BIG DATA AND HADOOP

Proof of Concept

Abstract

This document is a proof of concept for Pig, Hive, HBase and Sqoop learnt during Big Data and Hadoop training at Edureka.

Abhishek

abh1sh3k@yahoo.com

TABLE OF CONTENTS

1. INTRODUCTION

This project is being done as "Proof of Concept" to showcase the skills learnt during the 5 week training program of **Big Data and Hadoop** development. This POC will demonstrate the analysis done on different datasets using Pig and Hive along with HBase and Sqoop.

2. PIG

2.1 DATASET

Source: https://data.cityofnewyork.us/Transportation/Parking-Tickets/yyiw-ypks

A listing of all NYC parking tickets sorted by summons number, license plate, issue date, violation, fine amount, and other categories. This data is provided by Department of Finance (DOF). This is a Microsoft Access file (mdb) named **"Parking Ticktes.mdb"** and contains data for March 2010.

File Size: 37.7MB (378MB after unzip)
Number of records: 872,370

2.2 PROBLEM STATEMENT

Use Pig script to analyze the parking tickets dataset to find out the top 10 states with most fine samount collected for the month of March in 2010.

2.3 APPROACH

- Export Microsoft Access data to MySQL
- Export data from MySQL to HDFS using Sqoop
- Group the data by state and find total amount for each state
- Sort the amount in descending order and limit it to 10

2.4 CODE

Code: FineAmount.pig

```
-- Load summon records
tickets = load '/tickets/part-m-00000' using PigStorage(',') AS (summon_number:chararray,
plate:chararray, state:chararray, ws_type:chararray, issue_date:chararray,
violation_desc:chararray, violation_code:chararray, fine_amount:long, county:chararray,
front opposite:chararray, house no:chararray, street name:chararray,
intersect street name:chararray, street code1:chararray, street code2:chararray,
street code3:chararray);

-- Group summons by state
ticketsbystate = group tickets by state;

-- Sum total fine amount from each state
finebystate = foreach ticketsbystate generate group, SUM(tickets.fine amount);

-- Order the records from highest fine
topfineamount = order finebystate by $1 desc;

-- Select top 10 state by fine
top10states = limit topfineamount 10;
```

```
-- Store the result in hdfs
store top10states into '/ticketssoutput';
```

