Programming Assignment 3

Migrating Your Analysis to Amazon’s AWS Cluster: Using EMR, Glacier, and S3

Due: April 17th, 2015 By 5:00PM
No submission required

Objectives
The objectives of this programming assignment are to enable you to gain experience with:

- AWS’ Elastic Map Reduce (EMR) service
- Storage system (Glacier, S3)

0. Create your account
If you do not have AWS account, please create your account. Please follow the instructions at:
http://aws.amazon.com/getting-started/

If you have any problem creating your account, please let us know immediately.

1. Access the public data block stored in Glacier service
The input data that you have used for the assignment 2 will be uploaded as a public Glacier volume. A data stored in Glacier service is available within an Availability Zone. If your Availability zone is not the same as one of your Glacier data, you should change your zone.

2. Prepare the Storage for Input and Output
In this assignment, you should use S3 to store your input and output data. The output of your off-line computing should be stored in S3 and retrieved for the command-line software. The document that the authorship is unknown should be retrieved from Glacier and stored in S3 to test your software and the location should be passed as a parameter. More information is available at,
http://media.amazonaws.com/AWS_Amazon_EMR_Best_Practices.pdf

Your output stored in S3 should be human readable text.
3. Create your EMR cluster
Now, create your EMR cluster. You can do this by following instructions available at:
http://aws.amazon.com/elasticmapreduce/

4. Run Your Job
Run your MapReduce job on the EMR cluster.

5. Submission and Grading
Code submission is not required for PA3. You should schedule your demos with the GTA. Demo includes:
   (1) Setting up of the EMR cluster with input at Amazon’s AWS and input data from Glacier. (2 points)
   (2) Running the MapReduce jobs (assignment 2) (4 points)
   (3) Retrieving outputs from your S3 directory (2 points)
   (4) An interview (2-3 questions) about the technical details pertaining to this assignment (2 points)

If you could not finish PA2 successfully, please contact GTA.

This assignment will account for 10% of your final course grade. The grading will be done over a 10-point scale.