AWS and Apache Spark

ddebarr@uw.edu, 2014-04-21

This note covers how to …

1) Set up an Amazon Web Services (AWS) account [if you don't already have one]
2) Apply for a $100 AWS educational credit
3) Set up an Apache Spark server

Set up an AWS Account

1) Navigate to https://aws.amazon.com.
2) Click the "Create an AWS Account" button, in the upper right-hand corner of the page.
3) Fill in your name@uw.edu email address, click the "I am a new user" radio button, and click the "Sign in using our secure server" button.
4) Fill in your name, your email address twice, and your new password twice; then click the "Create account" button.
5) Enter your payment details. Please keep track of your charges carefully. When you are done with your server, you should make sure you have copied off all data you wish to keep and terminate the server.

Apply for a $100 AWS Educational Credit

1) Navigate to https://aws.amazon.com/education/awseducate/apply/ and apply for your $100 AWS credit as a student of the University of Washington's Machine Learning Certificate program: https://www.pce.uw.edu/certificates/machine-learning. Getting approved for the credit may take a few days.
2) After you receive an email from AWS with your "Credit Code": navigate to https://aws.amazon.com, select "Account Settings" from the "My Account" menu (at the top of the page, on the right-hand side), select "Credits" from the navigation pane, enter your "Promo Code", and click the "Redeem" button.

Set up an Apache Spark server

1) If you already have a linux machine with a few gigabytes of memory, then you can skip to step 8.
2) Navigate to https://aws.amazon.com.
3) Select "AWS Management Console" from the "My Account" menu.
4) Under "AWS Services", click the "EC2" (Elastic Compute Cloud) link under the "Compute" heading.
5) This is the interface for creating and terminating instances:
    a. If you have a non-zero value for "Running Instances", and you wish to terminate a server, then …
        i. Click the "Running Instances" link.
        ii. Right-click the instance.
        iii. Choose "Terminate" from the "Instance State" menu.

b. If you wish to create a server [that costs around $2.75 per day], then …
    i. Click the "Launch Instance" button.
    ii. Click the "Select" button next to the "Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9abea4fb" option.
    iii. Click the radio button for Type "t2.large" [according to http://calculator.s3.amazonaws.com/index.html, this server costs $0.104 per hour to run].
    iv. Click the "Next: Configure Instance Details" button.
    v. Click the "Next: Add Storage" button.
    vi. Change the size from "8" [GiB (gigabytes)] to "64" [according to https://aws.amazon.com/ebs/pricing/, this storage costs $0.10 per GB-month of provisioned storage (around $0.010 per hour)].
    vii. Click the "Review and Launch" button.
    viii. Click the "Launch" button.
    ix. Click the "Create a new key pair" button.
    x. Enter "ml_server_login_key" for the "Key pair name". Store this file in a safe place, as this is effectively the password for your server.
    xi. Click the "Download Key Pair" button.
    xii. Click the "Launch Instances" button.

6) Click the "i" link next to the "The following instance launches have been initiated", and take note of the Domain Name Service (DNS) name for your server [found next the "Public DNS" label]. The name of my server was "ec2-IP1-IP2-IP3-IP4.us-west-2.compute.amazonaws.com".
7) Connect to your AWS server
    a. If you're using Windows to connect to your server, download and run "A Windows MSI installer package for everything except PuTTYtel" [from http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html]
    b. From the PuTTY apps menu, run "PuTTYgen" [this only needs to be done once]
        i. Click the "Load" button, browse to your "Downloads" directory, and select the "ml_server_login_key.pem" file.
        ii. Click the "Save private key" button. This creates a ".ppk" version of the private key, for use by the Putty SSH client.
    c. From a Windows command prompt, you can start the Secure Shell (SSH) client with the following command:
    "C:\Program Files (x86)\PuTTY\putty.exe" –i
    "C:\Users\dadebarr\Downloads\ml_server_login_key.ppk"
    ubuntu@DNS_name_for_server
    d. Login as username "ubuntu" (if prompted; happens when omitted from the command line).
8) Install support software with the following commands
    a. sudo apt-get update
    b. sudo apt-get install unzip
    c. sudo apt-get install default-jdk
    d. sudo apt-get install python-sklearn

9) Setup the Apache Spark server
    a. wget http://d3kbcqa49mib13.cloudfront.net/spark-1.6.1-bin-hadoop2.6.tgz
    b. gzip -dc spark-1.6.1-bin-hadoop2.6.tgz | tar xvf -
    c. cd spark-1.6.1-bin-hadoop2.6
    d. sed -e 's/rootCategory=INFO/rootCategory=ERROR/' conf/log4j.properties.template > conf/log4j.properties
    e. sbin/start-master.sh
    f. sbin/start-slave.sh spark://127.0.0.1:7077