Lab 3

February 6, 2017
Big Data 2017

In this lab we will learn how to use Amazon's Elastic MapReduce to run MapReduce jobs.

First, log in to the AWS Management Console here: http://console.aws.amazon.com.

Step 1: Creating an S3 bucket

Amazon S3 is a web service that provides storage. We will use S3 to store our code, input data, and output files for mapreduce jobs that we run using AWS.

1. In the AWS Management console, click **Services** and then click **S3**.
2. Click **Create Bucket**.
3. In the **Bucket Name** field, type **emrbucket-** followed by your initials and the date (for example, emrbucket-ecc2617). You don't have to stick to this naming scheme, you could name your bucket whatever you want (but it has to be unique, lower case only, and can't contain spaces, underscores, or periods).
4. Select **US Standard** as the region.
5. Click **Create**.

Step 2: Launch an EMR cluster with a sample job

The next step is to create and launch the EMR cluster. EMR provisions Amazon EC2 instances (virtual servers) to run jobs. These EC2 instances are preloaded with Hadoop and other needed libraries.

1. Return to the AWS Management console.
2. Click **Services** and then click **EMR**.
3. Click **Create Cluster.**
4. For **Name**, type *Wordcount*.
5. Leave logging enabled. For the logging S3 folder, enter *s3://<your-bucket-name>/logs/*
6. Under **Software configuration**, leave the default options selected.
7. Under **Hardware configuration**, leave the default options selected.

8. Under **Security and access**, in the **EC2 Key Pair** field, select **Proceed without an EC2 key pair**. (In the future, if you want to SSH to the master node of your cluster, you will need to use your key pair we generated in Lab 1).
9. Click **Create cluster**.
   Your cluster is now being started. This will take around 10 minutes until the cluster is up and running and available for computation.
10. We will now run an example mapreduce job. Click the **Add step** button.
11. For **Step type**, click **Streaming program**.
12. In the **Mapper** field, enter
    s3://elasticmapreduce/samples/wordcount/wordSplitter.py
13. In the **Reducer** field, enter: *aggregate*
    Note that here we are using the default "reduce" function in Hadoop streaming.
14. In the **Input S3 location** field, enter
    s3://elasticmapreduce/samples/wordcount/input
15. In the **output S3** field, type *s3://<your-bucket-name>/output/*
16. Click **Add**.

After your cluster starts up, your Hadoop streaming job will be run. When the cluster has finished processing the sample Wordcount job, it will store the results in a folder called output in the S3 bucket you created.

You can monitor your cluster while it is running:

- Click **Cluster List** on the left.
- Click the arrow to the left of cluster name ("My cluster") to see more details.

Step 3: Viewing the output

1. Go to Amazon S3 (you can get there by returning to the AWS Management console and clicking **S3** under **Services**).
2. Click on the name of your bucket.
3. You should see two folders, logs and output. Click on the **output** folder.
4. You should see the output of the Mapreduce job in a number of files named "part-xxxxx" (there is one file per reduce task).
5. You can download each file to your computer by right-clicking it and selecting **Download.**
6. You can then open the files using your favorite text editing program.

Step 4: Terminate your cluster

It is very important that you terminate your cluster through the EMR console once you are finished working. Otherwise, it will keep running and your account will be charged!

1. Go to the EMR console.
2. Select the box to the left of your cluster and click the **Terminate** button.

Deliverable: (due Monday, February 13, 6pm)

1. Go to Amazon **S3** and click on the bucket you created for this lab.
2. Check the box on the left next to the output folder.
3. Click **Actions** and select **Make Public**. Click **Ok** on the popup.
4. In the Lab 3 assignment on NYU Classes, submit a link to your public output folder in your S3 bucket, e.g., "https://s3.amazonaws.com/<your-bucket-name>/output/". (You can just write the url in the text box).