Project of CS 444: Big Data Systems
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 at most three students (including yourself).
  2. Install Hadoop/Oozie and any other packages you might need 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 following website: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7
  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 are executed in the 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 scale up the system to the maximum allowed number of VMs for at least 5 increment steps (e.g., 5 steps with an increment of 2 VMs: 2 → 4 → 6 → 8 → 10), 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: LastNamesOfAllTeamMembers.zip):
  1. A commands.txt text file that lists all the commands you used to run your code (similar to the command list provided in the tutorials) and produce the required results in the fully distributed mode
  2. An output.txt text file that stores the final results from all workflow runs
  3. The source code of your MapReduce programs (including the JAR files) and any other programs you might have developed and included in your workflow
  4. The Oozie workflow configuration 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 algorithms 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