CS 643, Cloud Computing - Fall ‘16

Programming Assignment – Due 10/17, 5PM

Goal: The purpose of this individual assignment is to learn how to use the Amazon Web Services (AWS) platform and to learn how to develop parallel applications using the Hadoop programming framework.

Description: The assignment is divided in 2 tasks:

1. Build your own Hadoop AMI, starting from a basic Ubuntu AMI. You have to use latest stable Hadoop release. You are required to store this AMI in S3, and its name must include your last name. This AMI will be tested with the application built for task 2. However, if your AMI doesn’t work you are allowed to use one of the pre-built Hadoop AMIs for task 2.

2. Write a Hadoop/Yarn MapReduce application that takes as input the 50 Wikipedia web pages dedicated to the US states (we will provide these files for consistency) and:
   a) Computes how many times the words “education”, “politics”, “sports”, and “agriculture” appear in each file. Then, the program outputs the number of states for which each of these words is dominant (i.e., appears more times than the other three words).
   b) Identify all states that have the same ranking of these four words. For example, NY, NJ, PA may have the ranking 1. Politics; 2. Sports. 3. Agriculture; 4. Education (meaning “politics” appears more times than “sports” in the Wikipedia file of the state, “sports” appears more times than “agriculture”, etc.)

   This program will be tested on a 4-node AWS cluster.

Submission: Email the following files to the TA (Jianchen Shan js622@njit.edu)

- AMI.txt – a step by step description of how you built your AMI. This file must also include the name of your AMI.
- The code of your Hadoop application together with a README file containing any instructions necessary to run the application as well as the name of the AMI you used to run it (in case you didn’t use the one you built).

Grading:

- Task 1	– 30 points
- Task 2.a	– 35 points
- Task 2.b	– 35 points