**Install Spark on your windows machine**

**0. Pre-requirement: Install Java in your environment.**

Please go to C:\Program Files to find the folder java. If you cannot find the java folder or the java version is not 1.8, you need to download Java JRE from

<https://www.java.com/en/download/>

**1. Download Spark 2 installation.**

<http://spark.apache.org/downloads.html> You may select the last build. Spark 3 was just published. Let’s use Spark 2.4.6 with prebuild Hadoop 2.7. Just click the download link (shown in “3. Download Spark: [spark-2.4.6-bin-hadoop2.7.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-2.4.2/spark-2.4.2-bin-hadoop2.7.tgz)”), and you will go to the Spark download page and click the http link to download Spark. It will take some time to download Spark.

**2. Unzip and extract your download into a local folder.**

As the installation is packaged by gzip then tar, you will need to unpack it using any zip tools to get another zip file "spark-2.4.6-bin-hadoop2.7" and unzip the file again to get all materials. One of the zip tools you can use is 7 zip (<https://www.7-zip.org/download.html>). If you use 7 zip, to unzip the zip file, you need to right click the file and select 7 zip -> “Extract files …”. You can now cut/paste the second spark-2.4.2-bin-hadoop2.7 folder that contains bin to your preferred location on the hard disc. Now, the bin folder of Spark will be in the folder C:\somewhere\spark-2.4.6-bin-hadoop2.7

**3. Download / Install Hadoop 2.7.1 binary for windows 10**

Download Hadoop binary (winutils.exe) from <https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>. Put the downloaded winutils.exe into a folder C:\somewhere\hadoop-2.7.1\bin\. You need to first create this folder and then add winutils.exe to it. (Please note, you need to have a bin folder under Hadoop-2.7.1 and put the winutils.exe file in the bin folder).

**4. Configure your environment.**

4.1 From search box next to the windows logo in the lower left corner of your desktop, type: "view advanced system settings” -> open “advanced system settings” ->click the button of "Environment Variables"

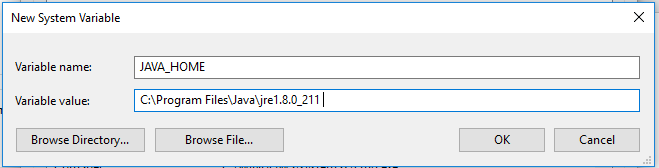
4.2 Add/change below environment variables according to your software version and location.

JAVA\_HOME= C:\Program Files\Java\jre1.8.0\_211 (where your java is installed)

SPARK\_HOME=C:\somewhere\spark-2.4.6-bin-hadoop2.7 (where your spark bin folder can be found)

HADOOP\_HOME=C:\somewhere\hadoop-2.7.1 (where the bin folder that contains winutils.exe can be found)

To add the environment variables, you click “New…” below the “System variables” box and input the variable name such as “JAVA\_HOME” into the variable name textbox and the path into the variable value box, and then click OK. Please see the example below.

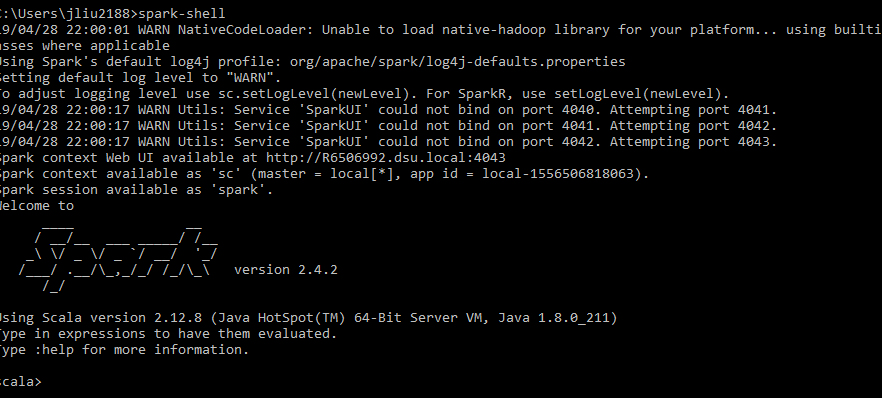


4.3 Append below variable into "Path":

Click the variable “Path” in the “System variables” panel, click edit, and you will see a popup window with the title “Edit environment variable”. Click “New” and add C:\somewhere\spark-2.4.2-bin-hadoop2.7\bin (the Spark bin folder). Click OK to close the popup windows.

**5. Try to run Spark using command line.**

Go to the search box in the lower left corner of your desktop and type cmd to open Command Prompt. Type "spark-shell” to run spark. If you see the following screen, congratulations, you have successfully installed Spark on your windows machine.



# **6. Setup Jupyter Notebook for Apache Spark**

Next, we want to set up Jupyter Notebook for PySpark (the Spark Python API is called PySpark, which is included in the Spark package you just installed).

6.1. Download Anaconda for python programing. <https://www.anaconda.com/distribution/#download-section>. Please select the link “Windows” on the top of the page and make sure you are downloading **anaconda 3** for Windows installer. Install Anaconda using mostly default settings (except that in one step, you need to check a check box to add python to path)

6.2. Add/Change below environment variables according to your software version and location. (If you have installed Anaconda in a different location, the path could be different.)

PYSPARK\_PYTHON = C:\ProgramData\Anaconda3\python.exe

PYSPARK\_DRIVER\_PYTHON = C:\ProgramData\Anaconda3\Scripts\jupyter.exe

PYSPARK\_DRIVER\_PYTHON\_OPTS = notebook

6.3. Run Jupyter Notebook with PySpark:

Click the windows icon in the lower left corner-> click anaconda 3 -> click Anaconda Prompt. In the Anaconda Prompt window, type **pyspark** in the command line. If you see Jupyter Notebook open in your preferred web browser, you have successfully set up the Jupyter Notebook with PySpark.