Spark installation

Tuesday, August 28, 2018 12:31 PM

Use cases :-
https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa

https://github.com/apache/spark/tree/master/examples/src/main/python/mllib

https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a

Instructions
Cd c:\spark>

Spark-submit <python spark code>

https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a

```
(base) C:\Users\anshuman_mahapatra>cd C:\SparkCourse

(base) C:\SparkCourse>spark-submit spark-linear-regression-mycode.py
```

```
(base) C:\Users\anshuman_mahapatra>cd c:\spark

(base) c:\spark>dir
```

```
(base) c:\spark>pyspark
Python 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:32:41) [MSC v.1900
Type "help", "copyright", "credits" or "license" for more information.
2018-09-25 13:19:14 WARN  NativeCodeLoader:62 - Unable to load native-hadoo
Java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLo
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/
```

Best Method

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 or 10 – INSTALL JAVA 8**. Spark is not compatible with Java 9 or newer.
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.exea 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. 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:
10. SPARK_HOME c:\spark
11. JAVA_HOME (the path you installed the JDK to in step 1, for example C:\JDK)
12. HADOOP HOME c:\winutils
13. Add the following paths to your PATH user variable:
    %SPARK_HOME%\bin

    %JAVA_HOME%\bin
14. Close the environment variable screen and the control panels.
15. Install the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default Don't install a Python 2.7 version!
16. Test it out!
17. Open up Canopy and select "Canopy Command Prompt" from the Tools menu.
18. Enter **cd c:\spark** and then **dir** to get a directory listing.
19. Look for a text file we can play with, like README.md or CHANGES.txt
20. Enter **pyspark**
21. At this point you should have a >>> prompt. If not, double check the steps above.
22. Enter **rdd = sc.textFile("README.md")** (or whatever text file you've found)
    Enter **rdd.count()**
23. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
24. Enter **quit()** to exit the spark shell, and close the console window
25. 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

Method B: Using Homebrew

26. 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)"
27. Enter **brew install apache-spark**
28. Create a log4j.properties file via
29. cd /usr/local/Cellar/apache-spark/2.0.0/libexec/conf (substitute 2.0.0 for the version actually installed)
30. cp log4j.properties.template log4j.properties
31. Edit the log4j.properties file and change the log level from INFO to ERROR on log4j.rootCategory.

Step 2: Install Canopy

Install the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default

Step 3: Test it out!

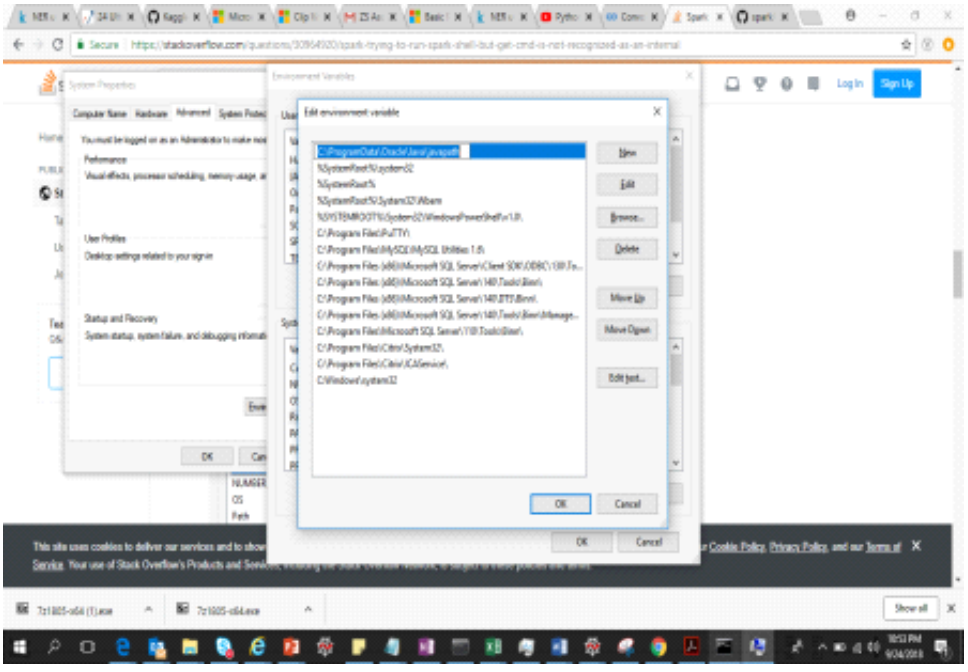
32. Open up a terminal
33. cd into the directory where you installed Spark, and then **ls** to get a directory listing.
34. Look for a text file we can play with, like README.md or CHANGES.txt
35. Enter **pyspark**
36. At this point you should have a >>> prompt. If not, double check the steps above.
37. Enter **rdd = sc.textFile("README.md")** (or whatever text file you've found) Enter **rdd.count()**
38. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
39. Enter quit() to exit the spark shell, and close the terminal window
40. You've got everything set up! Hooray!