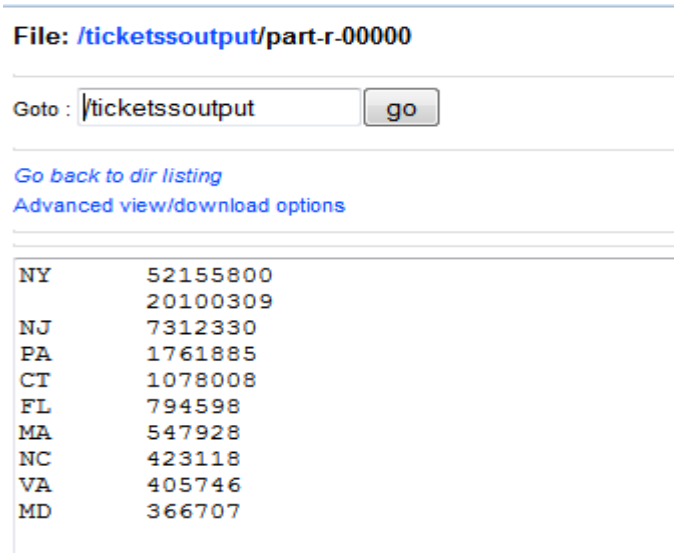
2.5 EXECUTION

```
cloudera@cloudera-vm:~$pig FineAmount.pig
2014-10-08 21:11:41,879 [main] INFO  org.apache.pig.Main - Logging error messages to:
/home/cloudera/pig_1412782901876.log
2014-10-08 21:11:42,345 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
at: hdfs://localhost:8020
2014-10-08 21:11:42,810 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job
tracker at: localhost:8021
2014-10-08 21:11:43,469 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
used in the script: GROUP_BY,ORDER_BY,LIMIT
2014-10-08 21:11:43,469 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - pig.usenewlogicalplan is set to
true. New logical plan will be used.
2014-10-08 21:11:43,988 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - (Name: topfineamount:
Store(/ticketssoutput:org.apache.pig.builtin.PigStorage) - scope-63 Operator Key: scope-63)
2014-10-08 21:11:44,023 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation
threshold: 100 optimistic? false
2014-10-08 21:11:44,107 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move
algebraic foreach to combiner
2014-10-08 21:11:44,181 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
before optimization: 3
2014-10-08 21:11:44,182 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size
after optimization: 3
2014-10-08 21:11:44,352 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script
settings are added to the job
2014-10-08 21:11:44,387 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2014-10-08 21:11:49,362 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up
single store job
2014-10-08 21:11:49,463 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=181452966
2014-10-08 21:11:49,463 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Neither
PARALLEL nor default parallelism is set for this job. Setting number of reducers to 1
2014-10-08 21:11:49,576 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce
job(s) waiting for submission.
2014-10-08 21:11:50,077 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2014-10-08 21:11:50,189 [Thread-4] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
Total input paths to process : 1
2014-10-08 21:11:50,190 [Thread-4] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2014-10-08 21:11:50,216 [Thread-4] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to
process : 3
2014-10-08 21:11:51,154 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId:
job_201410081944_0002
2014-10-08 21:11:51,154 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information
at: http://localhost:50030/jobdetails.jsp?jobid=job_201410081944_0002
```

```
2014-10-08 21:12:07,282 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1% complete
2014-10-08 21:12:10,296 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 2% complete
2014-10-08 21:12:13,308 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 4% complete
2014-10-08 21:12:16,340 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 6% complete
2014-10-08 21:12:19,352 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 7% complete
2014-10-08 21:12:22,363 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 9% complete
2014-10-08 21:12:25,377 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 10% complete
2014-10-08 21:12:39,464 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 12% complete
2014-10-08 21:12:45,497 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 13% complete
2014-10-08 21:12:48,509 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 15% complete
2014-10-08 21:12:49,011 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 19% complete
2014-10-08 21:12:56,571 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 33% complete
2014-10-08 21:13:01,273 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script
settings are added to the job
2014-10-08 21:13:01,275 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2014-10-08 21:13:05,673 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up
single store job
2014-10-08 21:13:05,703 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce
job(s) waiting for submission.
2014-10-08 21:13:05,851 [Thread-14] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
Total input paths to process : 1
2014-10-08 21:13:05,851 [Thread-14] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2014-10-08 21:13:05,853 [Thread-14] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to
process : 1
2014-10-08 21:13:06,204 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId:
job_201410081944_0003
2014-10-08 21:13:06,204 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information
at: http://localhost:50030/jobdetails.jsp?jobid=job_201410081944_0003
2014-10-08 21:13:11,239 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2014-10-08 21:13:23,334 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 66% complete
2014-10-08 21:13:26,376 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script
settings are added to the job
2014-10-08 21:13:26,376 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2014-10-08 21:13:31,584 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up
single store job
2014-10-08 21:13:31,650 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce
job(s) waiting for submission.
2014-10-08 21:13:31,823 [Thread-25] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
Total input paths to process : 1
2014-10-08 21:13:31,823 [Thread-25] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2014-10-08 21:13:31,825 [Thread-25] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to
process : 1
```

```
2014-10-08 21:13:32,152 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId:
job_201410081944_0004
2014-10-08 21:13:32,152 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - More information
at: http://localhost:50030/jobdetails.jsp?jobid=job_201410081944_0004
2014-10-08 21:13:37,686 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 83% complete
2014-10-08 21:13:52,303 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2014-10-08 21:13:52,311 [main] INFO  org.apache.pig.tools.pigstats.PigStats - Script Statistics:

HadoopVersion  PigVersion     UserId  StartedAt      FinishedAt     Features
0.20.2-cdh3u0  0.8.0-cdh3u0   cloudera       2014-10-08 21:11:44    2014-10-08 21:13:52
       GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime     MinMapTIme     AvgMapTime     MaxReduceTime  MinReduceTime
       AvgReduceTime  Alias   Feature Outputs
job_201410081944_0002  3       1       32      23      29      29      29      29
       finebystate,tickets,ticketsbystate    GROUP_BY,COMBINER
job_201410081944_0003  1       1       2       2       2       12      12      12      topfineamount
       SAMPLER
job_201410081944_0004  1       1       2       2       2       10      10      10      topfineamount
       ORDER_BY,COMBINER      /ticketssoutput,

Input(s):
Successfully read 872370 records (181462225 bytes) from: "/tickets/part-m-00000"

Output(s):
Successfully stored 10 records (105 bytes) in: "/ticketssoutput"

Counters:
Total records written : 10
Total bytes written : 105
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201410081944_0002  ->      job_201410081944_0003,
job_201410081944_0003  ->      job_201410081944_0004,
job_201410081944_0004


2014-10-08 21:13:52,356 [main] WARN
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered
Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 28 time(s).
2014-10-08 21:13:52,356 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

2.6 RESULT



3. HIVE

3.1 DATASET

Source: http://openflights.org/data.html

It has 2 datasets available as below:

Airlines: As of January 2012, the OpenFlights Airlines Database contains **5888** airlines. This is a data file (dat) named **"airlines.dat"** and contains the information like airline name, country etc.
File Size: 381KB
Number of records: 6048

Routes: OpenFlights/Airline Route Mapper Route Database contains **59036** routes between **3209** airports on **531** airlines spanning the globe. This is a data file (dat) named **"routes.dat"** and contains information like airline name, source airport, destination airport etc.

File Size: 2.26MB
Number of records: 67,663

Airports: The OpenFlights Airports Database contains **6977** airports spanning the globe. This is a data file (dat) named ***"airports.dat"*** and contains the information like airport Id, Name etc.

File Size: 831KB
Number of records: 8107

3.2 PROBLEM STATEMENT

Use HiveQL to analyze the mentioned datasets and find out below –

- Total flights run by each airlines from India which are active.
- Top 10 busiest source airport in India.

3.3 APPROACH

- Remove quotes (“”) and \N characters from the datasets using standard linux command
- Create a database and tables to load the datasets
- Find total number of flights with airline names from routes dataset by joining airlines dataset and filtering by country and its status as active
- Count total number of entries of source airports, filter it by country and limit to 10 based on counts

3.4 CODE

Code: airinfo.sql

```
-- Create database
CREATE DATABASE airinformation;

