Project of CS 644: Introduction to Big Data
Flight Data Analysis

In this project, you will develop an Oozie workflow to process and analyze a large volume of flight data.

- Instructions:
  1. Form a project team of two students (including yourself).
  2. Install Hadoop/Oozie on your AWS VMs.
  3. Download the Airline On-time Performance data set (flight data set) from the period of October 1987 to April 2008 on the Statistical Computing website: http://stat-computing.org/dataexpo/2009/the-data.html
  4. Design, implement, and run an Oozie workflow to find out
     a. the 3 airlines with the highest and lowest probability, respectively, for being on schedule;
     b. the 3 airports with the longest and shortest average taxi time per flight (both in and out), respectively; and
     c. the most common reason for flight cancellations.

- Requirements:
  1. Your workflow must contain at least three MapReduce jobs that run in fully distributed mode.
  2. Run your workflow to analyze the entire data set (total 22 years from 1987 to 2008) at one time on two VMs first and then gradually increase the system scale to the maximum allowed number of VMs for at least 5 increment steps, and measure each corresponding workflow execution time.
  3. Run your workflow to analyze the data in a progressive manner with an increment of 1 year, i.e. the first year (1987), the first 2 years (1987-1988), the first 3 years (1987-1989), …, and the total 22 years (1987-2008), on the maximum allowed number of VMs, and measure each corresponding workflow execution time.

- Submission (all in a zipped file: YourLastName_YourPartnerLastName.zip):
  1. A commands.txt text file that lists all the commands you used to run your code and produce the required results in fully distributed mode
  2. An output.txt text file that stores the final results from all the runs
  3. The source code of your MapReduce programs (including the JAR files) and any other programs you might have developed and included in the workflow
  4. The Oozie workflow XML file
  5. A project report in PDF that includes:
     a. A diagram that shows the structure of your Oozie workflow
     b. A detailed description of the algorithm you designed to solve each of the problems
     c. A performance measurement plot that compares the workflow execution time in response to an increasing number of VMs used for processing the entire data set (22 years) and an in-depth discussion on the observed performance comparison results
     d. A performance measurement plot that compares the workflow execution time in response to an increasing data size (from 1 year to 22 years) and an in-depth discussion on the observed performance comparison results