[Activity] Getting Set Up: Installing Python, a JDK, Spark, and its Dependencies.

# [Activity] Getting Set Up: Installing Python, a JDK, Spark, and its Dependencies.

**WARNING! Do not install Java 16, 17, or newer** while setting up your system. Apache Spark is only compatible with Java 8 or 11 right now. Java 11 may be downloaded from <https://www.oracle.com/java/technologies/javase-jdk11-downloads.html>

Resources for this lecture:

[winutils.exe](http://media.sundog-soft.com/Udemy/winutils.exe)

[Apache Spark](https://spark.apache.org/downloads.html)

[JDK](http://www.oracle.com/technetwork/java/javase/downloads/index.html)

# Installing Apache Spark and Python

## **Windows: (keep scrolling for MacOS and Linux)**

1. Install a JDK (Java Development Kit) from <http://www.oracle.com/technetwork/java/javase/downloads/index.html> [.](http://www.oracle.com/technetwork/java/javase/downloads/index.html) **You must install the JDK into a path with no spaces**, for example c:\jdk. Be sure to change the default location for the installation! **DO NOT INSTALL JAVA 16. SPARK IS ONLY COMPATIBLE WITH JAVA 8 OR 11.**
2. Download a **pre-built** version of **Apache Spark 3** from <https://spark.apache.org/downloads.html>
3. If necessary, download and install WinRAR so you can extract the .tgz file you downloaded. <http://www.rarlab.com/download.htm>
4. Extract the Spark archive, and copy its **contents** into **C:\spark**after creating that directory. You should end up with directories like c:\spark\bin, c:\spark\conf, etc.
5. Download winutils.exe from [https://sundog–s3.amazonaws.com/winutils.exe](https://sundog-spark.s3.amazonaws.com/winutils.exe) [a](https://sundog-spark.s3.amazonaws.com/winutils.exe)nd move it into a **C:\winutils\bin** folder that you’ve created. (note, this is a 64-bit application. If you are on a 32-bit version of Windows, you’ll need to search for a 32-bit build of winutils.exe for Hadoop.)
6. Create a **c:\tmp\hive** directory, and cd into c:\winutils\bin, and run **winutils.exe chmod 777 c:\tmp\hive**
7. Open the the **c:\spark\conf** folder, and make sure “File Name Extensions” is checked in the “view” tab of Windows Explorer. Rename the log4j.properties.template file to log4j.properties. Edit this file (using Wordpad or something similar) and change the error level from INFO to ERROR for log4j.rootCategory
8. Right-click your Windows menu, select Control Panel, System and Security, and then System. Click on “Advanced System Settings” and then the “Environment Variables” button.
9. Add the following new USER variables:
   1. SPARK\_HOME c:\spark
   2. JAVA\_HOME (the path you installed the JDK to in step 1, for example C:\JDK)
   3. HADOOP\_HOME c:\winutils
   4. PYSPARK\_PYTHON python
10. Add the following paths to your PATH user variable:

**%SPARK\_HOME%\bin**

**%JAVA\_HOME%\bin**

1. Close the environment variable screen and the control panels.
2. Install the latest **Anaconda for Python 3** from anaconda.com. Don’t install a Python 2.7 version! If you already use some other Python environment, that’s OK – you can use it instead, as long as it is a Python 3 environment.
3. Test it out!
   1. Open up your Start menu and select “Anaconda Prompt” from the Anaconda3 menu.
   2. Enter **cd c:\spark** and then **dir** to get a directory listing.
   3. Look for a text file we can play with, like README.md or CHANGES.txt
   4. Enter **pyspark**
   5. At this point you should have a >>> prompt. If not, double check the steps above.
   6. Enter **rdd = sc.textFile(“README.md”)** (or whatever text file you’ve found) Enter **rdd.count()**
   7. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
   8. Enter **quit()** to exit the spark shell, and close the console window
   9. You’ve got everything set up! Hooray!

## **MacOS**

### **Step 1: Install Apache Spark**

#### **Method A: By Hand**

If you’ve never used “homebrew,” this might be the better way to go for you. The best setup instructions for Spark on MacOS are at the following link:

[**https://medium.com/luckspark/installing-spark-2-3-0-on-macos-high-sierra-276a127b8b85**](https://medium.com/luckspark/installing-spark-2-3-0-on-macos-high-sierra-276a127b8b85)

Spark 2.3.0 is no longer available, but the same process should work with 2.4.4 or 3.x.

#### **Method B: Using Homebrew**

1. Install Homebrew if you don’t have it already by entering this from a terminal prompt: /usr/bin/ruby -e “$(curl -fsSL [https://raw.githubusercontent.com/Homebrew/install/master/install)”](https://raw.githubusercontent.com/Homebrew/install/master/install)
2. Enter **brew install apache-spark**
3. Create a log4j.properties file via
   1. cd /usr/local/Cellar/apache-spark/2.0.0/libexec/conf  (substitute 2.0.0 for the version actually installed)
   2. cp log4j.properties.template log4j.properties
4. Edit the log4j.properties file and change the log level from INFO to ERROR on log4j.rootCategory.It’s OK if Homebrew does not install Spark 3; the code in the course should work fine with recent 2.x releases as well.

### **Step 2: Install Anaconda**

Install the latest Anaconda for Python 3 from anaconda.com

### **Step 3: Test it out!**

1. Open up a terminal
2. cd into the directory where you installed Spark, and then **ls**to get a directory listing.
3. Look for a text file we can play with, like README.md or CHANGES.txt
4. Enter **pyspark**
5. At this point you should have a >>> prompt. If not, double check the steps above.
6. Enter **rdd = sc.textFile(“README.md”)** (or whatever text file you’ve found) Enter **rdd.count()**
7. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
8. Enter quit() to exit the spark shell, and close the terminal window
9. You’ve got everything set up! Hooray!

## **Linux**

1. Install Java, Scala, and Spark according to the particulars of your specific OS. A good starting point is [http://www.tutorialspoint.com/apache\_spark/apache\_spark\_installation.htm (](http://www.tutorialspoint.com/apache_spark/apache_spark_installation.htm)but be sure to install Spark 2.4.4 or newer)
2. Install the latest **Anaconda for Python 3** from anaconda.com
3. Test it out!
   1. Open up a terminal
   2. cd into the directory you installed Spark, and do an ls to see what’s in there.
   3. Look for a text file we can play with, like README.md or CHANGES.txt
   4. Enter **pyspark**
   5. At this point you should have a >>> prompt. If not, double check the steps above.
   6. Enter **rdd = sc.textFile(“README.md”)** (or whatever text file you’ve found) Enter **rdd.count()**
   7. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
   8. Enter quit() to exit the spark shell, and close the console window
   9. You’ve got everything set up! Hooray!