USE airinformation;

-- Create airlines table to store airline data
CREATE TABLE airlines (AirlineId STRING, Name STRING, Alias STRING, IATA STRING, ICAO STRING,
Callsign STRING, Country STRING, Active STRING)
ROW FORMAT
DELIMITED FIELDS
TERMINATED BY ','
STORED AS textfile;

-- Load airline data in airlines table
LOAD DATA LOCAL INPATH '/home/cloudera/airlines.dat' OVERWRITE INTO TABLE airlines;

--Create routes table to store data of airline routes information
CREATE TABLE routes (Airline STRING, AirlineId STRING, SourceAirport STRING, SourceAirportId
STRING, DestinationAirport STRING, DestinationAirportId STRING, Codeshare STRING, Stops STRING,
Equipment STRING)
ROW FORMAT
DELIMITED FIELDS
TERMINATED BY ','
STORED AS textfile;

-- Load routes data in routes table
LOAD DATA LOCAL INPATH '/home/cloudera/routes.dat' OVERWRITE INTO TABLE routes;

-- Create airports table to store airport data
CREATE TABLE airports (AirportId STRING, Name STRING, City STRING, Country STRING, IATA STRING,
ICAO STRING, Latitude STRING, Longitude STRING, Altitude STRING, Timezone STRING, DST STRING)
ROW FORMAT
DELIMITED FIELDS
TERMINATED BY ','
STORED AS textfile;

-- Load airport data to airports table
LOAD DATA LOCAL INPATH '/home/cloudera/airports.dat' OVERWRITE INTO TABLE airports;

-- Create table to store the result of total number of flights of each airline
CREATE TABLE airlinecount (AirlineName STRING, TotalFlights INT);

-- Load the result in airlinecount table
```

```
INSERT OVERWRITE TABLE airlinecount
SELECT airline data.Name, route data.cou
FROM
(
SELECT Name, AirlineId FROM airlines WHERE Country='India' AND Active='Y'
)
airline_data
JOIN (
SELECT AirlineId, COUNT(AirlineId) AS cou FROM routes GROUP BY AirlineId
)
route_data
ON route_data.AirlineId = airline_data.AirlineId;

-- Create table to store result of top 10 busiest airport
CREATE TABLE airportcount (AirportName STRING, TotalAirports INT);

-- Load result of top 10 busiest airports in airportcount table
INSERT OVERWRITE TABLE airportcount
SELECT a.SourceAirport, COUNT(a.SourceAirport) AS Total FROM routes a JOIN airlines b ON
a.AirlineId = b.AirlineId WHERE b.Country='India' GROUP BY a.SourceAirport ORDER BY Total DESC
LIMIT 10;
```

3.5 EXECUTION

```
cloudera@cloudera-vm:~$ sed -i 's/\\N//g' airlines.dat
cloudera@cloudera-vm:~$ sed -i 's/\\N//g' airports.dat
cloudera@cloudera-vm:~$ sed -i 's/"//g' airlines.dat
cloudera@cloudera-vm:~$ sed -i 's/"//g' airports.dat
cloudera@cloudera-vm:~$sudo hive -f airinfo.sql
Hive history file=/tmp/root/hive_job_log_root_201410082319_337787744.txt
OK
Time taken: 4.113 seconds
OK
Time taken: 0.013 seconds
OK
Time taken: 0.588 seconds
Copying data from file:/home/cloudera/airlines.dat
Copying file: file:/home/cloudera/airlines.dat
Loading data to table airinformation.airlines
Deleted hdfs://localhost/user/hive/warehouse/airinformation.db/airlines
OK
Time taken: 1.167 seconds
OK
Time taken: 0.061 seconds
Copying data from file:/home/cloudera/routes.dat
Copying file: file:/home/cloudera/routes.dat
Loading data to table airinformation.routes
Deleted hdfs://localhost/user/hive/warehouse/airinformation.db/routes
OK
Time taken: 0.831 seconds
OK
Time taken: 0.069 seconds
Copying data from file:/home/cloudera/airports.dat
Copying file: file:/home/cloudera/airports.dat
Loading data to table airinformation.airports
Deleted hdfs://localhost/user/hive/warehouse/airinformation.db/airports
OK
Time taken: 0.312 seconds
OK
Time taken: 0.051 seconds
Total MapReduce jobs = 2
Launching Job 1 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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0001, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0001
2014-10-08 23:19:47,811 Stage-1 map = 0%,  reduce = 0%
2014-10-08 23:19:58,011 Stage-1 map = 100%,  reduce = 0%
2014-10-08 23:20:10,139 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0001
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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0002, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0002
2014-10-08 23:20:18,188 Stage-2 map = 0%,  reduce = 0%
2014-10-08 23:20:24,257 Stage-2 map = 50%,  reduce = 0%
2014-10-08 23:20:25,273 Stage-2 map = 100%,  reduce = 0%
2014-10-08 23:20:38,365 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0002
Loading data to table airinformation.airlinecount
Deleted hdfs://localhost/user/hive/warehouse/airinformation.db/airlinecount
Table airinformation.airlinecount stats: [num_partitions: 0, num_files: 1, num_rows: 0,
total_size: 117]
OK
Time taken: 68.147 seconds
OK
Time taken: 0.094 seconds
Total MapReduce jobs = 4
Launching Job 1 out of 4
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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0003, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0003
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0003
2014-10-08 23:20:48,936 Stage-1 map = 0%,  reduce = 0%
2014-10-08 23:20:54,997 Stage-1 map = 50%,  reduce = 0%
2014-10-08 23:20:56,003 Stage-1 map = 100%,  reduce = 0%
2014-10-08 23:21:06,146 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0003
Launching Job 2 out of 4
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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0004, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0004
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0004
2014-10-08 23:21:14,973 Stage-2 map = 0%,  reduce = 0%
2014-10-08 23:21:19,342 Stage-2 map = 50%,  reduce = 0%
2014-10-08 23:21:20,358 Stage-2 map = 100%,  reduce = 0%
2014-10-08 23:21:29,414 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0004
```

```
Launching Job 3 out of 4
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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0005, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0005
2014-10-08 23:21:38,600 Stage-3 map = 0%,  reduce = 0%
2014-10-08 23:21:43,637 Stage-3 map = 100%,  reduce = 0%
2014-10-08 23:21:53,779 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0005
Launching Job 4 out of 4
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 mapred.reduce.tasks=<number>
Starting Job = job_201410082157_0006, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410082157_0006
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill
job_201410082157_0006
2014-10-08 23:22:01,748 Stage-4 map = 0%,  reduce = 0%
2014-10-08 23:22:03,775 Stage-4 map = 100%,  reduce = 0%
2014-10-08 23:22:18,857 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_201410082157_0006
Loading data to table airinformation.airportcount
Deleted hdfs://localhost/user/hive/warehouse/airinformation.db/airportcount
Table airinformation.airportcount stats: [num_partitions: 0, num_files: 1, num_rows: 0,
total_size: 72]
10 Rows loaded to airportcount
OK
Time taken: 99.102 seconds
```

3.6 RESULT

Result of total number of flights with active airline names from India:

File: /user/hive/warehouse/airinformation.db/airlinecount/000000_0

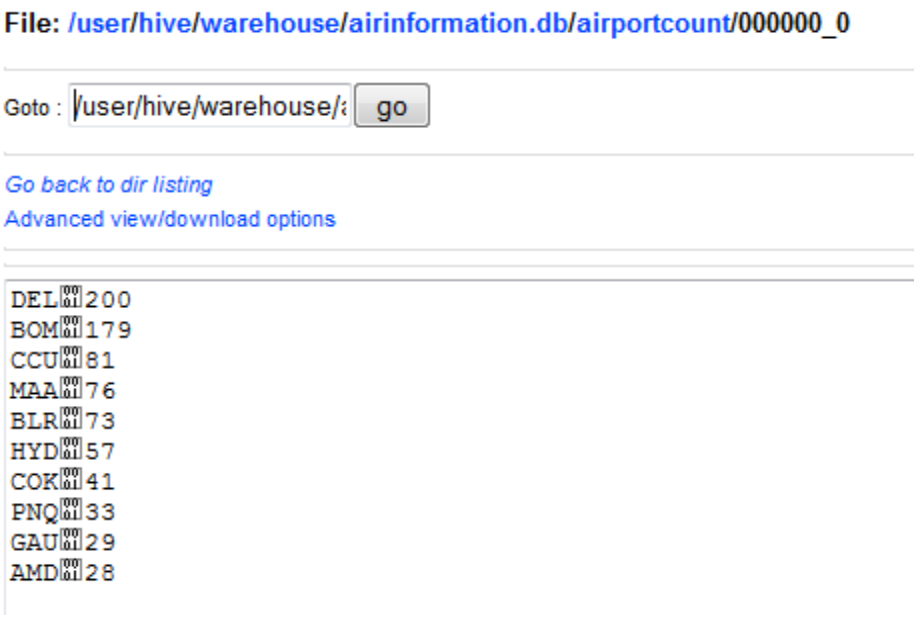
Goto : /user/hive/warehouse/a go

Go back to dir listing

Advanced view/download options

```
Air India Limited393
Air Sahara180
Go Air77
IndiGo Airlines227
Jet Airways292
Spicejet196
Air India Express76
```

Result of top 10 busiest source airports:



4. HBASE

4.1 DATASET

Source: https://data.cityofnewyork.us/Business/Times-Square-Entertainment-Venues/jxdc-hnze

Directory of entertainment venues in the Times Square area.

File Size: 6KB
Number of records: 78

4.2 PROBLEM STATEMENT

Load the records based on sub-industry 'Comedy Club' to HBase using Hive.

4.3 APPROACH

- Create a table in Hive called *clubs* and load the dataset
- Create a table *comedyclub* in Hive to map with HBase
- Load data from *clubs to comedyclub* by filtering clubs as 'Comedy Club'

4.4 CODE

Code:

```
-- Create table to load dataset
```

```
CREATE TABLE clubs (CompanyName STRING, SubIndustry STRING, SubSubIndustry STRING, Phone STRING,
Website STRING, Location1 STRING)
ROW FORMAT
DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS textfile;

-- Load dataset in clubs table
LOAD DATA LOCAL INPATH '/home/cloudera/Times_Square_Entertainment_Venues.csv' OVERWRITE INTO
TABLE clubs;

