Data: for this homework we will be using the flight data from /ds410/flightdata/2010-summary.csv in the HDFS file system. Every line contains an ORIGIN_COUNTRY_NAME, a DEST_COUNTRY_NAME and a count which is the number of daily flights from the origin country to the destination country.

Query: for each pair of countries (A, B), we are interested in how many ways there are of going from A to B in exactly 2 flights. For example, if the flight data looks like this:

DEST_COUNTRY_NAME	ORIGIN_COUNTRY_NAME	count
Canada	United States	1
Mexico	United States	2
Germany	Mexico	3
Germany	Canada	4
France	Germany	6
Germany	United States	5

then the output would look like:

United States	France	30
United States	Germany	10
Mexico	France	18
Canada	France	24

Because, for example, you can go from United States to Germany using exactly two flights in the following ways:

- 2 flights from United States to Mexico and 3 flights from Mexico to Germany, giving 2x3=6 choices
- 1 flight from United States to Canada and 4 flights from Canada to Germany, giving 1x4=4 choices

So overall, there are 6+4=10 possibilities.

Assignment: implement this in mapreduce. It involves a join (between flight data and itself) and possibly 2 steps (2 map/reduce pairs). Hint: think of how you would do it using 1 or maybe 2 sql queries (what exactly do you need to join on?) and then do it in mapreduce.

Upload: you will be producing 2 files:

1. Your mapreduce code. Call it flight.py
2. Your output. Use the "cat" command to save the entire output locally (hdfs dfs -cat my_output_directory/* > result.txt). This will create a file called "result.txt" that has your answers. Upload this to canvas.