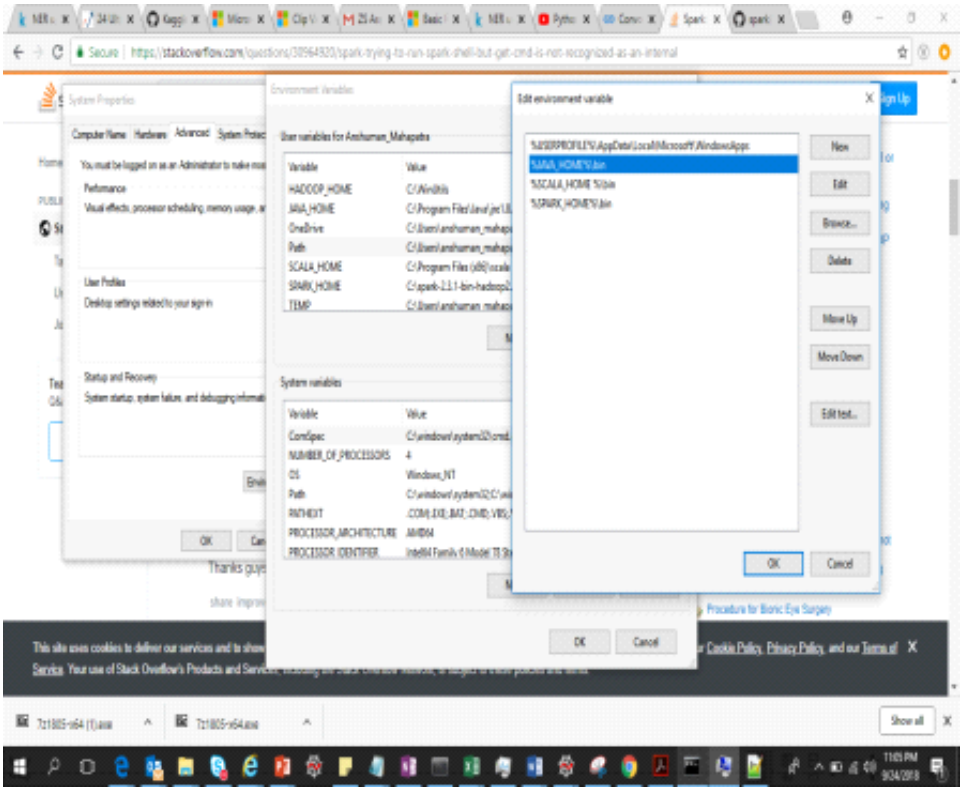
Linux

41. 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 (but be sure to install Spark 2.0 or newer)
42. Install the latest **Enthought Canopy for Python 3.5** from https://store.enthought.com/downloads/#default3. Test it out!
43. Open up a terminal
44. cd into the directory you installed Spark, and do an ls to see what's in there.
45. Look for a text file we can play with, like README.md or CHANGES.txt
46. Enter **pyspark**
47. At this point you should have a >>> prompt. If not, double check the steps above.
48. Enter **rdd = sc.textFile("README.md")** (or whatever text file you've found) Enter **rdd.count()**
49. You should get a count of the number of lines in that file! Congratulations, you just ran your first Spark program!
50. Enter quit() to exit the spark shell, and close the console window
51. You've got everything set up! Hooray!

From <https://sundog-education.com/spark-python/>

https://guendouz.wordpress.com/2017/07/18/how-to-install-apache-spark-on-windows-10/
https://medium.com/@loldja/installing-apache-spark-pyspark-the-missing-quick-start-guide-for-windows-ad81702ba62d




Edit environment variable
%SCALA_HOME%\bin
%SPARK_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel