Pyspark 安裝教學

Step 1:

下載安裝 python3.5.x 或 python 3.6.x，並加入環境變數 (建議不要用 Anaconda 下載)

Step 2:

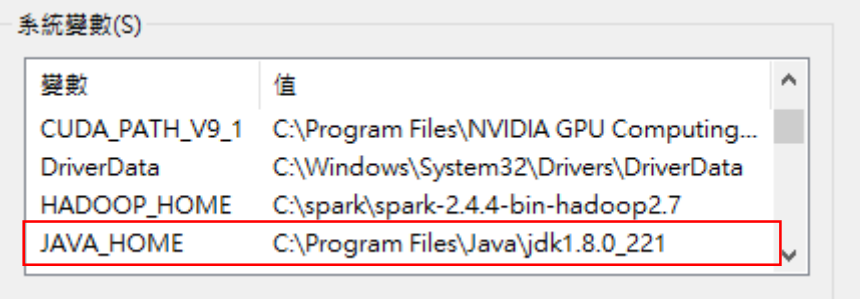
(1) 安裝 Java 8 (注意 Java 版本)

https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

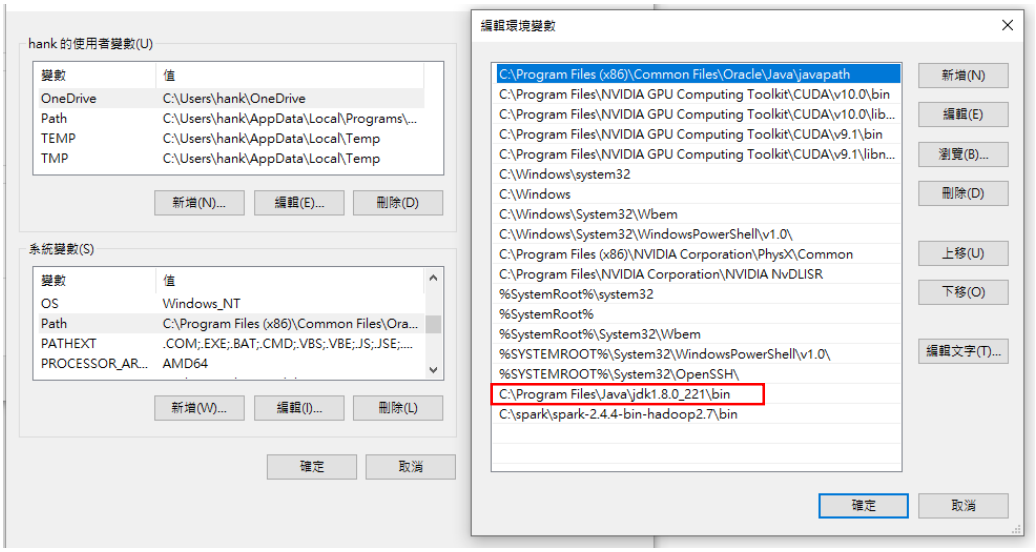
Windows x86	202.73 MB	jdk-8u221-windows-i586.exe
Windows x64	215.35 MB	jdk-8u221-windows-x64.exe

(2) 設定環境變數(JAVA_HOME & path)