-- Create a table which will be mapped to table in HBase
CREATE TABLE comedyclub (CompanyName STRING, SubIndustry STRING)
STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,info:c1")
TBLPROPERTIES ("hbase.table.name" = "clubinformation");

-- Load data from hive to HBase table
INSERT OVERWRITE TABLE comedyclub SELECT CompanyName, SubIndustry FROM clubs where
SubIndustry='Comedy Club';

-- Verify the data
scan "clubinfo"
```

4.5 EXECUTION

```
hive> CREATE TABLE clubs (CompanyName STRING, SubIndustry STRING, SubSubIndustry
STRING, Phone STRING, Website STRING, Location1 STRING)
    > ROW FORMAT
    > DELIMITED FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS textfile;
OK
Time taken: 9.776 seconds
hive> LOAD DATA LOCAL INPATH '/home/cloudera/Times_Square_Entertainment_Venues.csv'
OVERWRITE INTO TABLE clubs;
Copying data from file:/home/cloudera/Times_Square_Entertainment_Venues.csv
Copying file: file:/home/cloudera/Times_Square_Entertainment_Venues.csv
Loading data to table default.clubs
Deleted hdfs://localhost/user/hive/warehouse/clubs
OK
Time taken: 1.246 seconds
hive> CREATE TABLE comedyclub (CompanyName STRING, SubIndustry STRING)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,info:c1")
    > TBLPROPERTIES ("hbase.table.name" = "clubinformation");
OK
Time taken: 0.468 seconds
hive> INSERT OVERWRITE TABLE comedyclub SELECT CompanyName, SubIndustry FROM clubs
where SubIndustry='Comedy Club';
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201410091851_0007, Tracking URL =
http://localhost:50030/jobdetails.jsp?jobid=job_201410091851_0007
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -
kill job_201410091851_0007
2014-10-09 22:44:21,754 Stage-0 map = 0%,  reduce = 0%
2014-10-09 22:44:36,052 Stage-0 map = 50%,  reduce = 0%
2014-10-09 22:44:37,119 Stage-0 map = 100%,  reduce = 0%
2014-10-09 22:44:40,164 Stage-0 map = 100%,  reduce = 100%
```

```
Ended Job = job_201410091851_0007
4 Rows loaded to comedyclub
OK
Time taken: 40.09 seconds
hive> select * from comedyclub;
OK
Carolines on Broadway      Comedy Club
Ha! Comedy Club      Comedy Club
The World Famous Laugh Factory    Comedy Club
Times Square Art Center    Comedy Club
Time taken: 0.944 seconds
```

4.6 RESULT

```
cloudera@cloudera-vm:~$ hbase shell
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.90.1-cdh3u0, r, Fri Mar 25 16:10:51 PDT 2011

hbase(main):001:0> scan "clubinformation"
ROW                                    COLUMN+CELL
 Carolines on Broadway                 column=info:c1, timestamp=1412874875912, value=Comedy Club
 Ha! Comedy Club                       column=info:c1, timestamp=1412874875494, value=Comedy Club
 The World Famous Laugh Factory        column=info:c1, timestamp=1412874875494, value=Comedy Club
 Times Square Art Center               column=info:c1, timestamp=1412874875912, value=Comedy Club
4 row(s) in 0.9080 seconds

hbase(main):002:0> exit
```

5. SQOOP

5.1 DATASET

- Use PIG dataset used to export the data to HDFS.
- Use HIVE solution dataset to import the data from HDFS.

5.2 PROBLEM STATEMENT

- Using Sqoop import the dataset of Pig i.e. **"Parking Tickets.mdb**" from RDBMS table to HDFS.
- Export the airline record created in Hive solution from Hive to RDBMS table.

5.3 APPROACH

Import:

- Create a database and table in MySQL
- Export the data from MSAccess to MySQL using MS Access export wizard
- Run sqoop import to import the data from MySQL to HDFS

Export:

- Create a database and table in MySQL
- Run sqoop export to export the data from Hive to MySQL

5.4 CODE

Code: Import data from MySQL to HDFS

```
bin/sqoop import --connect jdbc:mysql://192.168.124.1/parkingtickets --table summons --username root -P --target-dir /tickets -m 1
```

Code: Export data from Hive to MySQL

```
bin/sqoop export --connect jdbc:mysql://192.168.124.1/airinformation --table airlinecount -m 1 --username root -P --export-dir /user/hive/warehouse/airinformation.db/airlinecount/000000_0 --input-fields-terminated-by '\001'
```

5.5 EXECUTION

Import Execution:

