**Bluemix**

**Terminal**

ssh nitesh@ bi-hadoop-prod-4252.bi.services.us-south.bluemix.net

hadoop123456

**Ambari**

nitesh

hadoop123456

RAW\_DATA = LOAD '/user/zkhan7/Project/2008.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_7 = LOAD '/user/zkhan7/Project/2007.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_6 = LOAD '/user/zkhan7/Project/2006.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_6 = LOAD '/user/zkhan7/Project/2006.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_5 = LOAD '/user/zkhan7/Project/2005.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_4 = LOAD '/user/zkhan7/Project/2004.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_3 = LOAD '/user/zkhan7/Project/2003.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_2 = LOAD '/user/zkhan7/Project/2002.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_1 = LOAD '/user/zkhan7/Project/2001.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_0 = LOAD '/user/zkhan7/Project/2000.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_99 = LOAD '/user/zkhan7/Project/1999.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_98 = LOAD '/user/zkhan7/Project/1998.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_97 = LOAD '/user/zkhan7/Project/1997.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_96 = LOAD '/user/zkhan7/Project/1996.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_95 = LOAD '/user/zkhan7/Project/1995.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_94 = LOAD '/user/zkhan7/Project/1994.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_93 = LOAD '/user/zkhan7/Project/1993.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_92 = LOAD '/user/zkhan7/Project/1992.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_91 = LOAD '/user/zkhan7/Project/1991.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_90 = LOAD '/user/zkhan7/Project/1990.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_89 = LOAD '/user/zkhan7/Project/1989.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_88 = LOAD '/user/zkhan7/Project/1988.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

RAW\_DATA\_87 = LOAD '/user/zkhan7/Project/1987.csv.bz2' USING PigStorage(',') AS

(year: int, month: int, day: int, dow: int,

dtime: int, sdtime: int, arrtime: int, satime: int,

carrier: chararray, fn: int, tn: chararray,

etime: int, setime: int, airtime: int,

adelay: int, ddelay: int,

scode: chararray, dcode: chararray, dist: int,

tintime: int, touttime: int,

cancel: chararray, cancelcode: chararray, diverted: int,

cdelay: int, wdelay: int, ndelay: int, sdelay: int, latedelay: int);

all\_joined = UNION RAW\_DATA, RAW\_DATA\_7, RAW\_DATA\_6, RAW\_DATA\_5, RAW\_DATA\_4, RAW\_DATA\_3, RAW\_DATA\_2, RAW\_DATA\_1, RAW\_DATA\_0, RAW\_DATA\_99, RAW\_DATA\_98, RAW\_DATA\_97, RAW\_DATA\_96, RAW\_DATA\_95, RAW\_DATA\_94, RAW\_DATA\_93, RAW\_DATA\_92, RAW\_DATA\_91, RAW\_DATA\_90, RAW\_DATA\_89, RAW\_DATA\_88, RAW\_DATA\_87;

Important – read the notes for each query on website

<https://github.com/michiard/CLOUDS-LAB/tree/master/labs/pig-lab/sample-solutions/AIRLINE>

Also, Refer supplement data here:

* Airport IATA Codes to City names and Coordinates mapping: <http://stat-computing.org/dataexpo/2009/airports.csv>
* Carrier codes to Full name mapping: <http://stat-computing.org/dataexpo/2009/carriers.csv>
* Information about individual planes: <http://stat-computing.org/dataexpo/2009/plane-data.csv>

In the following, we propose a series of exercises in the form of Queries.

The following are the queries for the analysis that we are going to do:

**Most Popular Airport**

Copy and paste the following query

**CARRIER\_DATA = FOREACH all\_joined GENERATE month AS m, carrier AS cname;**

**GROUP\_CARRIERS = GROUP CARRIER\_DATA BY (m,cname);**

**COUNT\_CARRIERS = FOREACH GROUP\_CARRIERS GENERATE FLATTEN(group), LOG10(COUNT(CARRIER\_DATA)) AS popularity;**

**dump COUNT\_CARRIERS -- we must save the result instead of dumping**

**STORE COUNT\_CARRIERS INTO '/user/schakra2/output/final/COUNT\_CARRIERS' USING PigStorage(',');**

**Top monthly outbound from LAX**

Copy and paste the following query

**OUTBOUND = FOREACH all\_joined GENERATE month AS m, scode AS s;**

**GROUP\_OUTBOUND = GROUP OUTBOUND BY (m,s);**

**COUNT\_OUTBOUND = FOREACH GROUP\_OUTBOUND GENERATE FLATTEN(group), COUNT(OUTBOUND) AS count;**

**GROUP\_COUNT\_OUTBOUND = GROUP COUNT\_OUTBOUND BY m;**

**topMonthlyOutbound = FOREACH GROUP\_COUNT\_OUTBOUND {**

**result = TOP(20, 2, COUNT\_OUTBOUND);**

**GENERATE FLATTEN(result);**

**}**

**STORE topMonthlyOutbound INTO '/user/schakra2/output/final/OUTBOUND-TOP' USING PigStorage(',');**

**Monthly Traffic**

UNION\_TRAFFIC = UNION COUNT\_INBOUND, COUNT\_OUTBOUND;

GROUP\_UNION\_TRAFFIC = GROUP UNION\_TRAFFIC BY (m,d);

TOTAL\_TRAFFIC = FOREACH GROUP\_UNION\_TRAFFIC GENERATE FLATTEN(group) AS (m,code),

SUM(UNION\_TRAFFIC.count) AS total;

TOTAL\_MONTHLY = GROUP TOTAL\_TRAFFIC BY m;

topMonthlyTraffic = FOREACH TOTAL\_MONTHLY {

result = TOP(20, 2, TOTAL\_TRAFFIC);

GENERATE FLATTEN(result) AS (month, iata, traffic);

}

**STORE topMonthlyTraffic INTO '/user/schakra2/output/final/OUTBOUND-TOP' USING PigStorage(',');**

**s**

**Arrival and departure – LAX to other airports**

Copy and paste the following query

**A = FOREACH all\_joined GENERATE scode AS s, dcode AS d;**

**B = GROUP A by (s,d);**

**COUNT = FOREACH B GENERATE group, COUNT(A);**

**DUMP CONT ---- we must save the result instead of dumping**

**STORE COUNT INTO '/user/schakra2/output/final/COUNT' USING PigStorage(',');**

**Average Delay**

Copy and paste the following query

**X= FOREACH all\_joined GENERATE carrier, scode AS s, dcode AS d, float(adelay-ddelay) AS y;**

**Z = GROUP X BY carrier;**

**AVG\_DELAY = FOREACH Z {**

**FILTER X BY (y >= 15);**

**GENERATE carrier, AVG(X.y); }**

**DUMP AVG\_DELAY;**

**STORE AVG\_DELAY INTO '/user/schakra2/output/final/COUNT2' USING PigStorage(',');**

**Longest flight by airtime**

Copy and paste the following query

**A = FOREACH all\_joined GENERATE scode AS s, dcode AS d, arrtime AS x;**

**B = GROUP A BY (s,d,x);**

**LONGEST = FOREACH B GENERATE group, COUNT(x);**

**DUMP LONGEST;**

**STORE LONGEST INTO '/user/schakra2/output/final/COUNT' USING PigStorage(',');**