Documentation

Pig has more functionality than can be discussed in one class. The purpose of this lab is to gain hands-on experience with additional pig functionality. Documentation for pig can be found in http://pig.apache.org/docs/r0.16.0/basic.html. You may need to use the documentation to successfully complete this lab.

Running Pig

Pig can be run interactively with the grunt shell by typing "pig" at the command line. You can also run a pig script by typing "pig nameofscript.pig"

What to upload:

a file called "piglab.pig"

Dataset

The data is in hdfs in the directory /ds410/flightdata/2010-summary.csv

Assignment:

In your pig file, add commands to:

- Count the number of destination countries for each origin. Store the result in 'destcount'
- Count the total number of outgoing flights for each origin. Store the result in 'outcount'
- List the origins that have at least 3 different destinations for which they have at least 100 flights (for example, "United States" is one such origin because it has 477 flights to Costa Rica, 390 flights to Italy, and 118 flights to Iceland).

Note: the input file has a header so you will need to get rid of the header as well using your pig script.