Sundog Education with Frank Kane

TAMING BIG DATA WITH APACHE SPARK AND PYTHON – GETTING STARTED

Join the Community

If you're on Facebook, you're invited to join the Facebook Group for this course! It's a great way to stay connected with your fellow students and collaborate.

https://www.facebook.com/groups/1026867157456171

Installing Apache Spark and Python

Windows

1. Install a JDK (Java Development Kit) from 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 9 – INSTALL JAVA 8**. Spark is not compatible with Java 9.
2. Download a **pre-built** version of Apache Spark 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 and 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. 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

7. 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.
8. 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
9. Add the following paths to your PATH user variable:

%SPARK_HOME%\bin

%JAVA_HOME%\bin

10. Close the environment variable screen and the control panels.
11. Install the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default Don't install a Python 2.7 version!
12. Test it out!
    1. Open up Canopy and select "Canopy Command Prompt" from the Tools 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

1. Install Apache Spark 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)"
    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.
2. Install the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default
3. Test it out!
    1. Open up a terminal
    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 terminal window
9. You've got everything set up! Hooray!

JX

:all Java, Scala, and Spark according to the particulars of your specific OS. A good starting point is
p://www.tutorialspoint.com/apache_spark/apache_spark_installation.htm (but be sure to install Spark 2.0 or
ver)
:all the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default 3. Test
ut!
Open up a terminal
Enter **cd c:\spark** and then **dir** to get a directory listing.
Look for a text file we can play with, like README.md or CHANGES.txt
Enter **pyspark**
At this point you should have a >>> prompt. If not, double check the steps above.
Enter **rdd = sc.textFile("README.md")** (or whatever text file you've found) Enter **rdd.count()**
You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!

ø. Enter quit() to exit the spark shell, and close the console window
9. You've got everything set up! Hooray!

Course Materials

On Udemy, you'll find the materials attached to each lecture as resources. If you'd like to get them all at once, you can grab it all from http://media.sundog-soft.com/Udemy/SparkCourse.zip

Optional: Join Our List

Join our low-frequency mailing list to stay informed on new courses and promotions from Sundog Education. As a thank you, we'll send you a **free course on Deep Learning and Neural Networks with Python**, and **discounts on all of Sundog Education's other courses!** Just click the button to get started.