Amazon Web Services Setup Guidelines for Homework 5

Goals:

1. Create an AWS instance (to get access to EC2, Elastic MapReduce and S3 storage).
2. Create storage buckets on S3 (to save outputs and logs of MR jobs)
3. Create a key pair (required for running MR jobs on EC2)
4. Get Access keys (also required for running MR jobs on EC2)
5. Get and redeem your free credit (worth $100) (and monitor how much you have left). You have to **contact Qianzhou for this step**, so please do it early.
6. Familiarize yourself with S3, EC2 and EMR (by doing a sample MR run)

1. Create an AWS account

a. Go to http://aws.amazon.com/account and sign up for an account, if you don't have one already.
b. Follow all the required steps and enter all the details. Assume you don't have any promotional credit---hence you'll have to enter payment details (you'll need **a valid credit/debit card**). You'll also have to validate using your phone.
c. Choose the 'basic' plan on the 'AWS Support plan' screen that is displayed after you validate your identity.
d. Once everything has been verified and created, you should have access to the AWS management Console.

2. Create storage buckets on S3

In the AWS Management Console click on "S3" under Storage & Content Delivery. We need S3 for two reasons: (1) an EMR workflow requires the input data to be on S3; and (2) EMR workflow output is always saved to S3.

Data (or objects) in S3 are stored in what we call "buckets" (essentially directories). For this assignment, the data you will process is in a public bucket called:

s3n://cs4604-2014-vt-cs-data/livejournal/

You will see how to reference this for EMR input later on. In the meanwhile you will need some buckets of your own (to store output and log files if you want to debug your runs).

For creating a 'log' bucket:

a. In the S3 console, click on 'Create Bucket'.
b. All S3 buckets need to have unique names, so call your logging bucket *cs4604-vt-cs-YOURVTID-logging*. Importantly, pick 'US Standard' for the Region dropdown. Click on "Create" (not on "Set-up Logging")
c. Your new bucket will appear in the S3 console. Clicking on it will tell you it is empty.

Now we will create our main bucket:

a. Again, create bucket. Name it *cs4604-vt-cs-YOURVTID*. Again pick US Standard time. Now we to link our logging bucket to this one---so click on 'Set Up Logging >>'.
b. "Enable Logging" and start typing in the name of your logging bucket. It should appear in the drop down menu. Select it and 'Create'.

3. Create a key pair

When you run jobs on EMR, you will need a valid public/private key pair for authentication. To create your first key pair:

a. Click on "EC2" under Compute and Networking section in the AWS management console.
b. Select the region, on the top right as US East or US Standard. The page will refresh.
c. On the refreshed page you should see a link stating '0 Key Pairs'. Click on this.
d. You will be given an option top 'Create Key Pair'. Name your key pair as you wish.
e. Upon providing a name and clicking on 'Create', your private key (a .pem file), will automatically begin downloading---choose a safe place so that you can retrieve it again.
f. If you need to access your public key, you will be able to find it in the same place where you found your account credentials. Amazon will not keep a record of your private key (as expected), so if you lose it, you will need to generate a new set.

Note: You would not really need to access your private key if you use the AWS Management Console, but you will be asked to name your key pair each time you run an EMR job. If you wish to log into the master node running your MapReduce job, you will need your .pem file. To log on to the master node (you can find the address of the master node from the MapReduce dashboard), you will need to do the following:

$ ssh hadoop@<master-node-address> -i <path-to-pem-file>

4. Get Access keys

Go to your Security Credentials from the AWS management console. The link to this page is in the dropdown under your name on the top right corner of the AWS management console. Under the Access Credentials section, check your Access Keys list. Click on the Create a new Access Key link. On clicking it will create a new access key, please download it. Now you are ready to run a MapReduce job.

