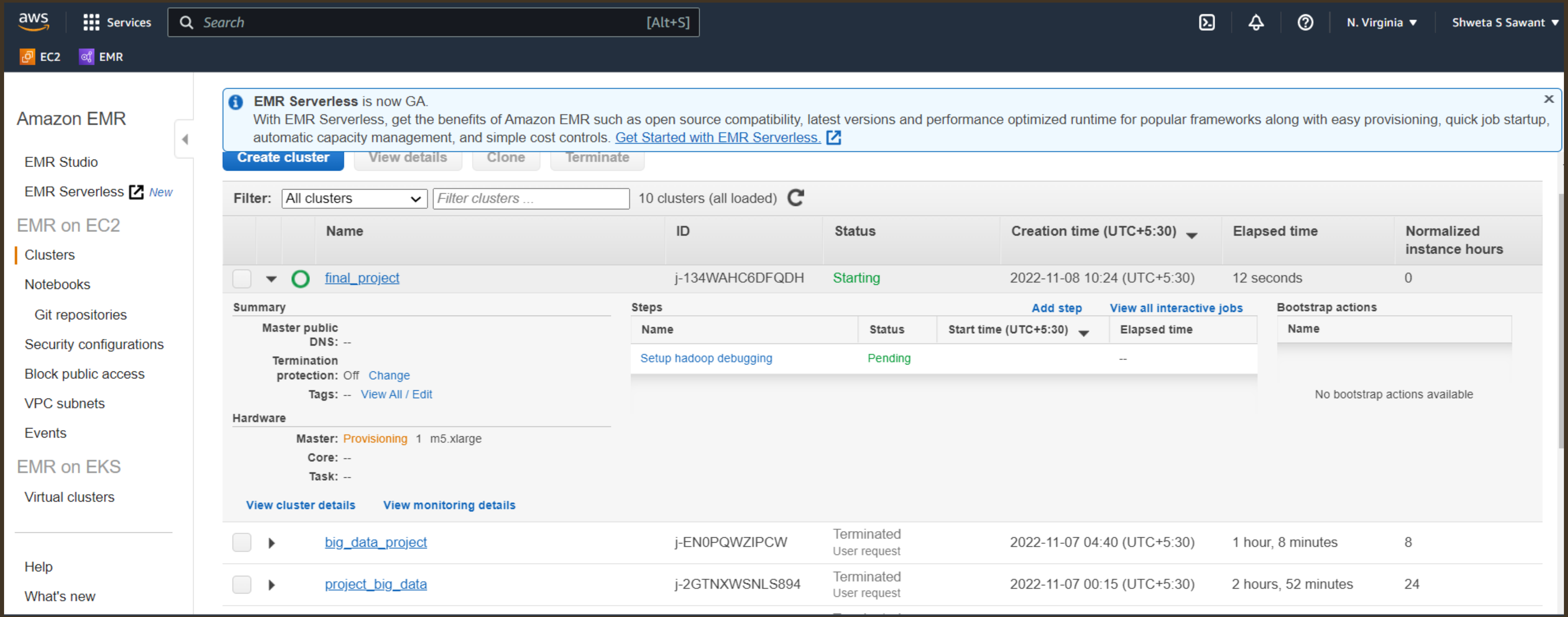
Network and hardware

Availability zone: us-east-1f
Subnet ID: [redacted]
Master: Bootstrapping 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: Omkar_1987
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: [redacted] (ElasticMapReduce-master)
Security groups for Core & Task: [redacted] (ElasticMapReduce-slave)

CREATING THE CLUSTER



EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. Get Started with EMR Serverless.

Filter: All clusters | Filter clusters ... | 10 clusters (all loaded)

Name	ID	Status	Creation time (UTC+5:30)	Elapsed time	Normalized instance hours
final_project	j-134WAHC6DFQDH	Starting	2022-11-08 10:24 (UTC+5:30)	12 seconds	0
big_data_project	j-EN0PQWZIPCW	Terminated User request	2022-11-07 04:40 (UTC+5:30)	1 hour, 8 minutes	8
project_big_data	j-2GTNXWSNLS894	Terminated User request	2022-11-07 00:15 (UTC+5:30)	2 hours, 52 minutes	24

Summary

Master public DNS: --
Termination protection: Off Change
Tags: -- View All / Edit

