Installation Options

Let’s get you all set up!

Python and Spark

- It is now time to get your system setup with Python and Spark!
- This section has lectures covering four installation options, you are free to choose whichever you want (or go with your own personal setup)

Python and Spark

- All of the installation options will work on any OS, because we will either link you online to a Linux based system, or use VirtualBox to set-up a Linux based system locally.
- Let’s explain why we do this.

Python and Spark

- Realistically Spark won't be running on a single machine, it will run on a cluster on a service, like AWS.
- These cluster services will pretty much always be a Linux based system.

Python and Spark

- Understanding the very basics of setting up something through a Linux Command Line is essential to getting Spark going in the “real-world”.
- We will go over various options based on Linux (instead of Windows or Mac).

Python and Spark

- But to be clear-all these options will work for you regardless of your Operating System.
- They’re Linux (Ubuntu) based, and will work either locally or through an online connection.

Python and Spark

- The four methods we will cover:
    - Ubuntu+Spark+Python on VirtualBox
    - Amazon EC2 with Python and Spark
    - Databricks Notebook System
    - AWS EMR Notebook **(not free!)**
- Let's explain each option!

Python and Spark

- Ubuntu on a VirtualBox
- This option will setup a VirtualBox on your local computer (any OS) and then walk through having Ubuntu,Spark, and Python all installed locally on this virtual machine.

Python and Spark

- Amazon EC2 with Python and Spark
- This will walk through setting up a free “micro” instance on AWS EC2 which you can connect to online through SSH.
- This falls under the AWS 1-year free trial limits.

Python and Spark

- Databricks Notebook System
- Databricks is a company founded by the creator of Spark.
- Currently they have a freely hosted Notebook platform that supports a variety of Spark APIs.

Python and Spark

- AWS EMR Notebook
- The AWS Elastic MapReduce Platform allows you to quickly set up clusters.
- This is **not** a free service.
- But it allows for a very quick setup of a large cluster.

Python and Spark

- Jump to the curriculum page and choose the installation lecture you prefer.
- You don’t need to go through all of them.
- Personally, I recommend the VirtualBox so you can run everything locally.

Python and Spark

- Don’t feel restricted to just these 4 options!
- The code we work through will work on any platform that has support for Spark and Python.

Python and Spark

- Let’s get started!
- We’ll begin auto-playing through the installation lectures, jump to whichever lecture you want!
- Leave yourself 20-30 minutes for the setup, although some are much faster!