5. Get and Redeem your free credit

In order to get your credit, you will need a unique credit code. Please send email to qiand12@vt.edu with the subject header 'CS4604: AWS Code' and he will mail you one (follow subject header **strictly**). Once you have your unique credit code go to 'Billing and Cost Management' link in the dropdown under your name, on the top right corner of the AWS management console---click on 'Credits' on the left side menu bar. Enter your code and click on 'Redeem'---if it does not work, email us asap. You will be able to see all your 'Credit' information such as usage on this page.

Important: There is only so much credit we can give---so you should always check how much credit you have left by clicking on the 'Account Activity' link from your account page. Sometimes this takes a while to update. Always make sure to test your mappers and reducers on some sample local data before using the AWS.

6. Familiarize yourself with S3, EC2, and EMR

Run the sample word-count application that comes with AWS.
To do this, follow these steps:

a. Click on the Elastic MapReduce link in the Analytics Section of the AWS management console. This will take you to the EMR Cluster page.
b. Click on the Create cluster link, you will be taken to 'cluster configuration' page.
c. Click on 'Configure Sample Application' on top right corner. Follow the directions to run the sample application 'Word Count (Streaming)'.

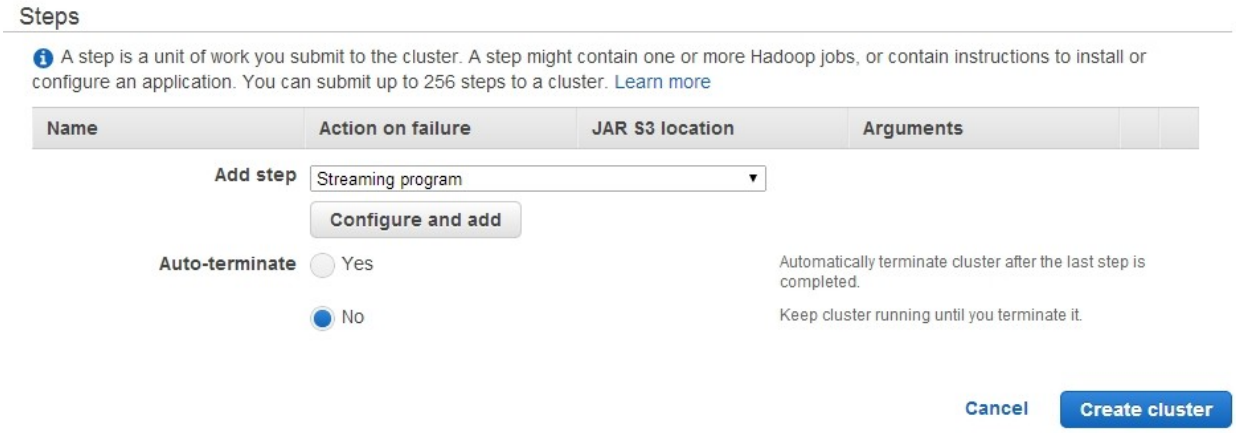
Most of the directions are clear and stick to the default values (except your output and logging bucket which you will need to specify to get the output). Once you create your cluster, it will

take 5-6 minutes to finish and the output will be in your S3 bucket. More step-by-step information on how-to use this sample application is given here:

