Spark and Python With VirtualBox

Let’s learn something!

PIERIAN DATA

Python and Spark

- This lecture will walk through how to download and set-up VirtualBox with Ubuntu.
- Then we will walk through installing Spark,Python, and the Jupyter Notebook on this VirtualBox Unbtunu.

Python and Spark

- The resources for this lecture has a link to a written form of these instructions that you can reference.

Python and Spark

- First you will need to download:
    - VirtualBox
    - Ubuntu
    - Let’s show you how you can find these.

Spark and Python With VirtualBox Part 2

Let’s learn something!

PIERIAN DATA

Python and Spark

- In Part 2 we will install Python,Spark,and the Jupyter Notebook onto our Ubuntu Machine.
- If Ubuntu was already your local OS, you skipped Part 1 and came here!

Python and Spark

- A quick note:
    - Spark 2.1 is incompatible with Python 3.6
    - This will be fixed when Spark 2.2 is released, in the meantime, we'll stick to Python 3.5 to avoid issues!

DataBricks Setup

Let’s learn something!

PIERIAN DATA

Python and Spark

- Databricks is a company that provides clusters that run on top of AWS and adds a convience of having a Notebook System already set up and the ability to quickly add files.

Python and Spark

- It has a free community version that supports a 6 GB cluster.
- It also has its own storage format known as DBFS.
- This “Table” format needs to be accessed in a particular way.

Python and Spark

- We’re going to walk through how to setup a Databricks account and how you can upload data and set it as a DataFrame.
- I recommend Databricks for people who want to quickest online setup.

Python and Spark

- Go to:

  https://databricks.com/try-databricks

AWS EMR Setup

Let’s learn something!

PIERIAN DATA

Python and Spark

- If you want to quickly set up a cluster with a Notebook Interface, AWS EMR is a good choice.
- Please note, what we will show in this lecture **does <u>not</u> fall under the free trial of AWS.**

Python and Spark

- In this lecture we will will walk through setting up the Zeppelin Notebook.
- We will also discuss some security options, I highly recommended you watch the EC2 Instance Lectures first before watching this to learn about SSH!

Python and Spark

- The Zeppelin Notebook is a fairly new environment that mimics Jupyter Notebook's capabilities but was created specifically with Big Data (Spark, Hadoop, etc...) in mind.

Python and Spark

- Let’s quickly explore what the Zeppelin Notebook looks like on their official docs and then go through the process of setting up our own on AWS EMR.
- Creating the cluster on AWS EMR can take awhile to intialize!

Python and Spark

- We'll also talk about security settings as we set-up the Zeppelin Notebook running on EMR, make sure to read through the resource documentation to choose the best security settings for you or your organization!

Python and Spark

- Let’s get started and jump into the docs:
  - zeppelin.apache.org
- Afterwards we will log into to our AWS:
  - aws.amazon.com

Spark and Python on AWS EC2

PIERIAN DATA

Python and Spark

- This lecture will walk through how to set-up Python, Spark, and a Jupyter Notebook on AWS EC2
- Before we begin let's discuss a few things to keep in mind!

Python and Spark

- This is by far the most tedious set-up option out of the four.
- While everything we will show is within the free (one-year) tier for AWS, you still need a credit card to set up an account.

Python and Spark

- If you create an EC2 instance using different parameters than what is shown here, you may be liable for charges!
- Make sure to follow the directions shown in this video exactly, otherwise you will have to repeat the process all over again.

Python and Spark

- Leave yourself plenty of time to go through this process.
- If you feel uncomfortable with any of this, just go to the VirtualBox installation lectures, those are local, simpler, and 100% free!

Python and Spark

- Overall, this EC2 process is not too bad, but keep in mind that if you get an error during this process, it is because you made a mistake somewhere not following with the video lectures!
- Let’s get started walking you through the process of setting up!

AWS Account Set-Up

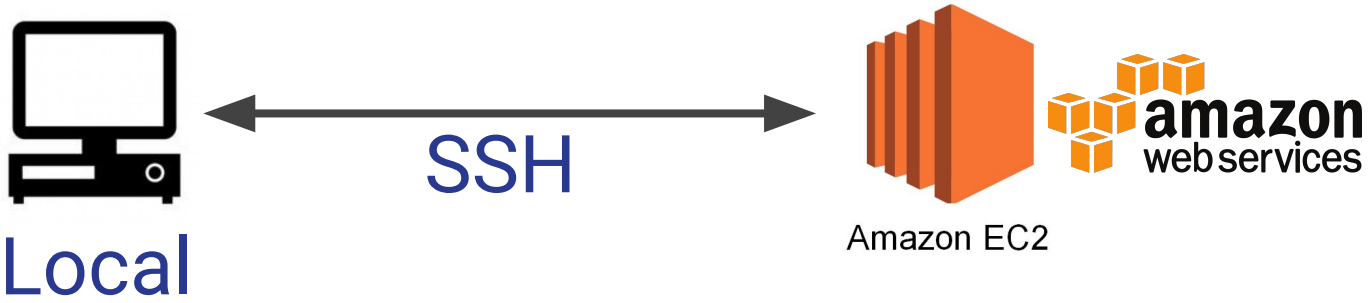
- Go to:
  - https://aws.amazon.com/free
- Then click on Create Free Account
- Sign up with an email address
- Then fill out the profile information

- The profile information:
  - Contact Information
  - Billing Information
  - ID Verification
  - Choose free support plan
- Next lecture we will explore AWS and create an EC2 instance

EC2 Instance Set-up

- Now that we have our AWS account we will create an EC2 Instance.
- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud.
- We can basically think of it as a virtual computer we can access through the internet.

- So here is our plan:
  - Create EC2 Instance on AWS
  - Use SSH to connect to EC2 over internet
  - SSH is different for Windows vs Mac/Linux
  - Set-up Spark and Juptyer on EC2 Instance



- SSH (Secure Shell Connection)
  - Watch this lecture all the way through for Windows
  - Skip to next lecture after EC2 Set-up for Mac/Linux
- Our goal is to remotely connect to the command line of our virtual machine on Amazon EC2

Login to your AWS console at: aws.amazon.com

PySpark Set-up

SSH with Mac and Linux

- Skip this lecture if you are on Windows, you should have connected to your instance already from the previous lecture
- We've created our EC2 instance using AWS console
- You should have also downloaded the .pem file
- Now we are going to connect to our instance through our terminal using SSH

- Make sure you can locate your .pem file
  - Recommend you relocate it to your Desktop
- Make sure you have the DNS address of your EC2 instance
- Check the resource link for the step by step instructions from Amazon
- Let’s get started by opening our terminal

Installations on EC2

Python and Spark

- By now you should have been able to SSH into your EC2 instance (PuTTY for Windows, directly for Mac/Linux)
- Everything we'll do now will be directly through this command line interface.
- Make sure to follow along carefully!

Python and Spark

- Our tasks for this EC2 instance:
    - Download and Install Spark
    - Install Jupyter Notebook
    - Connect with PySpark
    - Access EC2 Juptyer Notebook in our local browser!