Hardware

Master: Provisioning 1 m5.xlarge
Core: --
Task: --

Steps Add step View all interactive jobs

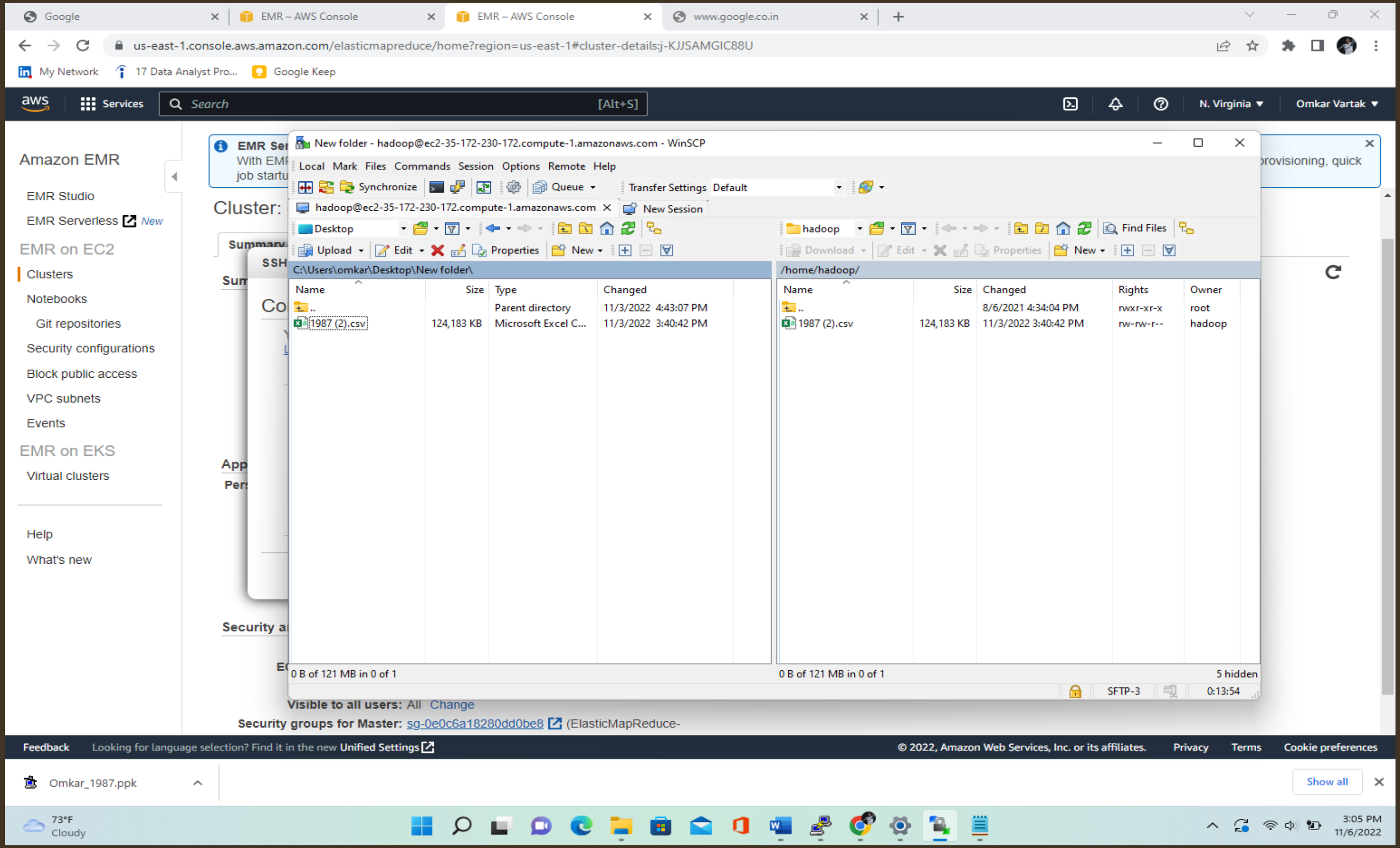
Name	Status	Start time (UTC+5:30)	Elapsed time
Setup hadoop debugging	Pending		--

Bootstrap actions

Name

No bootstrap actions available

USING WINSCP TO TRANSFER FILES



New folder - hadoop@ec2-35-172-230-172.compute-1.amazonaws.com - WinSCP

hadoop@ec2-35-172-230-172.compute-1.amazonaws.com

C:\Users\omkar\Desktop\New folder\

Name	Size	Type	Changed
..		Parent directory	11/3/2022 4:43:07 PM
1987 (2).csv	124,183 KB	Microsoft Excel C...	11/3/2022 3:40:42 PM

0 B of 121 MB in 0 of 1

/home/hadoop/

Name	Size	Changed	Rights	Owner
..		8/6/2021 4:34:04 PM	rwxr-xr-x	root
1987 (2).csv	124,183 KB	11/3/2022 3:40:42 PM	rw-rw-r--	hadoop

0 B of 121 MB in 0 of 1

5 hidden

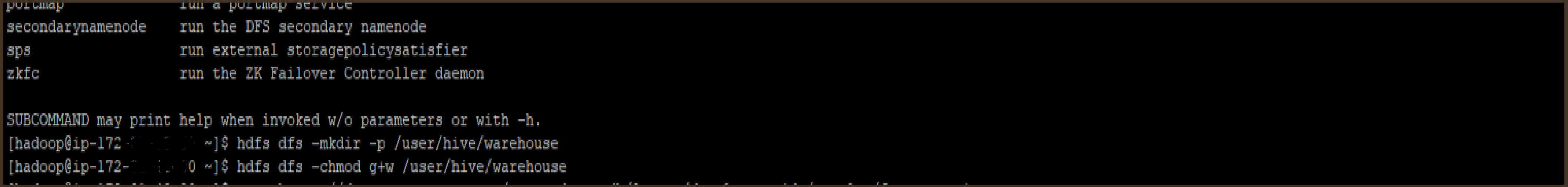
SFTP-3

0:13:54

Visible to all users: All Change

Security groups for Master: sg-0e0c6a18280dd0be8 (ElasticMapReduce-

MODIFY HDFS



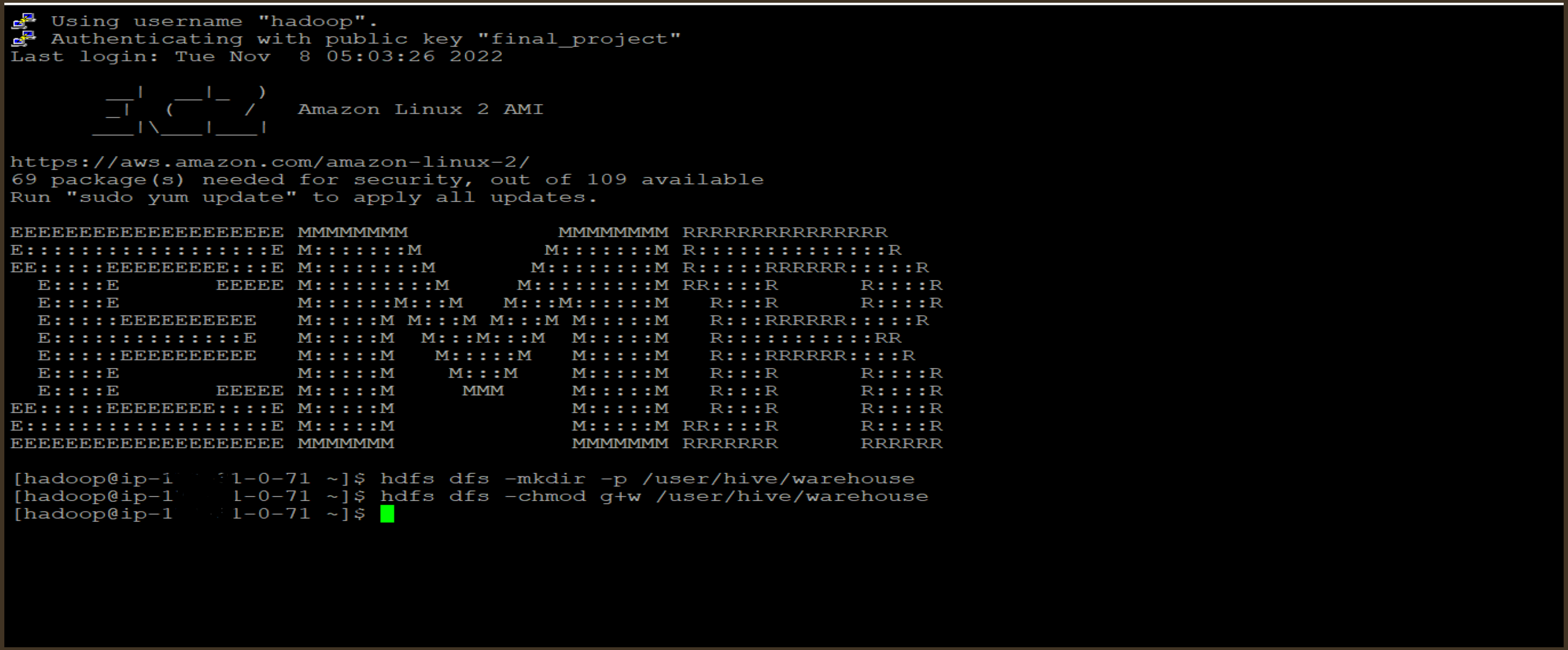
```
portmap              run a portmap service
secondarynamenode    run the DFS secondary namenode
sps                  run external storagepolicysatisfier
zkfc                 run the ZK Failover Controller daemon

SUBCOMMAND may print help when invoked w/o parameters or with -h.
[hadoop@ip-172[redacted] ~]$ hdfs dfs -mkdir -p /user/hive/warehouse
[hadoop@ip-172-[redacted]0 ~]$ hdfs dfs -chmod g+w /user/hive/warehouse
```