```
cloudera@cloudera-vm:/usr/lib/sqoop$ bin/sqoop import --connect
jdbc:mysql://192.168.124.1/parkingtickets --table summons --username root -P --target-dir
/tickets -m 1
Enter password:
14/10/08 19:58:59 INFO tool.CodeGenTool: Beginning code generation
14/10/08 19:59:00 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM `summons`
AS t LIMIT 1
14/10/08 19:59:00 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM `summons`
AS t LIMIT 1
14/10/08 19:59:00 INFO orm.CompilationManager: HADOOP_HOME is /usr/lib/hadoop
14/10/08 19:59:00 INFO orm.CompilationManager: Found hadoop core jar at: /usr/lib/hadoop/hadoop-
core.jar
Note: /tmp/sqoop-cloudera/compile/60949d9b991fa4e344de649d028e3cf8/summons.java uses or overrides
a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
14/10/08 19:59:02 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-
cloudera/compile/60949d9b991fa4e344de649d028e3cf8/summons.java to /usr/lib/sqoop/./summons.java
14/10/08 19:59:02 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-
cloudera/compile/60949d9b991fa4e344de649d028e3cf8/summons.jar
14/10/08 19:59:02 WARN manager.MySQLManager: It looks like you are importing from mysql.
14/10/08 19:59:02 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
14/10/08 19:59:02 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
14/10/08 19:59:02 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull
(mysql)
14/10/08 19:59:02 INFO mapreduce.ImportJobBase: Beginning import of summons
14/10/08 19:59:02 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM `summons`
AS t LIMIT 1
14/10/08 19:59:04 INFO mapred.JobClient: Running job: job_201410081944_0001
14/10/08 19:59:05 INFO mapred.JobClient:  map 0% reduce 0%
14/10/08 19:59:16 INFO mapred.JobClient:  map 100% reduce 0%
14/10/08 20:00:13 INFO mapred.JobClient: Job complete: job_201410081944_0001
14/10/08 20:00:13 INFO mapred.JobClient: Counters: 12
14/10/08 20:00:13 INFO mapred.JobClient:   Job Counters
14/10/08 20:00:13 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=66209
14/10/08 20:00:13 INFO mapred.JobClient:     Total time spent by all reduces waiting after
reserving slots (ms)=0
14/10/08 20:00:13 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving
slots (ms)=0
14/10/08 20:00:13 INFO mapred.JobClient:     Launched map tasks=1
14/10/08 20:00:13 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
14/10/08 20:00:13 INFO mapred.JobClient:   FileSystemCounters
14/10/08 20:00:13 INFO mapred.JobClient:     HDFS_BYTES_READ=87
14/10/08 20:00:13 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=59568
14/10/08 20:00:13 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=181452966
14/10/08 20:00:13 INFO mapred.JobClient:   Map-Reduce Framework
14/10/08 20:00:13 INFO mapred.JobClient:     Map input records=872370
14/10/08 20:00:13 INFO mapred.JobClient:     Spilled Records=0
```

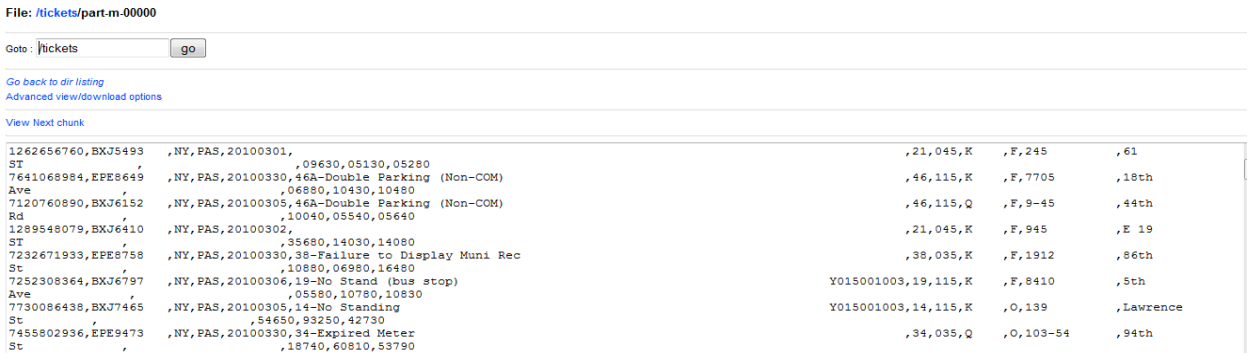
```
14/10/08 20:00:13 INFO mapred.JobClient:     Map output records=872370
14/10/08 20:00:13 INFO mapred.JobClient:     SPLIT_RAW_BYTES=87
14/10/08 20:00:13 INFO mapreduce.ImportJobBase: Transferred 173.047 MB in 69.9872 seconds (2.4726
MB/sec)
14/10/08 20:00:13 INFO mapreduce.ImportJobBase: Retrieved 872370 records.
```

Export Execution:

