Amazon Web Services (AWS) Setup Guidelines

For CSE6242 HW3, updated version of the guidelines by Diana Maclean
[Estimated time needed: 1 hour]

Note that important steps are highlighted in yellow.

What we will try to accomplish with this?

This guideline will help you get set up with the Amazon Web Services (AWS, a “cloud” platform) where you will run large-scale analysis on big data. Here are you will learn to do

1. **Create an AWS account** (to get access to EC2, Elastic MapReduce and S3 storage).
2. **Create storage buckets on S3** (to save outputs and logs of MapReduce jobs).
3. **Create a key pair** (required for running MapReduce jobs on EC2).
4. **Get Access Keys** (also required for running jobs on EC2).
5. **Redeem your free credit** (worth $100).
6. **Familiarize yourself with S3, EC2 and EMR** (by doing a sample MapReduce run).

1. **Create an AWS account**
   - Go to [http://aws.amazon.com](http://aws.amazon.com) and sign up for an account, if you do not have one already.
   - For now, please enter the required details, including payment details (you will need a valid credit card or debit card to sign up). Please follow Step 5 to redeem the $100 credits.
   - Validate your account with the identity verification through your phone.

Once your account has been created and your payment method verified, 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;
(2) EMR workflow output is always saved to S3. Data (or objects) in S3 are stored in what we call “buckets”. You can think of buckets as folders.

You will need to create some buckets of your own to (1) store your EMR output, and (2) store your log files if you wish to debug your EMR runs.

Once you have signed up, we will begin by creating the log bucket first.

i. In the S3 console, click on “Create Bucket”.
ii. All S3 buckets need to have unique names. You could name the logging bucket `cse6242-<gt-username>-logging`. **Important:** Please select “**US Standard**” in the Region dropdown. Click on “**Create**” (not on “Set Up Logging >>”).

iii. Your new bucket will appear in the S3 console. Clicking on it will show you that it is empty.

iv. Now we will create our main bucket. Go back to the main screen (clicking on “All Buckets”). Again, click on “Create Bucket”. Call this one `cse6242-<gt-username>`. Again, pick “**US Standard**” for the Region dropdown. Since we will link this bucket to our logging bucket, the regions for the two buckets should be the same. We will link our logging bucket to the one we are creating now, so click on “Set Up Logging >>”.

v. Click on “Enabled” to enable logging, and start typing in the name of your logging bucket. It should appear in the drop down menu, select it. Clear the “Target Prefix” field and click “Create”.

We are done creating buckets at this point.
3. Create a key pair

When you run jobs on EMR, you will need to have a valid public/private key pair. To create your first key pair, click on “EC2” under Compute in the AWS Management Console.

Select the region on the top right as U.S. West(Oregon) since the data bucket is stored in this region. You should see a link stating “0 Key Pairs” under Resources. Click on this.
You will be given an option to “Create Key Pair”. Name your key pair as you wish. Upon providing a name and clicking on “Create”, your private key (a .pem file) will automatically download. **Save it in a safe place where you will be able to find it again (IMPORTANT, do not lose this file).**

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 keeps no record of your private key, and if you lose it, you will need to generate a new set.

If your computer runs Windows, use the steps in the following link to convert your .pem file to a .ppk file for use with PuTTY. Read the section titled Converting Your Private Key Using PuTTYgen in the link below: [http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html)