After this we will extract the downloaded bz2 file using any extractor.

HDFS COMMANDS

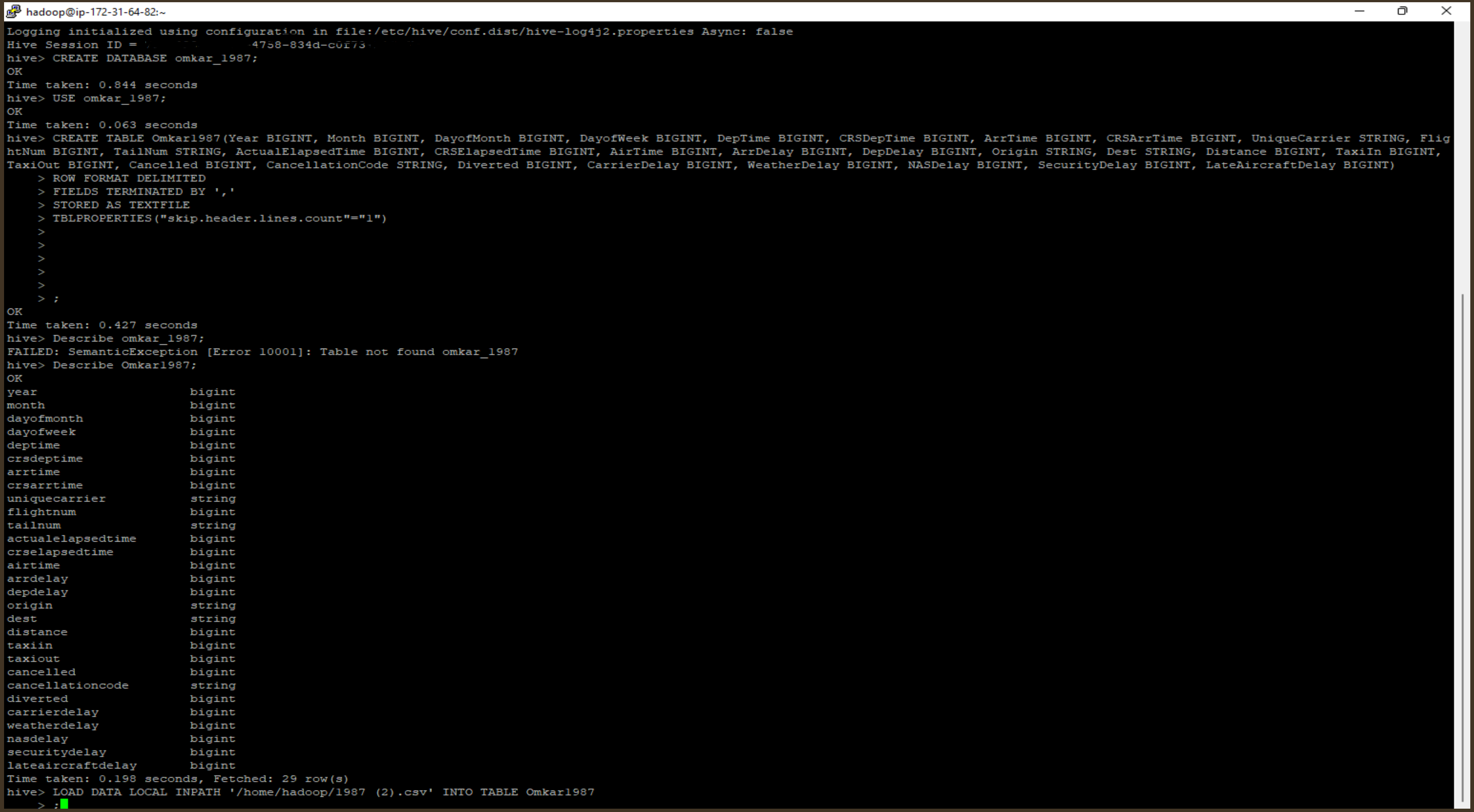
Below commands were used:

1. hdfs dfs -mkdir -p /user/hive/warehouse
2. hdfs dfs -chmod g+w /user/hive/warehouse