設定 JAVA_HOME 環境變數，jdk 位置請設定為自己下載的 jdk 的位置。

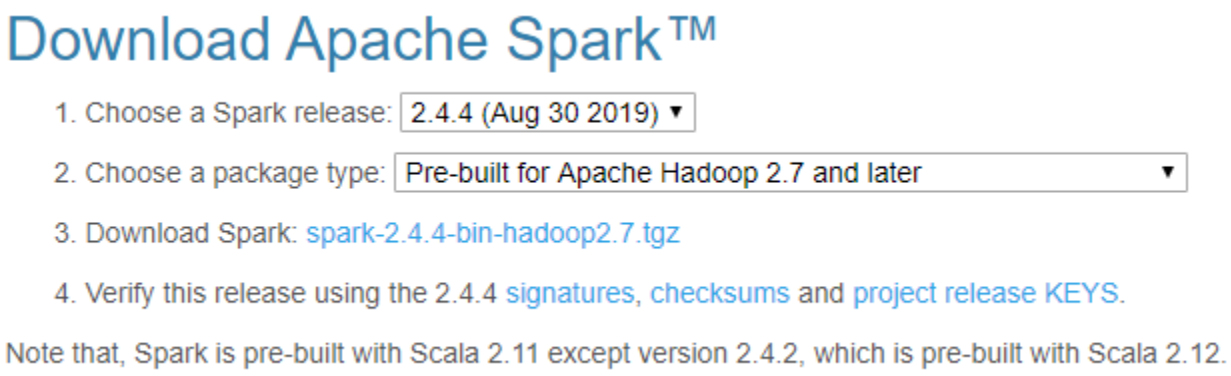


設定 path 環境變數，位置設為 jdk 裡面的 bin 資料夾。



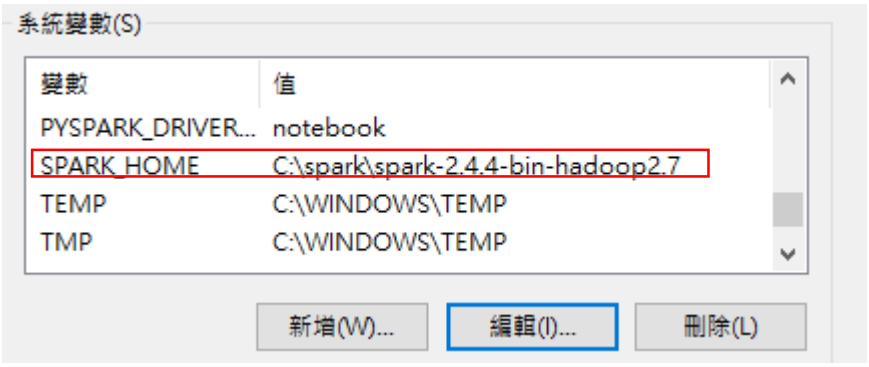
Step 3:

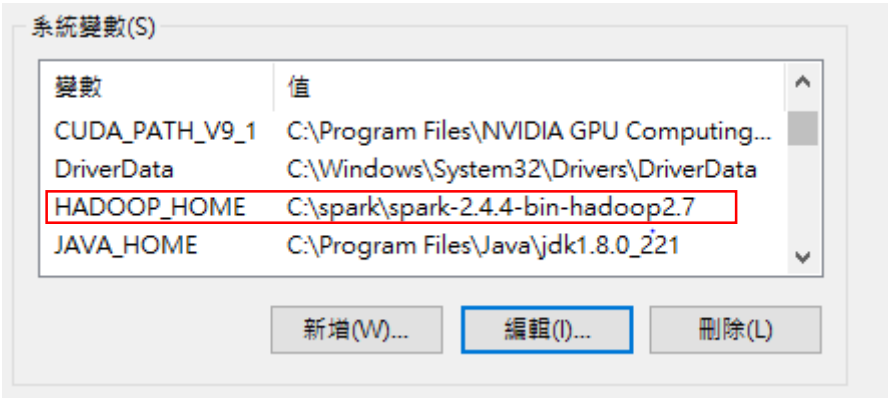
(1) 下載 spark，http://spark.apache.org/downloads.html



(2) 設定環境變數(SPARK_HOME & HADOOP_HOME)