**Note:** If you use the AWS Management Console, you would typically not be required to access your private key. However, you will be asked to name your access key pair and the private key 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 (you will need this in case you wish to run an interactive HIVE/PIG job flow). 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>/<pem-file-name>.pem
```
4. Get Access Keys (new site)

Click on “Security Credentials” under your username (top right). Click on “Continue ...”

Click on the **Create a new Access Key** link (under Access Keys), and download the Access Key file (do not lose this file). Now you are ready to run a MapReduce job.
5. Redeem your free credit

In order to add the credit to your account, you will need your unique Credit Code obtained after applying for the AWS Educate program for Students (follow steps listed at the start of HW3). Once you have your code, go to your account page (http://aws.amazon.com/account).

Click on “Credits”. Enter the Code into the Promo Code text box, and click Redeem.

Please email the CSE6242 instructors immediately if this does not work. You can check the credit remaining by clicking on the “Account Activity” link from your account page.
or by returning to this page. Sometimes this can take a while to update, so don’t be surprised if recent changes are not immediately apparent. We will set up a monitor in the next step which is triggered when you utilize half of the credit.

6. Set up a CloudWatch Usage Alert

Make sure your region (in the upper right corner of the screen) is set to: US West (Oregon). Test whether this email alert is working before scheduling in practice. That is, out of 100$, when your credit balance goes below say 95$, schedule a test alert and make sure it works. Remember this alert works only once. So once you got an alert for 95$, you schedule the next alert for 70$ and the next one for 60$ and so on. Also understand the difference between metrics `CPUCreditBalance` and `EstimatedCharges` before creating these alerts. To setup alerts using `EstimatedCharges` follow the link http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/free-tier-alarms.html.

Now we will turn on alerts.


2. Log In using your AWS credentials if necessary

3. Under Preferences, check the box labeled Receive Billing Alerts
Now we need to create a custom alarm so that it tells you when you have spent money.

1. Open the Amazon CloudWatch console. Click CloudWatch in the AWS Management Console.

2. In the navigation pane on the left, click Alarms, and then in the Alarms pane, click Create Alarm.
3. In the CloudWatch Metrics by Category pane, under Select Metrics, select EC2, followed by the CPUCreditBalance metric. If you do not see this metric, start an EC2 Instance and wait for 10 minutes.

4. Select USD then click the Absolute tab in the lower right. Enter the date dates below, select Sum, and “1 Day”. Then click next.
5. Fill out the alarm details and click **New List** next to “Send notification to:”:
Alarm Threshold

Provide the details and threshold for your alarm. Use the graph on the right to help set the appropriate threshold.

Name: Halfway Broke
Description: I have spent $50!

Whenever charges for: EstimatedCharges
is: $50

Actions

Define what actions are taken when your alarm changes state.

<table>
<thead>
<tr>
<th>Notification</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whenever this alarm:</td>
<td>State is ALARM</td>
</tr>
<tr>
<td>Send notification to:</td>
<td>Select a notification list</td>
</tr>
</tbody>
</table>

Enter your name and email.
You have now created an alert that will bother you when you pass $50. Consider making another alert which is activated when you use up $90 so that you do not get charged!

7. Familiarize yourself with S3, EC2 and EMR

We will now attempt to run a sample application of word count that comes with AWS. We will begin by clicking on the Elastic MapReduce (EMR) link in the Analytics section of the AWS Management Console. This will take you to the EMR Job Flows page. Click on the “Create Cluster” link.

Click on “Go to the advanced options”.

You have now created an alert that will bother you when you pass $50. Consider making another alert which is activated when you use up $90 so that you do not get charged!
Click on **Configure sample application**. Choose application **“Word count”**, then choose the output and logging bucket as the buckets you created previously. Click **“OK”** to configure the sample application.

The configure screen has many options, selecting a sample application will configure the cluster for you. We will show you how to do it manually, since you will need to do so for your assignment.
*Note the versions of the applications and the distribution may be different.

Make sure to select your logging bucket you made earlier under the log folder location.
Note: If your account supports only EC2-VPC, you can select the default VPC from the Network list i.e. you will not see “EC2-Classic”.

The costs listed(https://aws.amazon.com/ec2/pricing/) are charged on an hourly rate, based on the number and type of nodes in your cluster.

You don’t need to change these.

You don’t need to change these.

You don’t need to change these.

You will modify this to speed up computation, you can’t use more than 19 Cores.

Select your key pair here.

IF YOU CHOOSE NO: YOU HAVE TO SHUT DOWN THE CLUSTER YOURSELF!
Click “Configure and add” to add steps.

For the S3 output Location you should specify the bucket and an additional unique folder for each new run. It will help with organization.

Remember for the word count sample application it sets everything up for you. Scroll down to the end of the page, click on “Create cluster” to run the application.

You now can view the status of your application in “Cluster Details” screen. It takes several minutes for the whole process to run.

Provisioning - Amazon locates resources for your application
Bootstrapping - Amazon sets up and configures the nodes to run your application
Running - Runs and writes to your output bucket.
Terminating - Amazon deconstructs the setups you used for the application

You can track its progress once it’s been created.
After the application terminates, you could go back to the S3 output bucket you chose. The results will be written to the output folder. You should have several partxxxx files in the output folder. These are texts of the output! You have just successfully completed a MapReduce job flow on AWS and are ready for large scale data analytics!