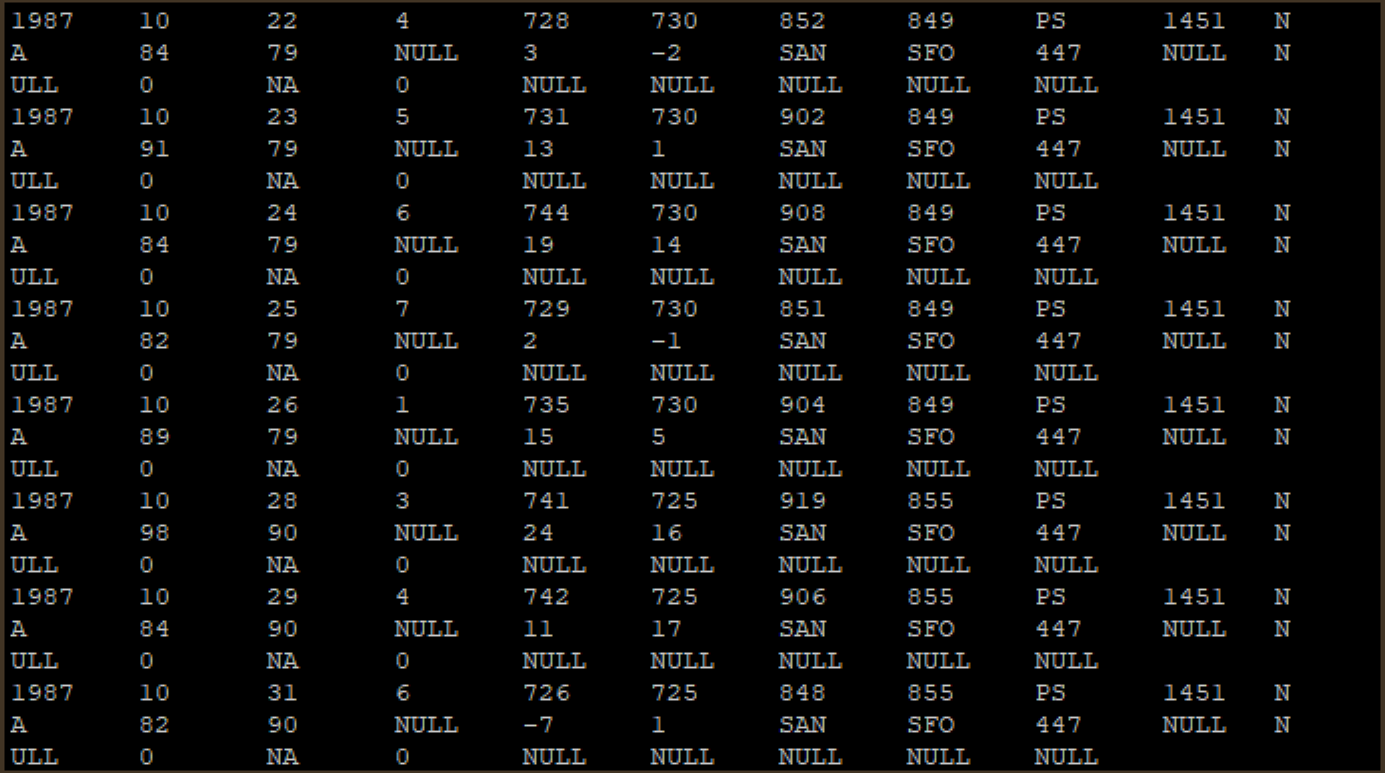
```
Using username "hadoop".
Authenticating with public key "final_project"
Last login: Tue Nov  8 05:03:26 2022

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

https://aws.amazon.com/amazon-linux-2/
69 package(s) needed for security, out of 109 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-1[redacted]1-0-71 ~]$ hdfs dfs -mkdir -p /user/hive/warehouse
[hadoop@ip-1[redacted]1-0-71 ~]$ hdfs dfs -chmod g+w /user/hive/warehouse
[hadoop@ip-1[redacted]1-0-71 ~]$ 
```

CREATING DATABASE AND TABLES, LOADING THE DATA.