```
cloudera@cloudera-vm:/usr/lib/sqoop$ hadoop dfs -ls
/user/hive/warehouse/airinformation.db/airportcount/000000_0
bin/sqoop export --connect jdbc:mysql://192.168.124.1/airinformation --table airlinecount -m 1 --
username root -P --export-dir /user/hive/warehouse/airinfo
ormation.db/airlinecount/000000_0 --input-fields-terminated-by '\001'
Enter password:
14/10/09 19:22:43 INFO tool.CodeGenTool: Beginning code generation
14/10/09 19:22:44 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM
`airlinecount` AS t LIMIT 1
14/10/09 19:22:44 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM
`airlinecount` AS t LIMIT 1
14/10/09 19:22:44 INFO orm.CompilationManager: HADOOP_HOME is /usr/lib/hadoop
14/10/09 19:22:44 INFO orm.CompilationManager: Found hadoop core jar at: /usr/lib/hadoop/hadoop-
core.jar
Note: /tmp/sqoop-cloudera/compile/0211d011400533df75514e98ed057f73/airlinecount.java uses or
overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
14/10/09 19:22:45 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-
cloudera/compile/0211d011400533df75514e98ed057f73/airlinecount.java to
/usr/lib/sqoop/./airlinecount.java
14/10/09 19:22:45 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-
cloudera/compile/0211d011400533df75514e98ed057f73/airlinecount.jar
14/10/09 19:22:45 INFO mapreduce.ExportJobBase: Beginning export of airlinecount
14/10/09 19:22:46 INFO manager.MySQLManager: Executing SQL statement: SELECT t.* FROM
`airlinecount` AS t LIMIT 1
14/10/09 19:22:47 INFO input.FileInputFormat: Total input paths to process : 1
14/10/09 19:22:47 INFO input.FileInputFormat: Total input paths to process : 1
14/10/09 19:22:47 INFO mapred.JobClient: Running job: job_201410091851_0004
14/10/09 19:22:48 INFO mapred.JobClient:  map 0% reduce 0%
14/10/09 19:22:55 INFO mapred.JobClient:  map 100% reduce 0%
14/10/09 19:22:56 INFO mapred.JobClient: Job complete: job_201410091851_0004
14/10/09 19:22:56 INFO mapred.JobClient: Counters: 12
14/10/09 19:22:56 INFO mapred.JobClient:   Job Counters
14/10/09 19:22:56 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=7324
14/10/09 19:22:56 INFO mapred.JobClient:     Total time spent by all reduces waiting after
reserving slots (ms)=0
14/10/09 19:22:56 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving
slots (ms)=0
14/10/09 19:22:56 INFO mapred.JobClient:     Launched map tasks=1
14/10/09 19:22:56 INFO mapred.JobClient:     Data-local map tasks=1
14/10/09 19:22:56 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
14/10/09 19:22:56 INFO mapred.JobClient:   FileSystemCounters
14/10/09 19:22:56 INFO mapred.JobClient:     HDFS_BYTES_READ=275
14/10/09 19:22:56 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=59405
14/10/09 19:22:56 INFO mapred.JobClient:   Map-Reduce Framework
14/10/09 19:22:56 INFO mapred.JobClient:     Map input records=7
14/10/09 19:22:56 INFO mapred.JobClient:     Spilled Records=0
14/10/09 19:22:56 INFO mapred.JobClient:     Map output records=7
14/10/09 19:22:56 INFO mapred.JobClient:     SPLIT_RAW_BYTES=152
14/10/09 19:22:56 INFO mapreduce.ExportJobBase: Transferred 275 bytes in 10.1044 seconds (27.2157
bytes/sec)
14/10/09 19:22:56 INFO mapreduce.ExportJobBase: Exported 7 records.
```

5.6 RESULT

Below screenshot shows the imported data from MySQL in HDFS:



File: /tickets/part-m-00000

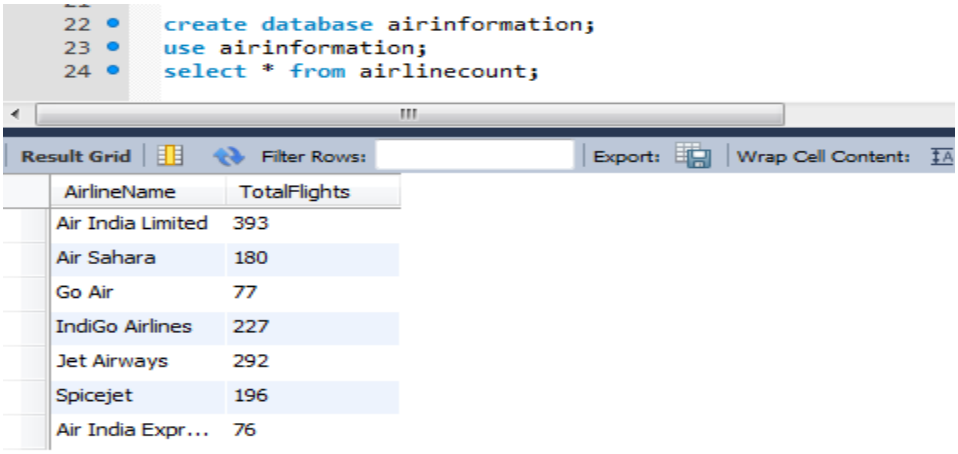
Goto : /tickets go

Go back to dir listing
Advanced view/download options

View Next chunk

```
1262656760,BXJ5493   ,NY,PAS,20100301,                                                    ,21,045,K    ,F,245         ,61
ST             ,                  ,09630,05130,05280
7641068984,EPE8649   ,NY,PAS,20100330,46A-Double Parking (Non-COM)                       ,46,115,K    ,F,7705        ,18th
Ave            ,                  ,06880,10430,10480
7120760890,BXJ6152   ,NY,PAS,20100305,46A-Double Parking (Non-COM)                       ,46,115,Q    ,F,9-45        ,44th
Rd             ,                  ,10040,05540,05640
1289548079,BXJ6410   ,NY,PAS,20100302,                                                    ,21,045,K    ,F,945         ,E 19
ST             ,                  ,35680,14030,14080
7232671933,EPE8758   ,NY,PAS,20100330,38-Failure to Display Muni Rec                     ,38,035,K    ,F,1912        ,86th
St             ,                  ,10880,06980,16480
7252308364,BXJ6797   ,NY,PAS,20100306,19-No Stand (bus stop)                  Y015001003,19,115,K    ,F,8410        ,5th
Ave            ,                  ,05580,10780,10830
7730086438,BXJ7465   ,NY,PAS,20100305,14-No Standing                          Y015001003,14,115,K    ,O,139         ,Lawrence
St             ,                  ,54650,93250,42730
7455802936,EPE9473   ,NY,PAS,20100330,34-Expired Meter                                   ,34,035,Q    ,O,103-54      ,94th
St             ,                  ,18740,60810,53790
```

Below screenshot shows the exported data from Hive in MySQL:



