**Installation Guide for Spark on MAC (draft version)**

Dependent on your Python set-up, there are a couple of possible approaches you might take to installing Spark on your local machine. Below we successfully tested both **manual** **download** and **command line** approaches. Please consider in the context of your own preferred Python set-up, other dependencies, etc. At the end, we provide a couple of links to methods described for Windows machines; we have not been able to test any of these.

***Method A - Installing Spark via manual download***

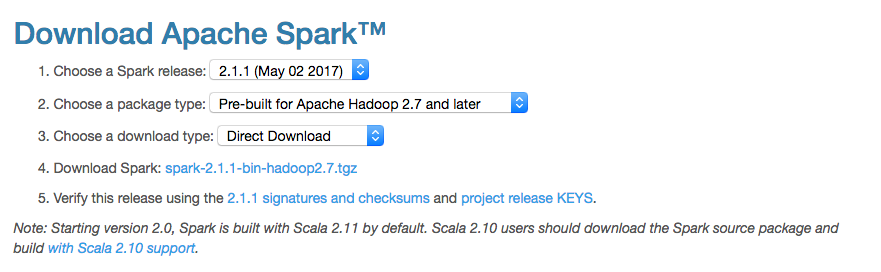
1. Make sure that java is installed on the computer. Java on IBM computers can be installed from Mac@IBM app store. Install the following from utilities:



1. Go to the official Apache Spark download website:

<http://spark.apache.org/downloads.html>

1. Fill in the following selection types:



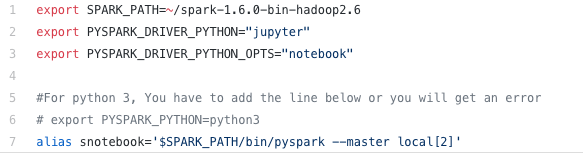
1. Click on [spark-2.1.1-bin-hadoop2.7.tgz](https://d3kbcqa49mib13.cloudfront.net/spark-2.1.1-bin-hadoop2.7.tgz)
2. This will trigger the direct download. Once this download is complete, navigate inside of the folder where the ‘spark-2.1.1-bin-hadoop2.7.tgz’ file has been downloaded.
3. Go to your home directory using the following command: **cd ~**
4. Use **pwd** to check for the path for the home directory
5. Use the **mv** command on the terminal to move this file from where it was downloaded to your home directory.

‘**mv /path/to/sparkfile /home/directory/.’**

1. Unzip the spark folder in your home directory using the following command:

**‘tar –zxvf spark-2.1.1-bin-hadoop2.7.tgz’**

1. Use the command **‘ls -a’** to check if the .bash\_profile file is present.
2. Next edit the .bash\_profile file using any editor (e.g. emacs, nano, vi)
3. Do not remove anything from the .bash\_profile file and add the following lines:



\*\*\* Replace the SPARK\_PATH with the name of the folder that was downloaded

\*\* The alias ‘snotebook’ can be replaced with any name

1. Type the following the terminal once you exit the editor:

**‘source .bash\_profile’**

1. Run the alias command to launch your jupyter notebook from anywhere on the terminal.

These steps were taken from the following link which is an extremely helpful tutorial for the installation of spark. It also contains a YouTube video tutorial:

<https://medium.com/@GalarnykMichael/install-spark-on-mac-pyspark-453f395f240b>

<https://www.youtube.com/watch?v=I5JtvpyM14U>

***Method B - Installing Spark using Homebrew (via command line)***

This method assumes you use Homebrew (a package manager for Mac) to install code packages on your machine, and that you have already installed a specific version of Python for code development i.e., that you do **NOT** use the default Apple-installed or Anaconda Python versions.

1. First, ensure that you have the latest **Java JDK** installed on your system (can be installed from Mac@IBM store). To check, at the terminal type: **java –version**
2. In preparation for Spark installation, you may also want to install the latest version of **Scala** on your machine. For example, to install using Homebrew, type:

**brew install scala**

1. Now install **Spark**. For example, to install using Homebrew, type:

**brew install apache-spark**

To check that this installed correctly, try running spark by typing:

**spark-shell**

This runs Spark in the terminal, using Scala as the coding language. To run using either Python or R, use the commands: **pyspark** or **sparkR.**

1. While you can now run **pyspark** from the Terminal, you will want to link **pyspark** directly to Jupyter notebook (or Spyder), for example. There are likely several ways to do this; the following method has been tested and seems to work.

First, locate the **/conf** directory in the installed pyspark package (if you installed via Homebrew, this should be located in **/usr/local/Cellar/apache-spark/2.1.1/libexec/**).

Then, create a file called ‘spark-env.sh’ inside this folder (copy/paste the spark-env.sh.template file already located there). Inside this environment file, add the following lines (modify the ‘python3’ variable appropriately for your own set-up):

**PYSPARK\_PYTHON=python3**

**PYSPARK\_DRIVER\_PYTHON=jupyter**

**PYSPARK\_DRIVER\_PYTHON\_OPTS=notebook**

1. Typing **pyspark** at the Terminal should now launch a Jupyter notebook in your browser. First check that pyspark is available in Jupyter, by running: **import pyspark.** Then test that everything is working fine, by running the following short test script (should output ~3.14):

**import pyspark**

**import random**

**num\_samples = 100000000**

**def inside(p):**

**x, y = random.random(), random.random()**

**return x\*x + y\*y < 1**

**count = sc.parallelize(range(0, num\_samples)).filter(inside).count()**

**pi = 4 \* count / num\_samples**

**print(pi)**

**sc.stop()**

***Links for Windows installation***

The Windows installation routine may overlap to some extent with the manual installation for Mac, in that you probably want to consider pre-installing Java and Scala, and you want to have a good sense of where your Python distribution came from and how it is organized.

The links below give some suggested routines for set-up of Spark locally on Windows. Please explore and document here on whatever method you take to installing Spark locally on Windows. You may be able to ignore some of the steps outlined (e.g., SBT or Hadoop steps) given that you are installing for local coding.

<http://nishutayaltech.blogspot.com/2015/04/how-to-run-apache-spark-on-windows7-in.html>

<https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c>