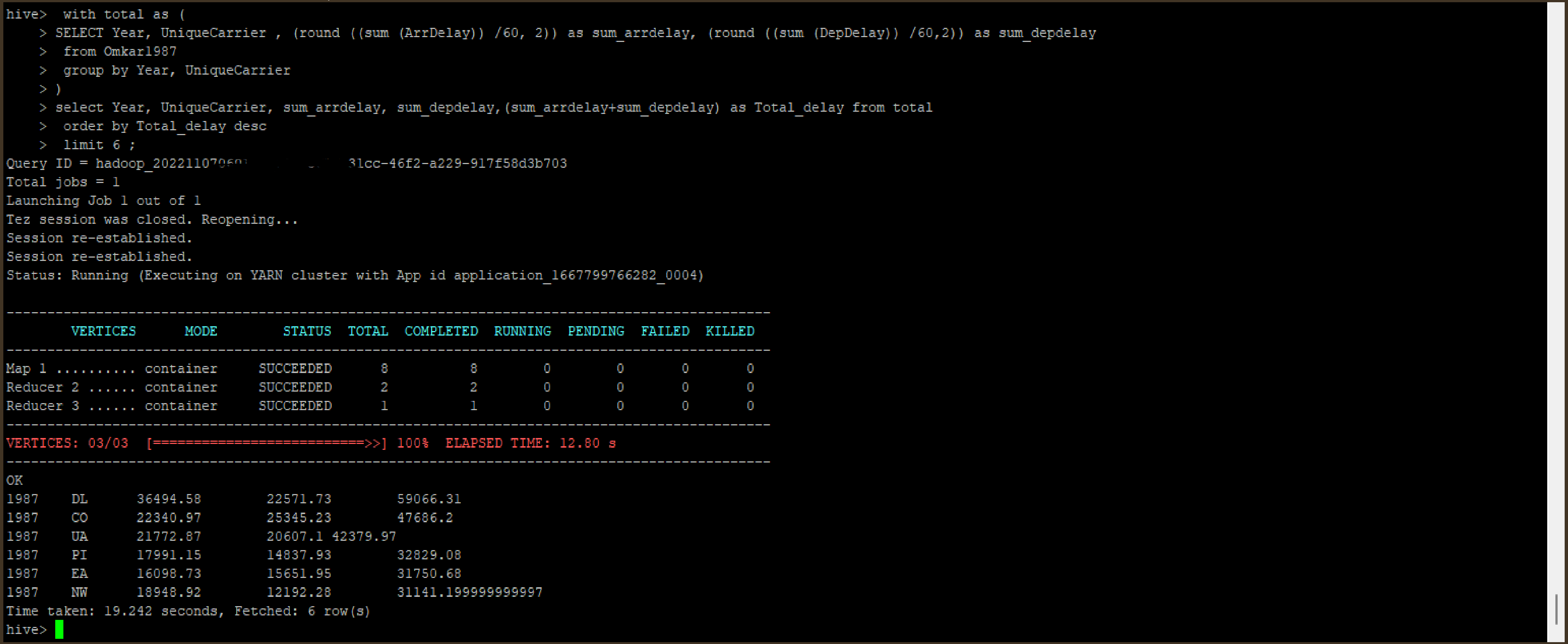
```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
Hive Session ID = [redacted]4758-834d-c0f73[redacted]
hive> CREATE DATABASE omkar_1987;
OK
Time taken: 0.844 seconds
hive> USE omkar_1987;
OK
Time taken: 0.063 seconds
hive> CREATE TABLE Omkar1987(Year BIGINT, Month BIGINT, DayofMonth BIGINT, DayofWeek BIGINT, DepTime BIGINT, CRSDepTime BIGINT, ArrTime BIGINT, CRSArrTime BIGINT, UniqueCarrier STRING, Flig
htNum BIGINT, TailNum STRING, ActualElapsedTime BIGINT, CRSElapsedTime BIGINT, AirTime BIGINT, ArrDelay BIGINT, DepDelay BIGINT, Origin STRING, Dest STRING, Distance BIGINT, TaxiIn BIGINT,
TaxiOut BIGINT, Cancelled BIGINT, CancellationCode STRING, Diverted BIGINT, CarrierDelay BIGINT, WeatherDelay BIGINT, NASDelay BIGINT, SecurityDelay BIGINT, LateAircraftDelay BIGINT)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > TBLPROPERTIES("skip.header.lines.count"="1")
    >
    >
    >
    >
    >
    > ;
OK
Time taken: 0.427 seconds
hive> Describe omkar_1987;
FAILED: SemanticException [Error 10001]: Table not found omkar_1987
hive> Describe Omkar1987;
OK
year                    bigint
month                   bigint
dayofmonth              bigint
dayofweek               bigint
deptime                 bigint
crsdeptime              bigint
arrtime                 bigint
crsarrtime              bigint
uniquecarrier           string
flightnum               bigint
tailnum                 string
actualelapsedtime       bigint
crselapsedtime          bigint
airtime                 bigint
arrdelay                bigint
depdelay                bigint
origin                  string
dest                    string
distance                bigint
taxiin                  bigint
taxiout                 bigint
cancelled               bigint
cancellationcode        string
diverted                bigint
carrierdelay            bigint
weatherdelay            bigint
nasdelay                bigint
securitydelay           bigint
lateaircraftdelay       bigint
Time taken: 0.198 seconds, Fetched: 29 row(s)
hive> LOAD DATA LOCAL INPATH '/home/hadoop/1987 (2).csv' INTO TABLE Omkar1987
    > ;
```

DISPLAYING THE DATA USING THE LIMIT FUNCTION:-



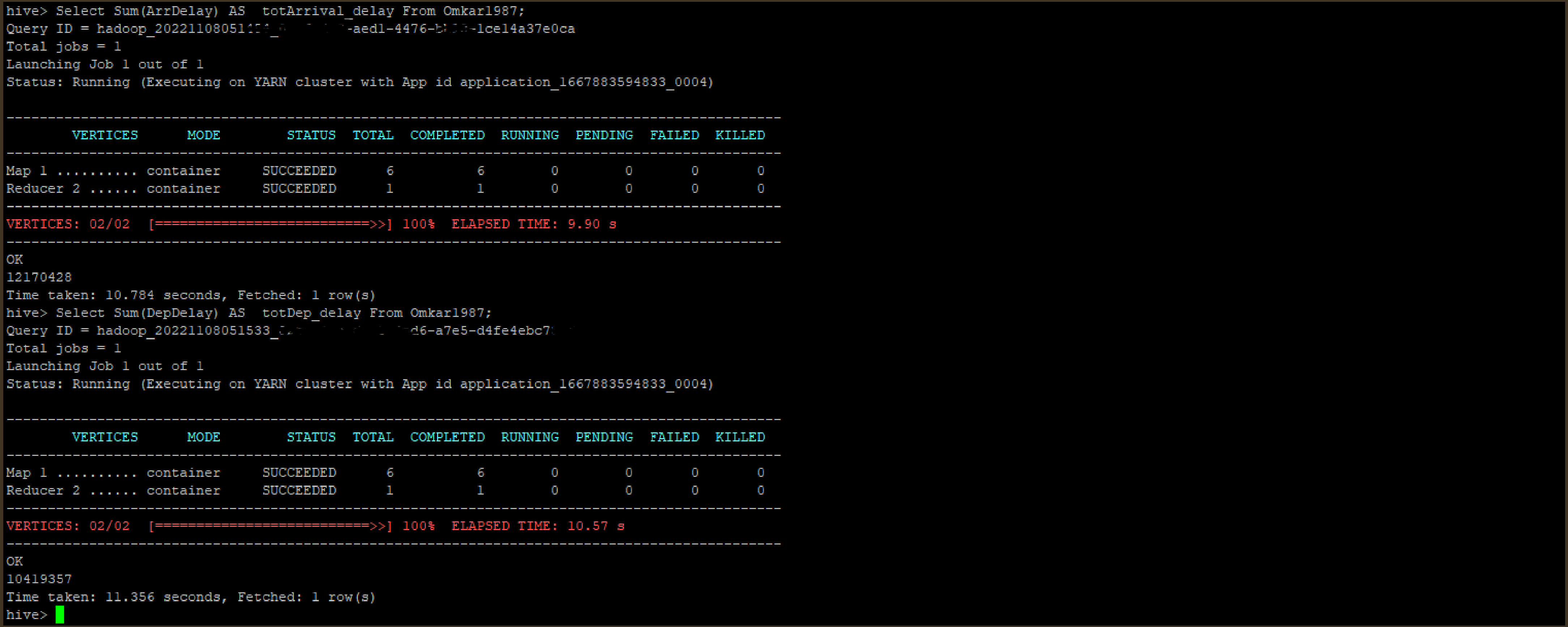
```
1987    10      22      4       728     730     852     849     PS      1451   N
A       84      79      NULL    3       -2      SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      23      5       731     730     902     849     PS      1451   N
A       91      79      NULL    13      1       SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      24      6       744     730     908     849     PS      1451   N
A       84      79      NULL    19      14      SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      25      7       729     730     851     849     PS      1451   N
A       82      79      NULL    2       -1      SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      26      1       735     730     904     849     PS      1451   N
A       89      79      NULL    15      5       SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      28      3       741     725     919     855     PS      1451   N
A       98      90      NULL    24      16      SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      29      4       742     725     906     855     PS      1451   N
A       84      90      NULL    11      17      SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
1987    10      31      6       726     725     848     855     PS      1451   N
A       82      90      NULL    -7      1       SAN     SFO     447     NULL   N
ULL     0       NA      0       NULL    NULL    NULL    NULL    NULL
```