```
22  create database airinformation;
23  use airinformation;
24  select * from airlinecount;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

AirlineName	TotalFlights
Air India Limited	393
Air Sahara	180
Go Air	77
IndiGo Airlines	227
Jet Airways	292
Spicejet	196
Air India Expr...	76

6. MAPREDUCE USING PYTHON

6.1 DATASET

Source: https://inventory.data.gov/dataset/032e19b4-5a90-41dc-83ff-6e4cd234f565/resource/38625c3d-5388-4c16-a30f-d105432553a4

This file contains directory information for every institution in the 2013 IPEDS universe. Includes name, address, city, state and zipcode.

File Size: 3.67MB
Number of records: 7770

6.2 PROBLEM STATEMENT

Find top 10 state and county with highest number of institutions.

6.3 APPROACH

- Using mapper to filter state and county as key-value pair
- Map state and county as key in reducer
- Count unique mapped keys and store as value
- Sort the key-value pair based on count and limit to 10
- Store the output in HDFS

6.4 CODE

Code: mapper.py

```
#!/usr/bin/python

import sys
from csv import reader

firstline = True
for line in reader(sys.stdin):
        if firstline:
                firstline = False
                continue
        results = [line[4], line[62]]
        print(",".join(results))
```

Code: reducer.py

```
#!/usr/bin/python

import sys
from operator import itemgetter

foundkey = ""
isFirst = 1
currentCount = 0
results = {}
i = 0

for line in sys.stdin:
        line = line.strip()
        state, county = line.split(",")

        currentkey = '%s,%s' % (state,county)

        if foundkey != currentkey:
                if isFirst == 0:
                        results.update({foundkey: currentCount})
                        currentCount = 0
                else:
                        isFirst = 0
                foundkey = currentkey
        currentCount += 1

for key, value in sorted(results.items(), key=itemgetter(1), reverse=True):
        if i < 10:
                print '%s,%d' % (key, value)
```

```
                i += 1
        else:
                break
```

Command to run Python scripts as map-reduce:

```
hadoop jar /usr/lib/hadoop/contrib/streaming/hadoop-streaming-0.20.2-cdh3u0.jar -file /home/cloudera/mapper.py -mapper /home/cloudera/mapper.py -file /home/cloudera/  reducer.py -reducer /home/cloudera/reducer.py -input /user/python/* -output /user/python-out
```

6.5 EXECUTION

```
cloudera@cloudera-vm:~$ hadoop fs -put postscndryunivsrvy2013dirinfo.csv /user/python
cloudera@cloudera-vm:~$ hadoop jar /usr/lib/hadoop/contrib/streaming/hadoop-streaming-0.20.2-cdh3u0.jar -file /home/cloudera/mapper.py -mapper /home/cloudera/mapper.py -file /home/cloudera/reducer.py -reducer /home/cloudera/reducer.py -input /user/python/* -output /user/python-out
packageJobJar: [/home/cloudera/mapper.py, /home/cloudera/reducer.py, /var/lib/hadoop-0.20/cache/cloudera/hadoop-unjar4555757051869222515/] [] /tmp/streamjob5125153062167765427.jar tmpDir=null
14/10/10 18:02:36 INFO mapred.FileInputFormat: Total input paths to process : 1
14/10/10 18:02:37 INFO streaming.StreamJob: getLocalDirs(): [/var/lib/hadoop-0.20/cache/cloudera/mapred/local]
14/10/10 18:02:37 INFO streaming.StreamJob: Running job: job_201410101748_0001
14/10/10 18:02:37 INFO streaming.StreamJob: To kill this job, run:
14/10/10 18:02:37 INFO streaming.StreamJob: /usr/lib/hadoop-0.20/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201410101748_0001
14/10/10 18:02:37 INFO streaming.StreamJob: Tracking URL: http://localhost.localdomain:50030/jobdetails.jsp?jobid=job_201410101748_0001
14/10/10 18:02:38 INFO streaming.StreamJob:  map 0%  reduce 0%
14/10/10 18:02:47 INFO streaming.StreamJob:  map 100%  reduce 0%
14/10/10 18:02:59 INFO streaming.StreamJob:  map 100%  reduce 100%
14/10/10 18:03:01 INFO streaming.StreamJob: Job complete: job_201410101748_0001
14/10/10 18:03:01 INFO streaming.StreamJob: Output: /user/python-out
```

6.6 RESULT

File: /user/python-out/part-00000

Goto : /user/python-out go

Go back to dir listing

Advanced view/download options

```
CA,Los Angeles County,232
IL,Cook County,146
AZ,Maricopa County,102
NY,New York County,97
TX,Harris County,83
TX,Dallas County,79
FL,Miami-Dade County,77
CA,Orange County,76
CA,San Diego County,71
NY,Kings County,51
```