

19162121031

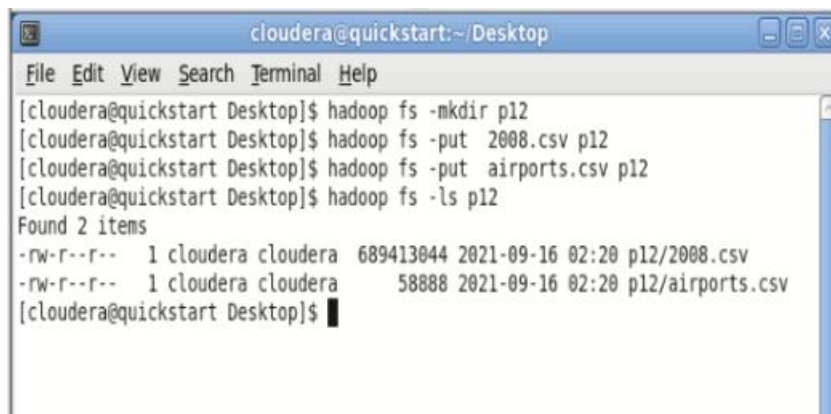
SMIT R PATEL

BDA

SEM 5

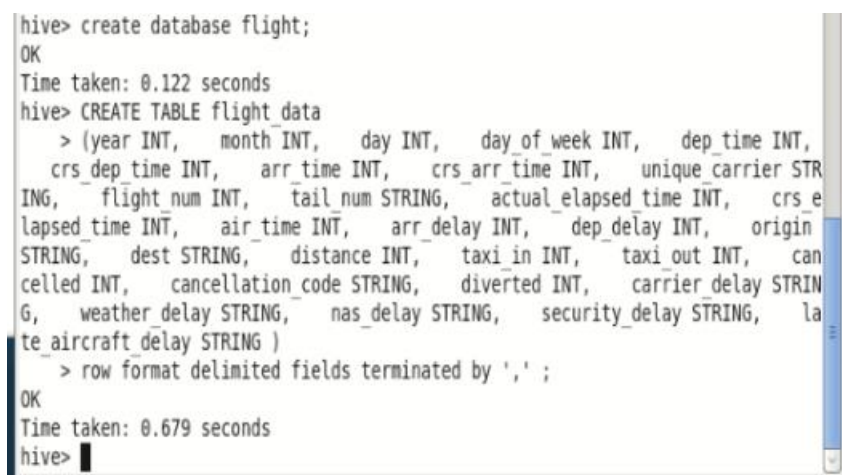
PRACTICAL 12

Here, first of all we make directory p12 in Hadoop and put file 2008.csv and airports.csv in p12 directory from Cloudera Desktop :



```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hadoop fs -mkdir p12
[cloudera@quickstart Desktop]$ hadoop fs -put 2008.csv p12
[cloudera@quickstart Desktop]$ hadoop fs -put airports.csv p12
[cloudera@quickstart Desktop]$ hadoop fs -ls p12
Found 2 items
-rw-r--r-- 1 cloudera cloudera 689413044 2021-09-16 02:20 p12/2008.csv
-rw-r--r-- 1 cloudera cloudera 58888 2021-09-16 02:20 p12/airports.csv
[cloudera@quickstart Desktop]$
```

1 Create hive table, flight_data:



```
hive> create database flight;
OK
Time taken: 0.122 seconds
hive> CREATE TABLE flight_data
> (year INT, month INT, day INT, day_of_week INT, dep_time INT,
crs_dep_time INT, arr_time INT, crs_arr_time INT, unique_carrier STR
ING, flight_num INT, tail_num STRING, actual_elapsed_time INT, crs_e
lapsed_time INT, air_time INT, arr_delay INT, dep_delay INT, origin
STRING, dest STRING, distance INT, taxi_in INT, taxi_out INT, can
celled INT, cancellation_code STRING, diverted INT, carrier_delay STRIN
G, weather_delay STRING, nas_delay STRING, security_delay STRING, la
te_aircraft_delay STRING )
> row format delimited fields terminated by ',' ;
OK
Time taken: 0.679 seconds
hive>
```

2 Load the data into the table:

```
hive> LOAD DATA LOCAL INPATH '2008.csv' OVERWRITE INTO TABLE flight_data;
Loading data to table default.flight_data
Table default.flight_data stats: [numFiles=1, numRows=0, totalSize=689413044, rawDataSize=0]
OK
Time taken: 10.343 seconds
hive>
```