QUERY DETERMINE THE THREE **CARRIERS** WITH THE HIGHEST DELAY TIME (IN HOURS)



```
hive>  with total as (
    > SELECT Year, UniqueCarrier , (round ((sum (ArrDelay)) /60, 2)) as sum_arrdelay, (round ((sum (DepDelay)) /60,2)) as sum_depdelay
    >  from Omkar1987
    >  group by Year, UniqueCarrier
    > )
    > select Year, UniqueCarrier, sum_arrdelay, sum_depdelay,(sum_arrdelay+sum_depdelay) as Total_delay from total
    >  order by Total_delay desc
    >  limit 6 ;
Query ID = hadoop_20221107[redacted]31cc-46f2-a229-917f58d3b703
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_1667799766282_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        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: 12.80 s
----------------------------------------------------------------------------------------------
OK
1987    DL      36494.58        22571.73        59066.31
1987    CO      22340.97        25345.23        47686.2
1987    UA      21772.87        20607.1 42379.97
1987    PI      17991.15        14837.93        32829.08
1987    EA      16098.73        15651.95        31750.68
1987    NW      18948.92        12192.28        31141.199999999997
Time taken: 19.242 seconds, Fetched: 6 row(s)
hive>
```

QUERY DETERMINE OVERALL WHICH TYPE OF DELAY (ARRIVALS OR DEPARTURES) IS THE LARGEST FOR **AIRPORTS**

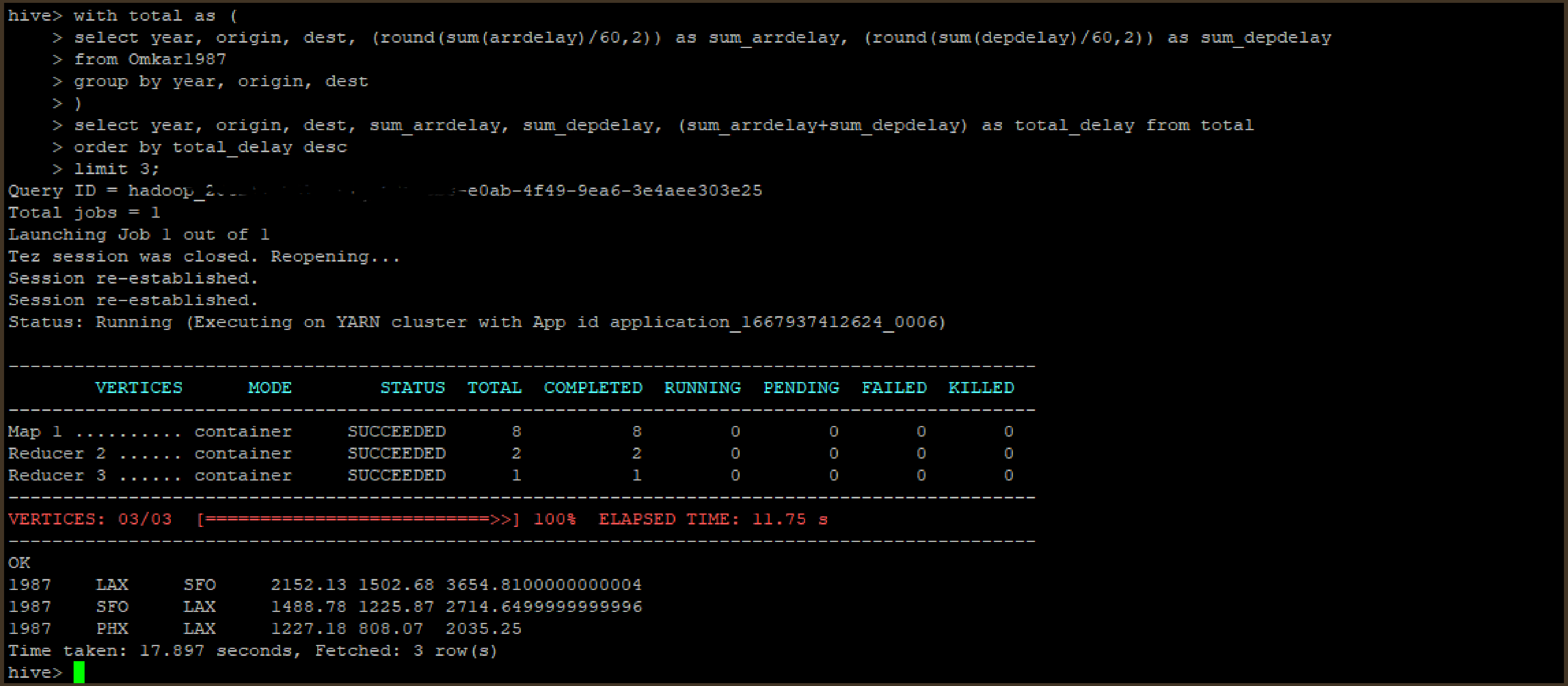


```
hive> Select Sum(ArrDelay) AS  totArrival_delay From Omkar1987;
Query ID = hadoop_20221108051[redacted]-aed1-4476-b[redacted]-1ce14a37e0ca
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667883594833_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.90 s
----------------------------------------------------------------------------------------------
OK
12170428
Time taken: 10.784 seconds, Fetched: 1 row(s)
hive> Select Sum(DepDelay) AS  totDep_delay From Omkar1987;
Query ID = hadoop_20221108051533_[redacted]d6-a7e5-d4fe4ebc7[redacted]
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667883594833_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 10.57 s
----------------------------------------------------------------------------------------------
OK
10419357
Time taken: 11.356 seconds, Fetched: 1 row(s)
hive>
```

