CS589: Machine Learning
Setting Up AWS for Spark on EMR

Benjamin M. Marlin

March 20, 2017

Warning: Please read these instructions carefully and completely before you start working with AWS. Amazon will automatically bill you if your charges exceed your available credits. You are solely responsible for any fees incurred.

Each UMass Amherst student can obtain $100 of Amazon Web Services (AWS) credits each year. If you have already used all of your credits for this year, please contact the course staff on Piazza. If you have not already used up your credits, but you have already obtained your credit code and applied it to your AWS account, you can skip to Step 2. If you have not obtained your credits for this year, begin at Step 0.

You must be connected to the Internet for many of the steps below. If you get errors related to downloading files or connecting to websites, verify that you have Internet access on the system you are using.

Step 0: Obtain AWS Credits

1. Go to `https://www.awseducate.com/Application` and complete the application as a student using your UMass Amherst email address. If you do not already have an Amazon account of any kind, you will need to create one. A full AWS account is recommended, but you can also use the new AWS Educate Starter Account option if you don't have access to a credit card to setup a full AWS account.
2. When your application is processed, Amazon will send you an email containing an AWS credit code if you applied using a regular AWS account. When you receive the credit code email, proceed to Step 1. If you applied for an AWS Educate Starter Account, you should be able to proceed directly to Step 3.

Step 1: Sign in, Redeem Credits

1. Go to `http://aws.amazon.com/` and log in by clicking "Sign in to Console".
2. At the top right, set your region to "N. Virginia".
3. Click your name in the top right, and then click "My Account."
4. On the bottom left of the Account page, click "Credits"
5. Enter the code you were sent in the "Promo Code" box.
6. Complete the security check and click the "Redeem" button.
7. When the page reloads, you should see $100 of AWS credits.

Step 2: Create Billing Alarm

1. Now click "Preferences" on the left of the Accounts page and check the box next to "Receive Billing Alerts."
2. In the description for "Receive Billing Alerts" click the "Manage Billing Alerts" link.
3. Click the "Create Alarm" button.
4. Select "Total Estimated Charge
5. Select "USD" and click "Next" at the bottom of the page.
6. Enter the name "No Credits" and the description "Out of credits".
7. Set the threshold to $0.01.
8. Select NotifyMe" under "Send notification to" or create a new notification list.
9. Click "Create Alarm"
10. **Warning:** Billing alerts are based on estimated costs. A threshold of 0.01 should trigger once you run out of credits, but the AWS cost estimates can lag behind actual usage. Please also keep track of usage yourself. The experiments we are running should cost a couple of dollars at most, but you are responsible for overages if you forget to shut down your cluster.

Step 3: Create SSH Key Pair

1. Go to `http://aws.amazon.com/` and log in by clicking "Sign in to Console".
2. Click "Services" in the top menu and then select "EC2."
3. At the top right, set your region to "N. Virginia".
4. In the far left menu select "Key Pairs" under "Network & Security"
5. Click "Create Key Pair."
6. Enter a name for the key pair. In this tutorial, we use the name: "CS589".
7. You will be prompted to download the private key file, "CS589.pem".
8. Move the SSH private key file "CS589.pem" from the location you downloaded it to your HW03/Code folder.
9. Finally, if you are on a Mac or Linux computer run the following command from your HW03/Code directory to set the permission on the private key file so that it is only readable by your account:

   ```
   chmod 400 CS589.pem
   ```

10. **Warning:** This is an SSH private key file. It gives the ability to log into any running instances that you have on EC2. Do not post it to any public repositories.

Step 4: Create a Spark Cluster using EMR

1. Go to `http://aws.amazon.com/` and log in by clicking "Sign in to Console".
2. At the top right, set your region to "N. Virginia".
3. Click "Services." In the search box enter "EMR" then click "EMR"
4. Click "Create cluster" and enter the following options:
    - Name cluster: 589exp
    - Disable logging
    - Launch mode: cluster
    - Vendor: Amazon
    - Release: emr-5.4.0
    - Applications: Spark
    - Instance Type: m4.large
    - Number of instances: 5
    - EC2 Key pair: CS589
    - Permissions: Default
5. Click "Create cluster" and wait until cluster has started (will take some time – 10mins). This step may fail the first time due to permissions issues. If so, try again. While you wait for this step to complete, you can configure the security group to allow access to the cluster after reading the warning below.
6. **Warning**: Your EMR cluster does not shutdown automatically if you close the log out of the cluster, close the AWS website, or suspend or shutdown your local computer. You need to follow the instructions below to terminate your EMR cluster when you're done with it, and verify that it has shut down. Otherwise, AWS will continue deducting credits and you'll eventually incur credit card charges if you're on a full AWS account or your account will be terminated if you're on an AWS Educate Starter account.

