REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = load '/home/acadgild/leena/ass/aviation/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

B = FOREACH A GENERATE (int)$1 as year, (int)$10 as flightNum, (chararray)$17 as origin,(chararray) $18 as dest;

Testing using limit 10

(2008,335,IAD,TPA)

(2008,3231,IAD,TPA)

(2008,448,IND,BWI)

(2008,3920,IND,BWI)

(2008,378,IND,JAX)

(2008,509,IND,LAS)

(2008,100,IND,MCO)

(2008,1333,IND,MCO)

(2008,2272,IND,MDW)

(2008,675,IND,PHX)

C = filter B by dest is not null;

grunt> C = filter B by dest is not null;

grunt> D = group C by dest;

grunt> E = FOREACH D GENERATE group, COUNT(C.dest);

grunt> F = ORDER E BY $1 DESC;

grunt> G = LIMIT F 5;

Output

(ORD,108984)

(ATL,106898)

(DFW,70657)

(DEN,63003)

(LAX,59969)

X = LOAD '/home/acadgild/leena/ass/aviation/airports.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Y = FOREACH X GENERATE (chararray)$0 as dest, (chararray)$2 as city, (chararray)$4 as country;

joinGY = JOIN G BY $0, Y BY dest;

(ATL,106898,ATL,Atlanta,USA)

(DEN,63003,DEN,Denver,USA)

(DFW,70657,DFW,Dallas-Fort Worth,USA)

(LAX,59969,LAX,Los Angeles,USA)

(ORD,108984,ORD,Chicago,USA)

2) Cancellations due to bad weather

REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = LOAD '/home/acadgild/leena/ass/aviation/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

B = FOREACH A GENERATE (int)$2 as month, (int)$10 as flightNum, (int)$22 as cancelled,(chararray)$23 as cancelcode;

Testing limit B 5

(1,335,0,N)

(1,3231,0,N)

(1,448,0,N)

(1,3920,0,N)

(1,378,0,N)

C = FILTER B BY cancelled == 1 AND cancelcode == 'B';

Testing

(12,170,1,B)

(12,1798,1,B)

(12,1484,1,B)

(12,1382,1,B)

(12,5335,1,B)

grunt> D = GROUP C BY month;

grunt> E = FOREACH D GENERATE group, COUNT(C.cancelled);

grunt> F = ORDER E BY $1 DESC;

grunt> G = LIMIT F 1;

Output (12,250).



3) Highest avg departure delay – top 10 origin

REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = LOAD '/home/acadgild/leena/ass/aviation/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

B = FOREACH A GENERATE (int)$16 AS depDelay, (chararray)$17 AS origin;

C = FILTER B BY (depDelay is not null) AND (origin is not null);

D = GROUP C BY origin;

E = FOREACH D GENERATE group, AVG(C.depDelay);

F = ORDER E BY $1 DESC;

G = LIMIT F 10;

Output

(CMX,116.1470588235294)

(PLN,93.76190476190476)

(SPI,83.84873949579831)

(ALO,82.2258064516129)

(MQT,79.55665024630542)

(ACY,79.3103448275862)

(MOT,78.66165413533835)

(HHH,76.53005464480874)

(EGE,74.12891986062718)

(BGM,73.15533980582525)

4) Which route has seen max diversion

REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = load '/home/acadgild/leena/ass/aviation/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

grunt> B = FOREACH A GENERATE (chararray)$17 AS origin,(chararray)$18 AS dest,(int)$24 AS diversion;

grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);

grunt> D = GROUP C BY (origin,dest);

grunt> E =FOREACH D GENERATE group, COUNT(C.diversion);

grunt> F = ORDER E BY $1 DESC;

grunt> G = LIMIT F 1;

Output