QUERY DETERMINES THE THREE **AIRPORTS** WITH THE HIGHEST DELAY TIME (IN HOURS)

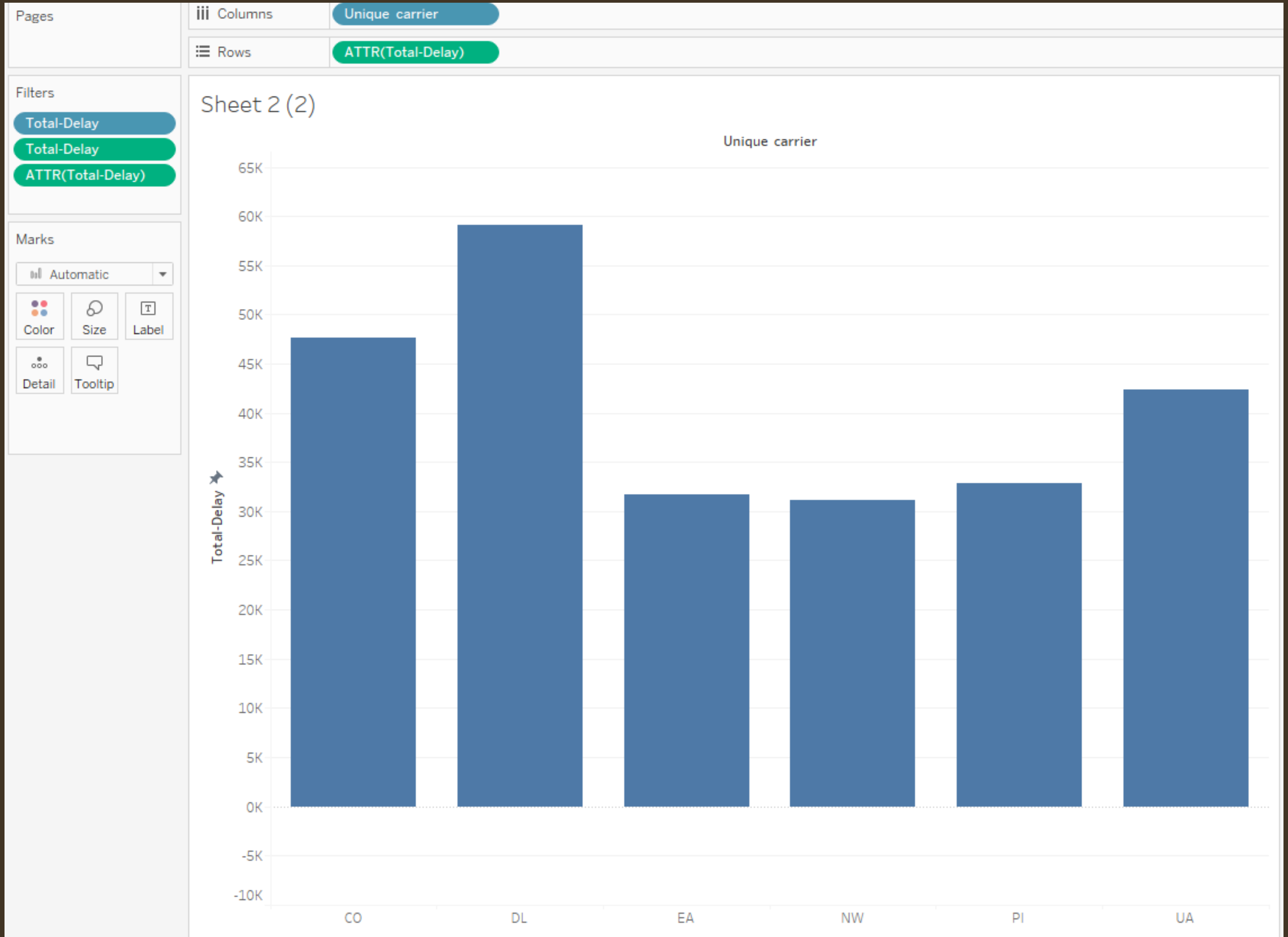
- We will divide the result by 60, In order to get the results in hours we will then use the round function.
- After that we will get the result in hours which can be used in the visualization process.
- Arr_delay = 12170428/60=202840.46
- Dep_delay=10419357/60=173655.95
- In conclusion Arrival delay is more than that of Departure delay

QUERY DETERMINES THE THREE **AIRPORTS** WITH THE HIGHEST DELAY TIME (IN HOURS)



```
hive> with total as (
    > select year, origin, dest, (round(sum(arrdelay)/60,2)) as sum_arrdelay, (round(sum(depdelay)/60,2)) as sum_depdelay
    > from Omkar1987
    > group by year, origin, dest
    > )
    > select year, origin, dest, sum_arrdelay, sum_depdelay, (sum_arrdelay+sum_depdelay) as total_delay from total
    > order by total_delay desc
    > limit 3;
Query ID = hadoop_2[redacted]-e0ab-4f49-9ea6-3e4aee303e25
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_1667937412624_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        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: 11.75 s
----------------------------------------------------------------------------------------------
OK
1987    LAX     SFO     2152.13 1502.68 3654.8100000000004
1987    SFO     LAX     1488.78 1225.87 2714.6499999999996
1987    PHX     LAX     1227.18 808.07  2035.25
Time taken: 17.897 seconds, Fetched: 3 row(s)
hive>
```

VISUALIZATION- TOOL USED TABLEAU FOR YEAR 1987. DETERMINE THE THREE **CARRIERS** WITH THE HIGHEST DELAY TIME (IN HOURS)



