Hadoop Pig Laboratory

The goal of this lab is to gain familiarity with Hadoop and the Pig Latin language to analyze data in different ways. In other words, to focus on "what to do" with your data: to perform some simple statistics and to mine useful information. Additional documentation that is useful for the exercises is available at: https://pig.apache.org/.

Useful tools for "debugging":

- **DESCRIBE** relation: this is very useful to understand the schema applied to each relation. Note that understanding schema propagation in Pig requires some time.
- **DUMP** relation: this command is similar to the STORE command, except that it outputs on `stdout` the selected relation.
- **ILLUSTRATE** relation: this command is useful to get a sample of the data in a relation.
- **EXPLAIN** generates (text and .dot) files that illustrate the DAG (directed acyclic graph) of the MapReduce jobs produced by Pig, and can be visualized by some graph-chart tools, such as GraphViz. This is very useful to grab an idea of what is going on under the hood.

Access to Hadoop

1) Option #1: Amazon Elastic MapReduce (Amazon EMR) is a web service that makes it easy to quickly and cost-effectively process vast amounts of data. Amazon EMR uses Hadoop, an open source framework, to distribute your data and processing across a resizable cluster of Amazon EC2 instances. Please register an Amazon AWS account using your IUPUI email and request student credit. You should be able to receive $40 credit for your account.
2) Option #2: Download and install on your desktop or laptop Hortonworks HDP Sandbox, a self-contained VM image pre-loaded with all the needed Hadoop software.
3) Option #3: Download and install on your desktop or laptop Clouder QuickStart Virtual Machines.

   If you choose Option #2 or Option #3, the machine you install Hadoop virtual machine is usually required to have at least 8G-10G of RAM.

Exercise 1:: Word Count

In this exercise, a text file ‘exercise1.txt’ (eBook of Pride and Prejudice) is provided. Please write Pig Latin scripts to

1) Count the occurrences of each word in exercise1.txt
2) List the 10 most popular words in exercise1.txt

Exercise 2:: Airflight Data Analysis

In this exercise, an exercise2.csv file is given. It has 29 fields which provide air flights information.

```
1       Year   actual departure year
2       Month  actual departure month 1-12
3       DayofMonth     actual departure day 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?
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
```

An carrier.csv file is given which provides carrier codes to full name mapping, and an airport.csv gives the airport code to full name mapping. Based on these files, write PIG scripts for the following queries:

1. Suppose a flight is called on time if the departure delay (attribute #16) and arrival delay(attribute #15) are both less than 15 minutes, compute the portion of the flights that are on time.
2. Compute the proportion of on-time flights by carrier, ranked by carrier
3. Compute the busiest routes. You can create a frequency table for the unordered pair (i,j) where i and j are distinct airport codes.
4. Find the names of the 10 most popular carriers
5. Find the names of the top 3 airports with the most inbound traffic.

All files (exercise1.txt, exercise2.csv.bz2, carrier.csv, airport.csv) can be found on Canvas under Files/Assignments.

This assignment should be completed independently; no collaboration with classmates is allowed. The output of all the queries should be stored into file(s).

Please submit your PIG scripts for all queries. Your scripts should be clearly written and ready for the TA to run.