Step 5: Configure Security

1. On the EMR cluster launch page under "Security and Access," click "Security groups for ... (ElasticMapReduce-Master)."
2. Select "ElasticMapReduce-master"
3. On the tabs shown toward the bottom of the page, click "Inbound."
4. Click "Edit," then click "Add Rule."
5. Configure the rule by setting "Type" to "All traffic" and "Source" to "My IP." Note: you need to do this every time the IP address of your computer changes.
6. Click "Save" and return to the browser tab with the EMR cluster launch page.
7. When the cluster has partially started, the "Mather public DNS" address will be available. Click the "SSH" link next to the public address, and follow the instructions to connect to the cluster using SSH.

Step 6: Accessing Your Spark Cluster

1. When the cluster has partially started, the public address of the master node will be available next to "Master public DNS" label at the top of the cluster status page.
2. Click the "SSH" link next to the "Master public DNS" address, and follow the instructions to connect to the cluster using SSH.
3. For linux and Mac users, you can use the system SSH. You need to specify the location of your CS589.pem file in the SSH command. The default location that EMR will suggest will not be correct.
4. Windows users will need to follow the suggested steps to download and configure puTTY with your CS589.pem file in order to access the EMR cluster.
5. In both cases, you should have saved your CS589.pem file in your HW03/Code directory.

Step 7: Running your code

1. Once you have connected to your cluster using SSH, the quickest way to test your code is to launch the nano editor by entering "nano aws_run_me.py" on the command line.
2. Now copy all of the code from your local version of aws_run_me.py and paste it into the nano editor.
3. To save the file, press "Control-o" to save and then 'Control-x" to exit.
4. to start your code running, enter the command shown below:

```
spark-submit  --num-executors 8 --executor-cores 1  --executor-memory 2G  aws_run_me.py
```

5. aws_run_me.py will run your regression code with between 1 and 8 workers and will report the total training and test time as well as the error rate achieved. You will also get a lot of output from the Spark job scheduler as it runs.
6. If your code takes longer than 10 minutes to run with one worker, you have some type of scalability problem. Stop the computation by pressing "Control-C" and start debugging.

Step 8: Terminate EMR Cluster

1. **Warning:** When you terminate a cluster, any user-created data/code/files stored on your cluster are lost and there is no way to recover it. Make sure you copy any results off your cluster before you terminate it.
2. To terminate your cluster, return to the EMR cluster list here: `https://console.aws.amazon.com/elasticmapreduce/`, select any active clusters and click "Terminate."
3. Keep refreshing the page until you're sure your cluster has been terminated. Note that EMR resources are region-specific. The cluster list only includes clusters in the currently selected region, so make sure your region is correctly selected.

More Notes

- Like any web service, AWS does not have perfect uptime. Service regions have outages periodically and it may take time to start-up a cluster of even 5 nodes at busy times. It is not a good idea to leave work on AWS to the last minute for these reasons, especially if you are out of late days.

- Once your cluster is running, you are charged a per-instance per-hour fee. The instances we're using have two cores each and 8GB RAM. The total cost is about $0.15 per-instance, per-hour. So, your cluster of 5 nodes costs about $0.75 per hour to operate.
- All billing times are rounded up to the nearest hour per instance, so if you start and terminate a cluster with 5 nodes four times in one hour, you will be charged for $5 times 4 = 20$ hours of usage at $0.15 per hour for a total of $3.00.
- To keep track of your usage, go here: `https://console.aws.amazon.com/billing` and look at your current bill. You can see the amount of credits that have been applied and the costs you've incurred, which will automatically balance until you've spent all of your credits.
- Note that billing is not updated in real time, so your current bill will often not reflect your usage on a given day until sometime on the next or following days.
- **Once your credits are spent, full AWS account users will have their credit cards automatically billed for the balance of their usage. AWS Educate Starter account users will have their accounts terminated.**