3 Ensure the table got created and loaded fine:

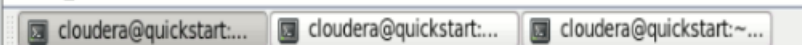
```
cloudera@quickstar
File Edit View Search Terminal Help
hive> show tables;
OK
ad
add
employee
flight_data
order
Time taken: 0.142 seconds, Fetched: 5 row(s)
hive> SELECT
> *
> FROM
> flight_data
> LIMIT 10;
OK
2008 1 3 4 2003 1955 2211 2225 WN 335 N
712SW 128 150 116 -14 8 IAD TPA 810 4 8
0 0 NA NA NA NA NA
2008 1 3 4 754 735 1002 1000 WN 3231 N
772SW 128 145 113 2 19 IAD TPA 810 5 1
0 0 0 NA NA NA NA NA
2008 1 3 4 628 620 804 750 WN 448 N
428WN 96 90 76 14 8 IND BWI 515 3 1
7 0 0 NA NA NA NA NA
2008 1 3 4 926 930 1054 1100 WN 1746 N
612SW 88 90 78 -6 -4 IND BWI 515 3 7
0 0 NA NA NA NA NA
2008 1 3 4 1829 1755 1959 1925 WN 3920 N
464WN 90 90 77 34 34 IND BWI 515 3 1
0 0 0 2 0 0 0 32
2008 1 3 4 1940 1915 2121 2110 WN 378 N
```

- 4 Query the table. Find average arrival delay for all flights departing SFO in January:

```
hive> SELECT
  > avg(arr_delay)
  > FROM
  > flight_data
  > WHERE
  > month=1
  > AND origin='SFO';
Query ID = cloudera_20210916022525_bb3453ac-b41f-4d85-8dde-c65c3c97c1b0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1631587610575_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1631587610575_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1631587610575_0001
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 1
2021-09-16 02:25:28,241 Stage-1 map = 0%, reduce = 0%
2021-09-16 02:26:29,176 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 27.31 sec
2021-09-16 02:26:36,301 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 82.62 sec
2021-09-16 02:26:37,337 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 113.86 sec
2021-09-16 02:26:38,426 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 148.49 sec
2021-09-16 02:26:42,605 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 150.15 sec
MapReduce Total cumulative CPU time: 2 minutes 30 seconds 150 msec
Ended Job = job_1631587610575_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3 Reduce: 1 Cumulative CPU: 150.15 sec HDFS Read: 689456416 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 30 seconds 150 msec
OK
28.669403949068094
Time taken: 102.413 seconds, Fetched: 1 row(s)
hive>
```

- 5 On hive shell: create the airports table

```
hive> create table airports
  > ( name STRING, country STRING, area_code INT, code STRING)
  > row format delimited fields terminated by ',' ;
OK
Time taken: 0.101 seconds
hive>
```

The image shows a screenshot of a terminal window with three tabs. The active tab is titled 'cloudera@quickstart:~...' and displays the Hive shell session from the previous block, including the 'create table airports' command and its successful execution. The other two tabs are also titled 'cloudera@quickstart:~...' but are not active.

6 Load data into airports table:

```
hive> LOAD DATA LOCAL INPATH 'airports.csv' OVERWRITE INTO TABLE airports;
Loading data to table default.airports
Table default.airports stats: [numFiles=1, numRows=0, totalSize=58888, rawDataSize=0]
OK
Time taken: 0.399 seconds
hive>
```

7 On hive shell, list some rows from the airports table:

```
hive> SELECT * FROM airports LIMIT 10;
OK
Key West Nas /Boca Chica Field (private U. S. Navy )   US      67      NQX
A L Mangham Jr. Regional                               US      67      OCH
AAF Heliport      US      67      AYE
Aberdeen Regional    US      67      ABR
Abilene Regional    US      67      ABI
Abraham Lincoln Capital US      67      SPI
Acadiana Regional    US      67      ARA
Accomack County US      67      MFV
Ada Municipal      US      67      ADT
Adak Island Ns     US      67      ADK
Time taken: 0.113 seconds, Fetched: 10 row(s)
hive>
```

8 On hive shell: run a join query to find the average delay in January 2008 for each airport and to print out the airport's name:

```
hive> SET hive.auto.convert.join=false;
hive> SELECT name, AVG(arr_delay) FROM flight_data f INNER JOIN airports a ON (f.origin = a.code) WHERE month=1 GROUP BY name;
Query ID = cloudera_20210916024242_84a2edde-a024-45aa-a3de-b7dcffd7c8bd
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 3
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_1631587610575_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1631587610575_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1631587610575_0003
Hadoop job information for Stage-1: number of mappers: 4; number of reducers: 3
2021-09-16 02:42:18,387 Stage-1 map = 0%, reduce = 0%
2021-09-16 02:42:28,096 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 1.99 sec
2021-09-16 02:42:30,212 Stage-1 map = 75%, reduce = 0%, Cumulative CPU 13.48 sec
2021-09-16 02:42:31,252 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 16.9 sec
2021-09-16 02:42:36,661 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 20.55 sec
2021-09-16 02:42:37,705 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 23.43 sec
2021-09-16 02:42:38,807 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 27.09 sec
MapReduce Total cumulative CPU time: 27 seconds 90 msec
Ended Job = job_1631587610575_0003
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1631587610575_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1631587610575_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1631587610575_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-09-16 02:42:45,593 Stage-2 map = 0%, reduce = 0%
2021-09-16 02:42:50,892 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.93 sec
2021-09-16 02:42:55,167 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.28 sec
MapReduce Total cumulative CPU time: 2 seconds 280 msec
cloudera@quickstart:~$ cloudera@quickstart:~$ cloudera@quickstart:~$
```

