CS499 Cloud Computing and Big Data
Assignment 3 - MapReduce in Docker

Due Date

Wed before class, March 1, 2017

Score

50

Questions and Directions

In this assignment, you need to write MapReduce jobs and run it in a Docker container.

1. Processing the Netflix Dataset

Netflix hosted a $1M competition a few years ago to find out the best movie recommendation algorithm (http://www.netflixprize.com/). We are going to use part of that dataset to practice the MapReduce.

Please download this dataset: https://s3-us-west-1.amazonaws.com/cs499-cc/a3-dataset.zip which contains the users' rating on a set of movies. Each row represents a rating of a movie by some user. The dataset contains 1821 movies and 28978 users in all. Ratings are integers from 1 to 5. The training set has 3.25 million ratings.

You need to write two MapReduce jobs to answer the following two questions:

1) What are the top 10 movies that have the highest average ratings? Tell us the titles (you can find the titles in a separated file).
2) Who are the top 10 users that posted the most reviews? Tell us their user ids.

Note that you don't have to output the final result directly from your Reducer. You can have a separate Java program to take the MapReduce output to figure out the rankings, the top 10, or find out the titles. You can do whatever you need to get the result, but you must have two MapReduce jobs written to accomplish the core tasks. This is a good tutorial to use for the basis of MapReduce:
https://examples.javacodegeeks.com/enterprise-java/apache-hadoop/hadoop-hello-world-example/

2. Run Your MapReduce in a Docker Container

You can use Eclipse to test your MapReduce program. However, you are also required to use Docker Container to run your MapReduce jobs. Thus, prepare some time to install and learn how to use Docker Container (https://www.docker.com/). There might be a little bit learning curve in the beginning, but it is worth the time, because Docker is very popular in the industry today.

I would recommend this Docker Image (https://hub.docker.com/r/nagasuga/docker-hive/), which has everything setup for you already. You need to put all your java programs and dataset into this container, and run your program using hadoop cluster and the HDFS. Look for online resources and tutorials on how to run a jar file with hadoop and hdfs.

Once you have done, you need to push the container that includes your own jar file and dataset to the Docker repository (https://docs.docker.com/engine/getstarted/step_six/). And of course, you will need to have a Docker account.

Submission

You need to submission the following:

1. The URL to your GitHub repo that contains 1) all the source code and 2) two text file that contains the answers to the 2 questions
2. The URL to the your own Docker Container Image that contains your MapReduce jar file in your Docker Hub, such as https://hub.docker.com/r/sunyu912/jhipsterdemo/.

Please use this Google Form (https://goo.gl/forms/REkksv3rsglzZ5ou2) to submit your URLs.