位置請改為自己下載 Spark 的

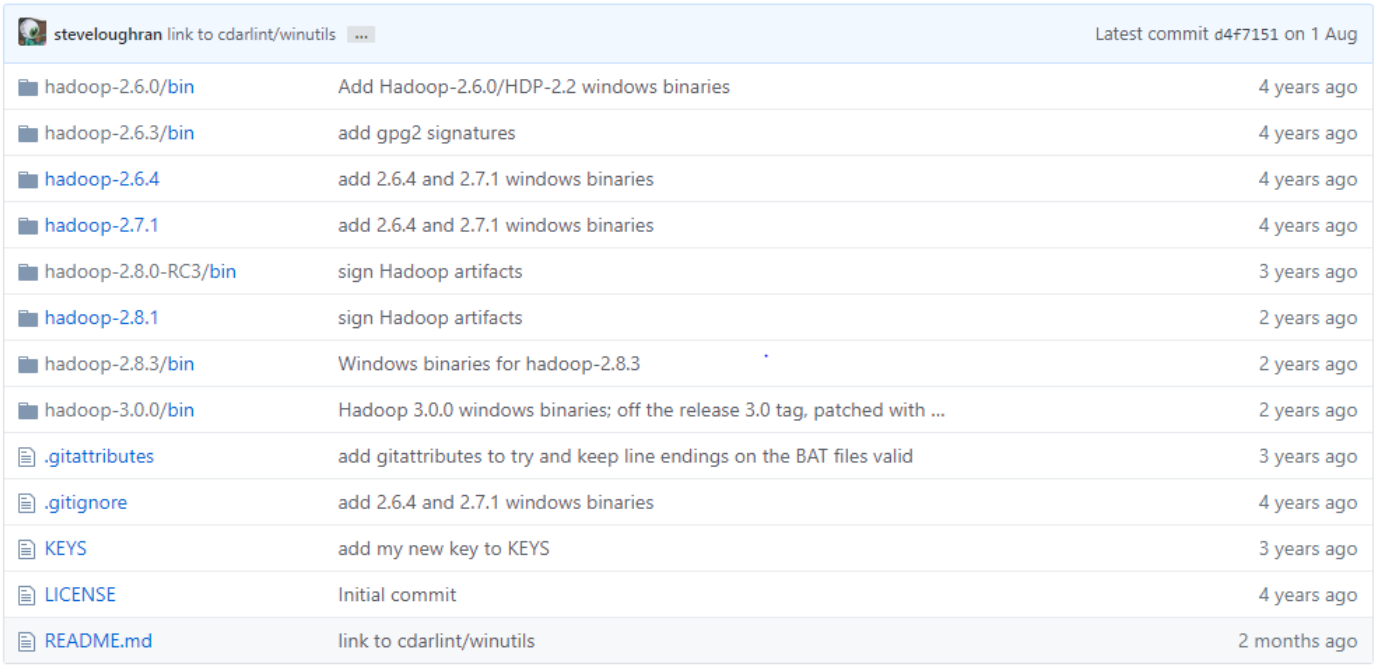




Step 4:

Windows 請下載對應的 hadoop 版本的 winutils.exe

https://github.com/steveloughran/winutils



steveloughran link to cdarlint/winutils ...		Latest commit d4f7151 on 1 Aug
hadoop-2.6.0/bin	Add Hadoop-2.6.0/HDP-2.2 windows binaries	4 years ago
hadoop-2.6.3/bin	add gpg2 signatures	4 years ago
hadoop-2.6.4	add 2.6.4 and 2.7.1 windows binaries	4 years ago
hadoop-2.7.1	add 2.6.4 and 2.7.1 windows binaries	4 years ago
hadoop-2.8.0-RC3/bin	sign Hadoop artifacts	3 years ago
hadoop-2.8.1	sign Hadoop artifacts	2 years ago
hadoop-2.8.3/bin	Windows binaries for hadoop-2.8.3	2 years ago
hadoop-3.0.0/bin	Hadoop 3.0.0 windows binaries; off the release 3.0 tag, patched with ...	2 years ago
.gitattributes	add gitattributes to try and keep line endings on the BAT files valid	3 years ago
.gitignore	add 2.6.4 and 2.7.1 windows binaries	4 years ago
KEYS	add my new key to KEYS	3 years ago
LICENSE	Initial commit	4 years ago
README.md	link to cdarlint/winutils	2 months ago

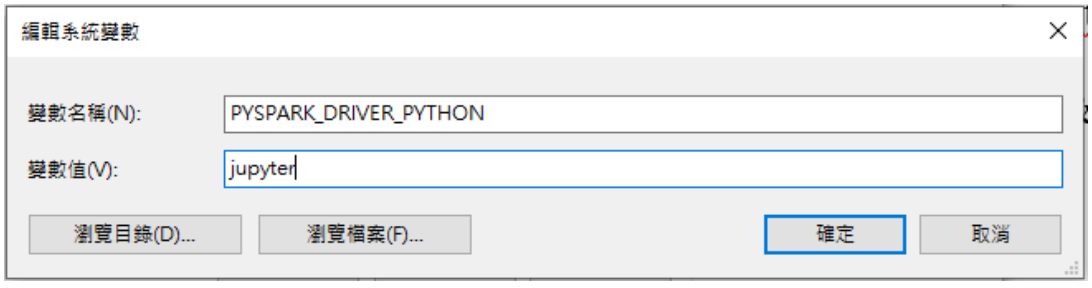
下載後放入 spark 中的 bin 資料夾內

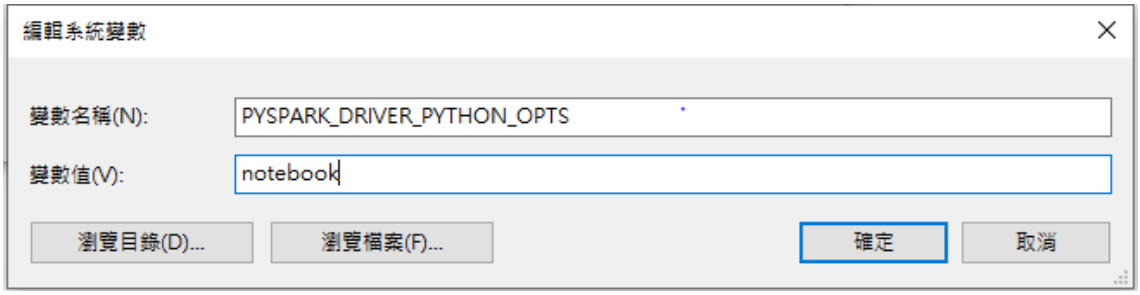
Step 5:

安裝 jupyter notebook，cmd 輸入 pip install jupyter

設定環境變數(PYSPARK_DRIVER_PYTHON &

PYSPARK_DRIVER_PYTHON_OPTS)





Step 6:

cmd 輸入 pyspark 會開啟 jupyter，可測試是否安裝成功