# MySQL

### **Table creation in MySQL database:**

CREATE TABLE flight\_input3(s\_no integer,
Year integer, Month LONGTEXT, DayofMonth integer,DayofWeek integer,
DepTime integer, CRSDepTime integer, ArrTime integer, CRSArrTime integer,
UniqueCarrier LONGTEXT, FlightNum integer, TailNum LONGTEXT,
ActualElapsedTime integer, CRSElapsedTime integer,
AirTime integer, ArrDelay integer,DepDelay integer, Origin LONGTEXT,
Dest LONGTEXT, Distance integer,TaxiIn integer, TaxiOut integer,
Cancelled integer, CancellationCode LONGTEXT);

#### Loading of data from local into MySQL table.

mysql -u hive -p --local\_infile=1 PDA -e "LOAD DATA LOCAL INFILE '/home/hduser/Downloads/DelayedFlights.csv' INTO TABLE flight\_input3 FIELDS TERMINATED BY ','";

```
mysql> show databases;
| Database
| information schema
 PDA
 amazon
 metastore
 mvsal
| performance schema
mysql> use PDA;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
Database changed
mysql> show tables;
| Tables in PDA |
| flight_input
| flight_input2
| flight_input3
3 rows in set (0.00 sec)
mysql> select count(*) from flight input3;
| count(*) |
  1044680 |
mysql>
```

# Hive

#### Table creation:

CREATE TABLE PDA.flight\_input3

(s\_no int, Year int, Month string, 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 string, Dest string, Distance int, TaxiIn int, axiOut int, Cancelled int, CancellationCode string)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (
"separatorChar" = ",", "quoteChar" = "\"")

stored as textfile
tblproperties ("skip.header.line.count"="1");

#### Loading data from MySQL into HDFS:

load data inpath '/flight\_input3/flight\_input3.csv' overwrite into table flight\_input3;

## Loading data from HDFS into hive using Sqoop:

sqoop import --connect jdbc:mysql://127.0.0.1/PDA --username hive --password admin --table flight\_input3 --target-dir /flight\_input3 -m 1;

```
hive> show tables;

OK
catable1
catable2
catable3
flight input
set hive
test_orc
fine taken: 0.153 seconds, Fetched: 8 row(s)
hive> select count(*) from flight_input3;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different
query ID = hduser_20190426214926_98419bb6-aad1-4226-a2fb-e41e3e633ce5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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 hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
    set hive.exec.reducers.max=<number>
Starting Job = job_i556311437514_0001, Tracking URL = http://PDA-VirtualBox:8088/proxy/application_1556311437514_0001/
Rill Command = /usr/local/hadoop/bin/hadoop job -kill job_1556311437514_0001
Radoop job information for stage-1: number of mappers: 1: number of reducers: 1
2019-04-26 21:501:37,708 Stage-1 map = 00%, reduce = 0%, cumulative CFU 12.08 sec
2019-04-26 21:501:37,708 Stage-1 map = 100%, reduce = 0%, cumulative CFU 14.6 sec
MapReduce Total cumulative CFU time: 14 seconds 600 msec
Ended Job = job_1556311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec
Ended Job = job_1556311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec

Ended Job = job_1556311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec

Ended Job = job_1556311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec

Ended Job = job_1556311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec

Ended Job = job_1566311437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec

Ended Job = job_156631437514_0001
MapReduce GPU Time Spent: 14 seconds 600 msec
```

#### **Hive Queries:**

# Query1: Flights that got cancelled maximum number of times during each month.

```
INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive_Q1'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
SELECT Month, flightnum, COUNT(cancelled) as t FROM flight_input3
WHERE cancelled= 1
GROUP BY Month, flightnum
ORDER BY t DESC
LIMIT 10;
```

```
hive> SELECT Month, flightnum, COUNT (cancelled) as t FROM flight_input3
> WHERE cancelled= 1
> GROUP BY Month, flightnum
> ORDER BY t DESC
> LIMIT 10;
```

```
Total MapReduce CPU Time Spent: 14 seconds 80 msec
OK
March
       511
               163
January 511
               127
February
               511
                       126
April 511
               124
               745
                    109
February
March 494
               97
               507
                       97
February
February
               351
                       94
May
       511
               91
March
       351
               90
Time taken: 94.522 seconds, Fetched: 10 row(s)
hive>
```

