Assignment 1-MapReduce Warm-up (15 points)

Description

For this assignment, you are given a large dataset of flight arrival and departure details for all commercial flights within the USA between the years 1987-2000. The original dataset is about 5.5 GB and is extracted from here: http://stat-computing.org/dataexpo/2009/ . The goal is to write a simple MapReduce programs which returns, for each unique carrier, the most frequent flight destination for each major airport.

What is the format of the dataset and how can you access it?

The dataset is a folder consisting of 9 comma-separated files where each file contains flight information for a particular year. For example 2000.csv contains all flight information for year 2000. I have uploaded the entire dataset on my s3 bucket, you can access it via the following path:

s3://class-data-set/flight-data

(Note: you will not be able to download the dataset from the above path, you just provide this path as input when running your program on EMR)

Each file in the data folder consists of 29 columns. The descriptions of the columns are as follows:

	Name	Description
1	Year	1987-2008

2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled? (0=false, 1=true)
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C =NAS, D = security)
24	Diverted	1 = yes, 0 = no

25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Each line in a csv file corresponds to a flight. For example a line of file 2000.csv could be as follows:

```
2000,1,29,6,1648,1647,1939,1859,HP,154,N653AW,291,252,239,40,1,ATL,PHX,
1587,5,47,0,NA,0,NA,NA,NA,NA,NA
```

This means: the flight year=2000, month=1, day of month=29, day of week=6, departure time=16:47, etc.

We don't need all the columns for each line, we are only interested in column 9 (the unique carrier), column 17 (origin), and column 18(destination).For example, for the above flight, carrier=HP, origin=ATL, and destination =PHX.

Note: The first line in each comma-separated file contains column headers. In your map function you need to check whether the line that you received as value is a header and if so do not emit anything for that line. In addition, there might be some missing values for origin or destination marked as "na". Your map function should not emit anything for those flights.

What you need to do:

Write a MapReduce program which returns, for each unique carrier and origin, the most frequent flight destination. Each line of your output must have the following format:

Unique_carrier origin most_frequent_destination frequency

Your output must be sorted by unique carrier and origin. Run your program in standalone mode on smaller sample of data. For example, run it only for file 2000.csv . You can download the file 2000.csv from the following URL:

http://stat-computing.org/dataexpo/2009/2000.csv.bz2

When I run my program on the input file 2000.csv, the first 10 lines of output generated are as follows.

```
AA    ABQ   DFW   2169
AA    ALB   ORD   1094
AA    AMA   DFW   1090
AA    ANC   SEA   30
AA    ATL   DFW   4440
AA    AUS   DFW   4700
AA    BDL   ORD   2137
AA    BHM   DFW   1096
AA    BNA   DFW   2910
AA    BOS   ORD   5001
```

This means for example, that the most frequent destination for American Airline (AA) flights departing from ABQ(Albuquerque) was DFW(Dallas International Airport) and that the number of AA flights from ABQ to Dallas was 2169.

Once you are confident that your program works correctly and produces the expected output, create a jar and run your program on Amazon EMR on the entire dataset. Please do not forget to terminate your cluster on EMR after your step has been completed otherwise that will use up all your credits for this class.

What you need to submit:

1- **The source code for your Mapper, reducer, and driver class. Please name your classes as follows: MostFrequentDestMapper.java, MostFrequentDestReducer.java, and MostFrequentDestDriver.java**
2- **Submit the first part of your output on EMR (part-0000)**

Good luck and please do not hesitate to email me if you have any questions.