COMP 543 Lab : AWS

1 Description

In this lab, you will:

1. Create an AWS EMR cluster
2. Upload a PySpark program onto the cluster
3. Run a Spark job
4. Check out the output on HDFS (Hadoop Distributed File System).

Note: this assumes you have previously signed up for an Amazon account. See Piazza!

2 Create your Key Pair

To connect to the cluster you will create later, you need a Key Pair as an identity.

1. Go to Amazon's AWS website (aws.amazon.com).
2. Sign in with your user name and password. Go to EC2 (you can reach EC2 from the AWS services search text box).
3. Click "key pairs".
4. Click "Create Key Pair".
5. Pick a key pair name that is likely unique to you (such as the name of your eldest child, or your last name, so that it is unlikely that you will forget it). Type it in, and click "Create".
6. This should create a ".pem" file that you can download. You will subsequently use this .pem file to connect securely to a machine over the internet.

3 Start Up a Cluster

Now it is time to create your cluster.

1. Click the AWS at the upper left of the dashboard to get back to the main menu. Then, search for or click "EMR". This stands for "Elastic Map Reduce."
2. Click "Create cluster". Choose "Go to advanced options".
3. Make sure that Hadoop and Spark are checked. Then click next.
4. On the next page, all of the defaults should be OK. For the Master node, you want an m4.large machine. If you are interested, you can find a list of all instance types at https://aws.amazon.com/ec2/instance-types/. Each m4.large machine has 4 CPU cores and 8GB of RAM. For the Core workers, you want 2 m4.large machines. Click next.
5. On the next page ("General options") everything should be OK, except you can unclick "termination protection". Click next.

6. On the next page, under EC2 key pair, it is really important to choose the EC2 key pair that you just created. This is important: if you do not do this, you won't be able to access your cluster.

7. Click "Create Cluster". Now your machines are being provisioned in the cloud! You will be taken to a page that will display your cluster status. It will take a bit of time for your cluster to come online. You can check the status of your cluster by clicking the little circular update arrow at the top right.

   Note: the very first time that create a cluster, it may take 15 minutes or more for the cluster to begin, and Amazon makes sure your account is valid. Take the opportunity to update your Facebook or chat with your neighbor. As soon as your master node changes to "bootstrapping", you are ready to go.

   Note: if you ever want to get back to the page that lists all of your EMR clusters, just click the "AWS" at the top left, then enter or click "EMR".

4 Connect to your cluster

Once your cluster is up and running, you will want to connect to the master node so that you run Spark jobs on it.

1. Click "Clusters", and then click on your cluster. First you need to make it so that you can connect via SSH. Under "Security and Access" (not a tab on the side, just a heading at the lower left), go to "Security groups for Master" and click on the link. In the new page, click on the row with Group Name = "ElasticMapReduce-master". At the bottom, click on the Inbound tab. Click on "Edit". Click "Add Rule". Then select "SSH" in the first box and "Anywhere" in the second. Click save.

2. Now you need to connect. Again go to your cluster Under "Hardware", you will see a list of "Instance Groups"... you will have two types of instance groups in your cluster... a "core" group (the workers) and the "master" group, which contains the machine that you will interact with. Click on the ID associated with the "master" group, and you will see a clickable link under "EC2 Instance Id". This is your master machine. Click on this link, and you will be taken to the EC2 dashboard where it will give you all sorts of info about your master node (if you ever want to get back to your cluster, just click on the AWS logo at the upper left of the console, which gets you back to the main menu; click EMR, and then choose your cluster and you will be back to your cluster once again). The thing that we are really interested in is the public IP. This will be a number such as 54.172.82.0.

3. Now that you have created your cluster, and identified the public IP of your master node, it is time to connect to the node and run a Spark job! To connect to your master node:

   Mac/Linux

   The following assumes that your .pem file is called MyFirstKeyPair.pem and that it is located in your working directory; replace this with the actual name and location of your file, assuming that you called your key pair something else. Type:

   ```
   chmod 500 MyFirstKeyPair.pem
   ```

   Now, you can connect to your master machine (replace "54.172.82.0" with the IP address of your own master machine):

   ```
   ssh -i "MyFirstKeyPair.pem" hadoop@54.172.82.0
   ```

This will give you a Linux prompt; you are connected to your master node.

Windows

In Windows, we'll assume that you are using the PuTTY suite of tools. First fire up PuTTYgen. Click "Load" and then in the file type drop-down menu, choose "all files". Then select "MyFirstKeyPair.pem" (your .pem file will have a different name, depending upon what you called your key pair). Then choose "save" and save your file as "MyFirstKeyPair" in an appropriate directory, where you can find it (again, use the name that you chose above; PuTTYgen will add a .ppk extension to the file you are saving) and "yes" to choose to save the file without paraphrase.

Next, fire up PuTTY. This will allow you to connect to your Amazon machine via SSH. In the left-hand side of the dialog that comes up, click "Connection" then "ssh" then "auth" and then click on "Browse" to select the private key file that you created above using PuTTYgen.

5 Run Spark jobs

Now, whether or not you are using Windows or Mac/Linux, you will have a Linux prompt to your master node. It is time to run a Spark job!

1. Download the file "topWords.py" from Canvas and save it in your work directory.

2. Transfer the PySpark program over to your master node so that you can run it.

   Mac/Linux

   In Mac or Linux, open up a terminal and go into the directory you've been working from. Type the following to fire up the secure ftp program:

   ```
   sftp -i "MyFirstKeyPair.pem" hadoop@54.172.82.0
   ```

   Then type "put topWords.py" to upload your jar. Type "exit" to exit the program.

   Windows

   Fire up WinSCP. Enter in the IP address, and the user name "hadoop", and then select the private key file created using PuTTYgen (this should be "MyFirstKeyPair.ppk" in the directory you saved). WinSCP will connect to the master, and you can use WinSCP's graphical user interface to transfer files to it. Transfer over "topWords.py".

3. Run the job! From the command line, simply type:

   ```
   spark-submit topWords.py
   ```

   This will launch the Spark job. A bunch of information will scroll by. After a few seconds, the computation is done.

4. Check out your results. Type:

   ```
   hadoop fs -ls output
   ```

   And you will see something like:

```
Found 5 items
-rw-r--r--   1 hadoop hadoop          0 2018-10-01 02:39 output/_SUCCESS
-rw-r--r--   1 hadoop hadoop      90587 2018-10-01 02:39 output/part-00000
-rw-r--r--   1 hadoop hadoop      95356 2018-10-01 02:39 output/part-00001
-rw-r--r--   1 hadoop hadoop     101621 2018-10-01 02:39 output/part-00002
-rw-r--r--   1 hadoop hadoop      92673 2018-10-01 02:39 output/part-00003
```

5. Copy the results from HDFS to the master node. Type:

   ```
   hadoop fs -get output
   ```

6. You can have a look at some of the results by typing:

   ```
   more output/part-00001
   ```

7. Show one of the graders this file to get checked off.

6 SHUT DOWN YOUR CLUSTER

Important: never leave your cluster up when you are not using it. You are being charged!

1. From the web page for your cluster, click "Terminate".
2. If "Termination Protection" is on, you will have to turn it off before you kill your machines.
3. Note: I've had mixed results actually killing machines in this way. After you kill them, make sure that they are dead. Click the cube, click "EC2" and click "Running Instances". There should not be any. If they are still there, click on "Running Instances". Then click the checkbox next to each of your machines, and under "Actions"-"Instance State" choose "terminate". Only log out after you have verified from the EC2 page that you have no running instance.