#### Query2: Flights that remained in air more time during each month.

SELECT Month, flightnum, SUM(AirTime) as d FROM flight\_input3 GROUP BY Month, flightnum
ORDER BY d DESC
LIMIT 10;

```
Total MapReduce CPU Time Spent: 15 seconds 890 msec
OK
February
                15
                        36622.0
January 73
                32027.0
March 73
May 15
               30565.0
                30274.0
February
                        29979.0
                73
January 15
               28429.0
March 267
                28036.0
March 15
May 73
March 1
               27811.0
                27487.0
                26820.0
Time taken: 73.654 seconds, Fetched: 10 row(s)
hive>
```

#### **Query 3: Top 20 most visited airports in Unites States**

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_Q3'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
SELECT dest,COUNT(dest) as p FROM flight\_input3
GROUP BY dest
ORDER BY p DESC
LIMIT 20;

```
hive> SELECT dest, COUNT (dest) as p FROM flight_input3
> GROUP BY dest
> ORDER BY p DESC
> LIMIT 20;
```

```
Total MapReduce CPU Time Spent: 11 seconds 800 msec
OK
ORD
        62367
ATL
       50426
DFW
        38176
DEN
       34154
LAX
       32500
EWR
       27865
LAS
       26640
PHX
       26270
SFO
       23556
IAH
       21985
DTW
       20942
LGA
       20739
MCO
       19166
MSP
       16529
SLC
       16259
       16188
BWI
PHL
       16163
BOS
       16141
JFK
       15789
SEA
       14880
Time taken: 69.289 seconds, Fetched: 20 row(s)
hive>
```

#### **Query 4: Top carriers in US airports**

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_Q4'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
SELECT UniqueCarrier, count(FlightNum) as d FROM flight\_input3
GROUP BY UniqueCarrier
ORDER BY d DESC
LIMIT 10;

```
hive> SELECT UniqueCarrier, count(FlightNum) as d FROM flight_input3
> GROUP BY UniqueCarrier
> ORDER BY d DESC
> LIMIT 10;
```

```
Total MapReduce CPU Time Spent: 11 seconds 780 msec
OK
WN
        213885
AA
        102651
MQ
        82135
UΑ
        81778
00
        73388
ΧE
        62246
US
        53684
DL
        48692
NW
        48265
CO
        44152
Time taken: 71.432 seconds, Fetched: 10 row(s)
```



# **MapReduce**

BI Query1: Average delay time caused by airline carriers.

Input- Extracted from MySQL database through MapR by programmatically accessing MySQL.

