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 set you up with the Amazon Web Services (AWS, a “cloud” platform) where you will run large-scale analyses 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 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). Amazon has generously agreed to provide each student with credit for this class; more on how to redeem this later.
   - 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.
You AWS Management Console should look like this:

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. For this assignment, we have put the data you will process in a public bucket called:

\textit{cse6242-gtcse-data}

You will see how to reference this for EMR input later on. In the meanwhile, you will need some buckets of your own to (1) store your EMR output in, and (2) store your log files if you want to debug your EMR runs.

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

i. In the S3 console, click on “Create Bucket”.

ii. All S3 buckets have to have unique names, so call your logging bucket `cse6242-<gt-username>-logging`. Importantly, pick “US Standard” for the Region dropdown. Click on “Create” (not on “Set Up Logging >>”).

![Image of the S3 console showing create bucket]

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

![Image of the S3 console showing an empty bucket]

iv. Now we will create our main bucket. 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. “Enable logging”, and start typing in the name of your logging bucket. It should appear in the drop down menu. Select it, and “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 & Networking in the AWS Management Console.

Select the region on the top right as U.S. East or U.S. Standard because U.S. Standard is where the data bucket is. You should see a link stating “0 Key Pairs”. 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.

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, however,
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 (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

Go to your Security Credentials from the management console (quick link here). Under the Access Credentials section, check your Access Keys list. Click on the Create a new Access Key link. Now you are ready to run a MapReduce job.
5. Redeem your free credit

In order to be credited, you will need your unique Credit Coupon Code. If you have not received this yet, e-mail me and I will issue you one. Once you have your code, go to your account page (http://aws.amazon.com/account)

Click on “Payment Method”. You will see your billing details; if you scroll down, you will also see an option to redeem a Credit Coupon.

Click on Redeem/View AWS Credits, and enter your code to get your credit. Please email the CSE6242 instructors immediately if this does not work. Unfortunately, we can only give
you so much free credit, so don’t go too wild! You can check on how much credit you have left by clicking on the “AccountActivity” link from your account page. Sometimes this can take a while to update, so don’t be surprised if recent changes are not immediately apparent.

6. 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 link in the Compute and Networking section of the AWS Management Console. This will take you to the EMR Job Flows page. Click on the Create New Job Flow link.

![Image of EMR Job Flows](image_url)

This will let you a create a new job flow. You need to specify the name of the job flow. For Create a Job Flow select “Run a sample application”. Select “Word Count (Streaming)” from the sample application and Continue.
Now you need to specify the location of the input (data), output, Mapper and Reducer. Since we are running a sample application, the input, Mapper and Reducer are already specified. You do need to specify the output location, which will be in your bucket.

You need to specify the number of EC2 instances you want running. Since we are running a small sample application, we will just use the default values. When we need to run larger job flows, we would need to specify more number of instances.
In the Advanced Options, you would need to specify your EC2 key-pair (section 3). You can “Enable Debugging” but you have specify a path to (one of) your S3 buckets for log files. Once the fields are filled, Continue.

We do not need any Bootstrap Actions so we will “Proceed with no Bootstrap Actions”.
This will bring us to the final Review page where we need to make sure all the information for this Job Flow is accurate. If you think you are ready, “Create Job Flow” and sit back. This should be done in about 5 minutes once the job has been initiated (which in itself takes a little bit of time).

Once you are done and the EMR console tells you that the job has completed, you can see the output in your S3 bucket. Open the ‘part-xxxxx’ files to see the output of the word count.
You have just successfully completed a MapReduce job flow on AWS and are ready for large scale data analytics!