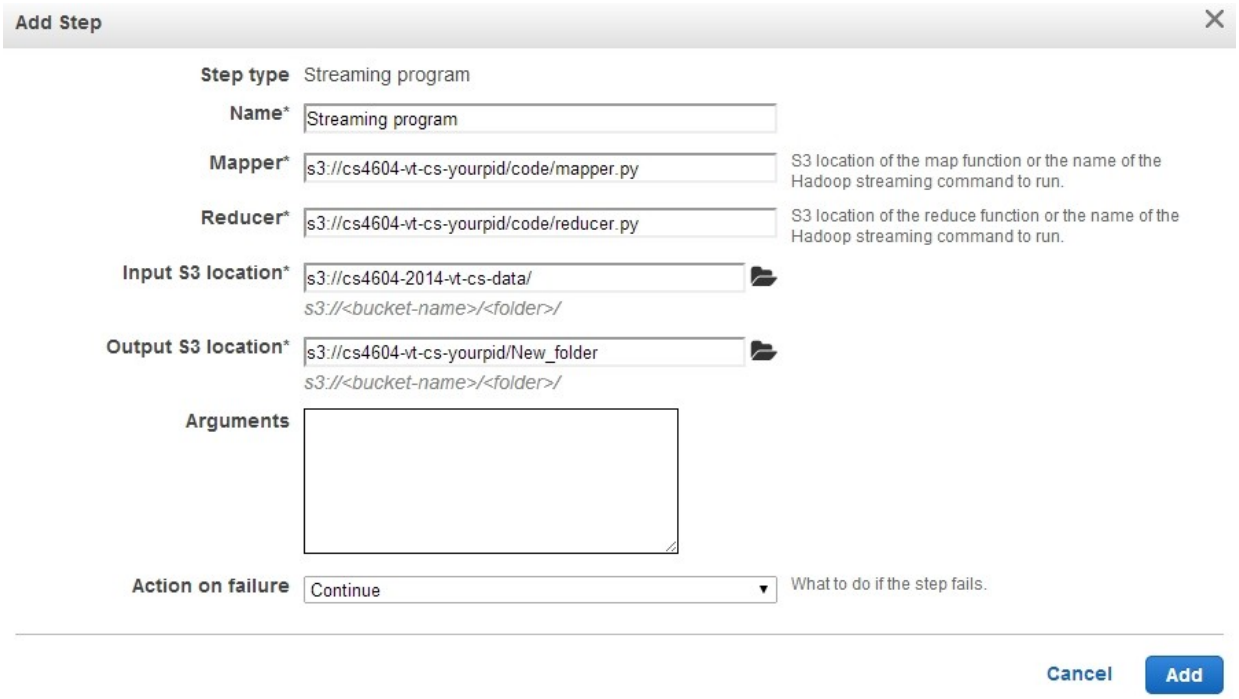
http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-get-started-count-words-step-5.html

Main clarifications wrt the steps in the above link:
Step 8: Choose "vpc-xxxxxx(default)" for "Network". For the number of EC2 instance, the default number is enough for word count. (You can use a different number here or for future job, up to 20)
Step 9: Choose the key pair created by yourself for "EC2 key pair" and then you can use the responding .pem file to login into the master node of Hadoop by using the command in Section 3 (if needed; typically you won't need to). For "IAM role", please set it as "No role found".

When you run your own mappers and reducers for HW5, there is little difference as follows:
1. Ignore the step of configure sample application.
2. In Step 11, you need to set up a 'step'. In our homework, we can set the step as "Streaming program" (easiest) or JAR (harder).



And then configure and add this step: (screen-shot in the next page)



Before you do this step, make sure you have updated your mapper and reducer code to the responding S3 bucket (folder).

Note that this is only for one mapper and reducer. If you want to run a series of MR jobs: Set-up and run the first MR job. Try to check the output if the format looks OK. Then set-up the second MR job by taking the output bucket of the first MR job and setting it as the input of the second MR job and so on.

You can run jobs using the AWS web-console like above in this assignment (the easier way); or you are also welcome to use the elastic-mapreduce command-line interface based on Ruby.

Follow Ruby-AWS instructions given here:
http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-cli-reference.html

and see a sample invocation here:
http://cs.smith.edu/dftwiki/index.php/Hadoop_Tutorial_3.2_--_Using_Your_Own_WordCount_program

Note: AWS has excellent documentation (see http://aws.amazon.com/documentation). So make sure to check it out in case of any further doubts.

Acks: Thanks to Amazon for generously providing free credits. Adapted from similar guidelines by Polo Chau (GTech) and Diana Maclean (Stanford).