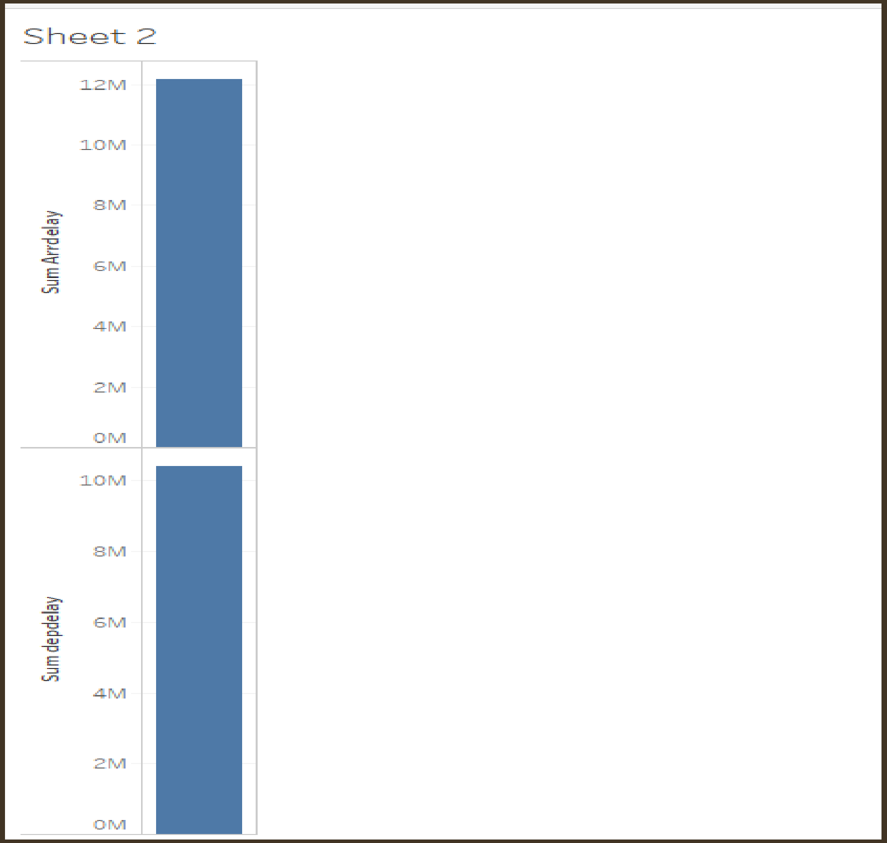
Columns: Unique carrier

Rows: ATTR(Total-Delay)

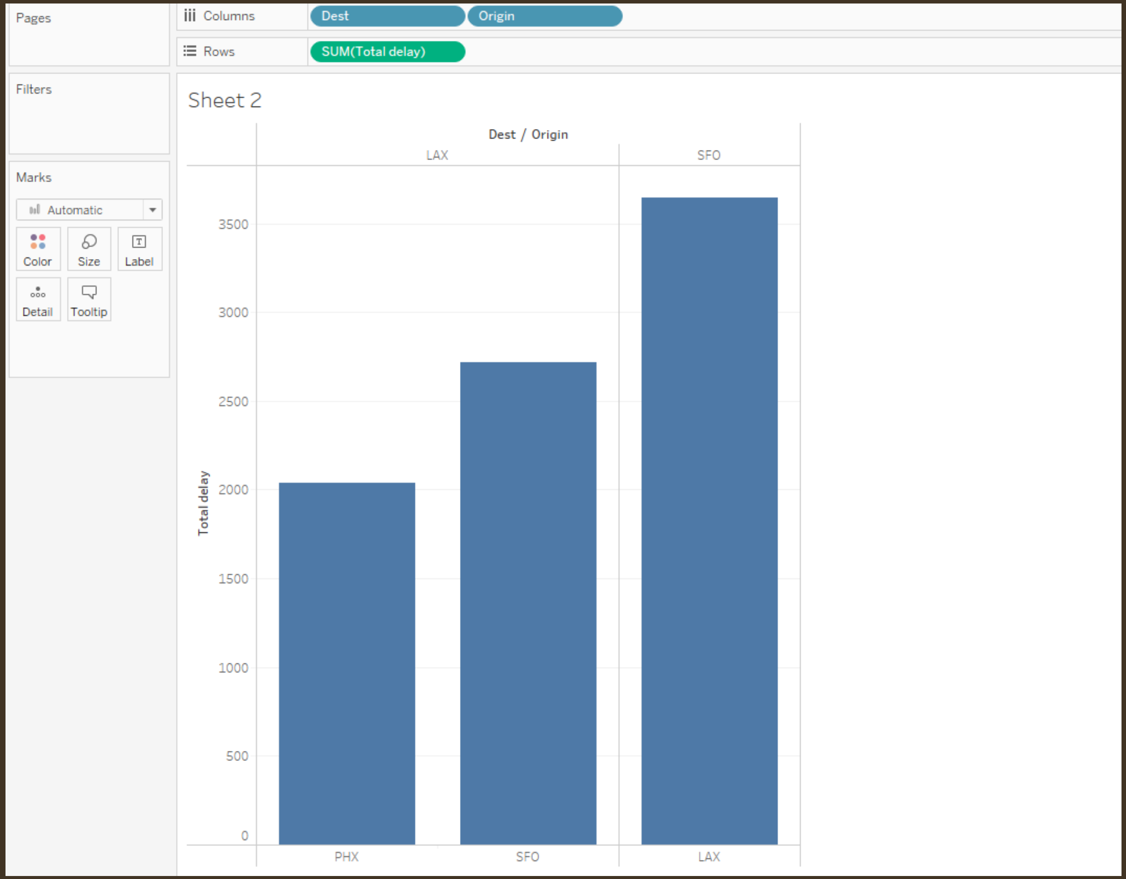
Filters: Total-Delay, Total-Delay, ATTR(Total-Delay)

Marks: Automatic

DETERMINE OVERALL WHICH TYPE OF DELAY (ARRIVALS OR DEPARTURES) IS THE LARGEST FOR **AIRPORTS**



DETERMINE THE THREE **AIRPORTS** WITH THE HIGHEST DELAY TIME (IN HOURS



Columns: Dest Origin

Rows: SUM(Total delay)

Sheet 2