## Output - MapR output generated in Hbase through MapR program.

```
hbase(main):008:0> scan 'flight'
ROW
                                                                  COLUMN+CELL
                                                                  column=airline:delay, timestamp=1556217629944, value=@?\x1F6g\xFF\xD2]
                                                                  column=airline:delay, timestamp=1556217629944, value=@?\xD3\xB5\xDD\xE9\x055
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                                                                                                 value=@%$\x89:\x16\x0B\xA4
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                                                                                                 value=08X\x0F\x0E\xD1\x0Ai
                                                                                                                                                Value=@8X\X0F\X0E\XD1\X0A1

value=@C\XD8V\X1D\X00C~

value=@1\XAB\X83$m\X10A

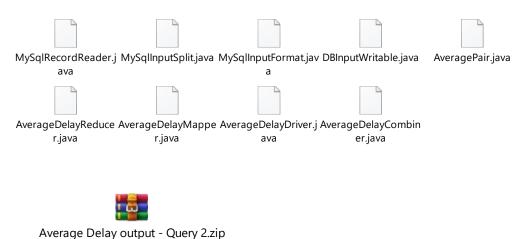
value=@7\XD00\XAF\X18\X98\XD1

value=@?\X91\XB13\X12\XDA\XA6

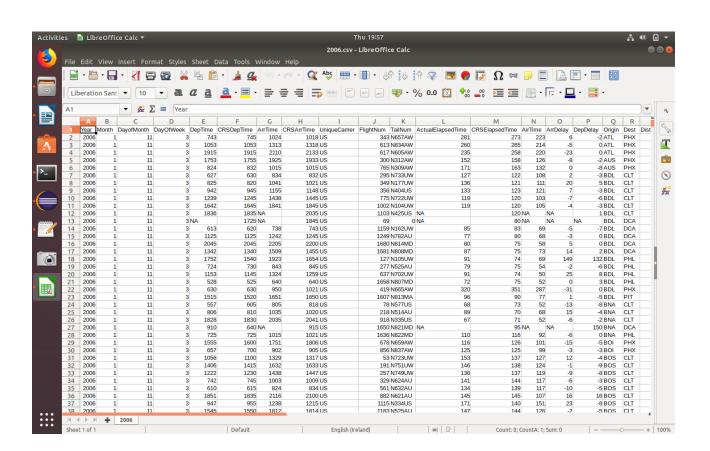
value=@1}\XC3A\XF0*\XBF

value=@<\XD9\X9E\X8B\XEFM%
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                  column=airline:delay,
                                                                                                      timestamp=1556217629944,
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                 column=airline:delay, timestamp=1556217629944, column=airline:delay, timestamp=1556217629944,
  "EV'
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                                                                                                 value=@0y\xD7\x91e\xE2\x0A
                                                                                                                                                 value=@>/\xFA\x8E)\xEA\xC9
value=@:\xFF\x08\x8Cv\xB1\xC1
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                  column=airline:delay, timestamp=1556217629944,
                                                                  column=airline:delay,
  "OH"
                                                                                                       timestamp=1556217629944,
                                                                                                                                                 value=04\xAC0\x86\xA9\x17l
                                                                 column=airline:delay, timestamp=1556217629944, Value=@4\XACU\X80\X84\X17\Column=airline:delay, timestamp=1556217629944, Value=@4\X80f8a\X83\X8F\Column=airline:delay, timestamp=1556217629944, Value=@0\X89\X16V[\x13\x80\column=airline:delay, timestamp=1556217629944, Value=@7\x11E\xFF\x94+:
column=airline:delay, timestamp=1556217629944, Value=@2\XA6>\X0F\XD1\x8E>
column=airline:delay, timestamp=1556217629944, Value=@0\xFA\XE2\x10V\x04\xEB\column=airline:delay, timestamp=1556217629944, Value=@7b\xE8\xAA9\xAC\xAD
  "00"
  "UA'
  "US
  "WN"
  "XE"
20 row(s) in 0.3790 seconds
```

#### **Java Files**

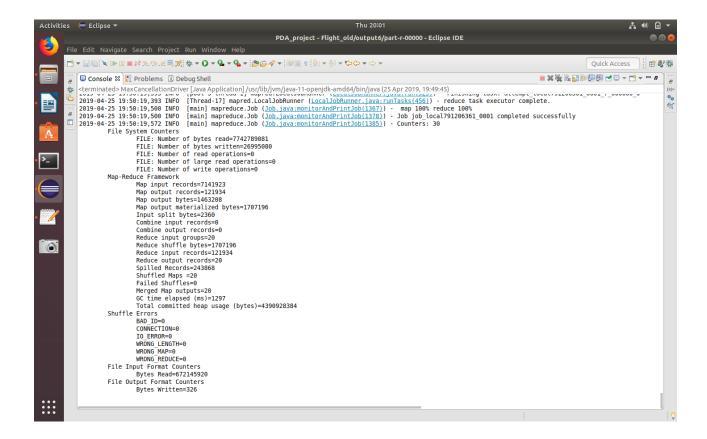


BI Query2: Maximum number of flights that got cancelled in each route.



## Output:

```
OGGtoDFW
                    261
 1 AA
 2 AQ
      LIHtoHNL
                    146
 3 AS
       SEAtoLAX
                    104
 4 B6
       BOStoJFK
                    56
 5 CO
      AUStoIAH
                    93
 6 DL
      ATLtoCHS
                    252
 7 EV
      ATLtoCRW
                    111
 8 F9
                    30
      DENtoSNA
 9 FL
      MLItoATL
                    67
10 HA
      HNLtoLIH
                    45
11 MQ
      DFWtoXNA
                    207
12 NW
      MSPto0RD
                    138
13 OH
      RDUtoLGA
                    153
14 00
      DENtoMTJ
                    165
15 TZ
                    52
      LAXto0GG
16 UA
      ORDtoPHL
                    278
17 US
                    179
       BOSTOPHL
18 WN
      LAXto0AK
                    457
19 XE
       ACKtoEWR
                    165
20 YV
                    140
      KOAtoHNL
```



### **Java Files**



MaxCancellationRedu MaxCancellationMap MaxCancellationDrive cer.java per.java r.java



Maximum Cancellation output - Query1.zip