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
2021-09-16 02:42:38,807 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 27.09 sec
MapReduce Total cumulative CPU time: 27 seconds 90 msec
Ended Job = job_1631587610575_0003
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1631587610575_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1631587610575_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1631587610575_0004
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-09-16 02:42:45,593 Stage-2 map = 0%, reduce = 0%
2021-09-16 02:42:50,892 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 0.93 sec
2021-09-16 02:42:55,167 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.28 sec
MapReduce Total cumulative CPU time: 2 seconds 280 msec
Ended Job = job_1631587610575_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4 Reduce: 3 Cumulative CPU: 27.09 sec HDFS Read: 689530621 HDFS Write: 16439 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.28 sec HDFS Read: 22288 HDFS Write: 11732 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 370 msec
OK
Abilene Regional 15.013043478260869
Abraham Lincoln Capital 42.75912408759124
Adak Island Ns -4.222222222222222
Akron/canton Regional 10.701408450704225
Albany International 7.412162162162162
Albert J Ellis 6.441558441558442
Albuquerque International 4.7077826725403815
Alexandria International 15.923857868020304
Arcata 41.61363636363637
Asheville Regional 11.830645161290322
Aspen 26.5869918699187
Atlantic City International 6.103448275862069
Augusta Regional 5.768421052631579
Austin Straubel International 27.900990099009903
Austin-bergstrom International 4.82597523943004
```

mouse pointer inside or press Ctrl+G.

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Syracuse Hancock International 11.02579365079365
T. F. Green State 4.485552115583076
Tallahassee Regional 4.7011494252873565
Tampa International 3.5917417417417417
Ted Stevens Anchorage International 4.9961330239752515
Telluride Regional 11.378378378378379
Texarkana Regional-Webb Field 11.936363636363636
Toledo Express 14.0
Tri-Cities 5.935483870967742
Tri-Cities Regional 3.231707317073171
Tucson International 6.2253968253968255
Tulsa International 8.675390035228988
Tupelo Regional -1.8
Tyler Pounds Regional Airport 8.77304964539007
University Of Illinois Willard 47.457013574660635
University Park Airport 18.37037037037037
Valdosta Regional 4.405063291139241
Valley International 7.620865139949109
Ventura County Camarillo -1.7410714285714286
Waco Regional 8.270408163265307
Walker Field 9.220670391061452
Washington Dulles International 13.39019517442754
Waterloo Regional 15.1
Westchester County 15.238210399032647
Wichita Mid-Continent 11.836092715231787
Wiley Post/W.Rogers M 4.75
Wilkes-barre/scranton International 22.433673469387756
Will Rogers World Airport 7.79559748427673
William P Hobby 7.228734659331359
Wilmington International 3.4597156398104265
Wrangell SPB 14.982142857142858
Yakima Air Terminal 17.25
Yakutat 17.625
Yampa Valley 18.2925
Yeager 12.234309623430962
Yellowstone Regional -2.5824175824175826
Yuma MCAS/Yuma International 8.438172043010752
Time taken: 44.769 seconds, Fetched: 285 row(s)
hive>
```

Extra

```
hive> describe airports;
OK
name          string
country       string
area_code     int
code          string
Time taken: 0.104 seconds, Fetched: 4 row(s)
hive>
```

```
hive> describe flight_data;
OK
year                int
month               int
day                 int
day_of_week         int
dep_time            int
crs_dep_time        int
arr_time            int
crs_arr_time        int
unique_carrier      string
flight_num          int
tail_num            string
actual_elapsed_time int
crs_elapsed_time    int
air_time            int
arr_delay           int
dep_delay           int
origin              string
dest                string
distance            int
taxi_in             int
taxi_out            int
cancelled           int
cancellation_code   string
diverted            int
carrier_delay       string
weather_delay       string
nas_delay           string
security_delay      string
late_aircraft_delay string
Time taken: 0.147 seconds, Fetched: 29 row(s)
```