SPARK With Jupyter Notebook

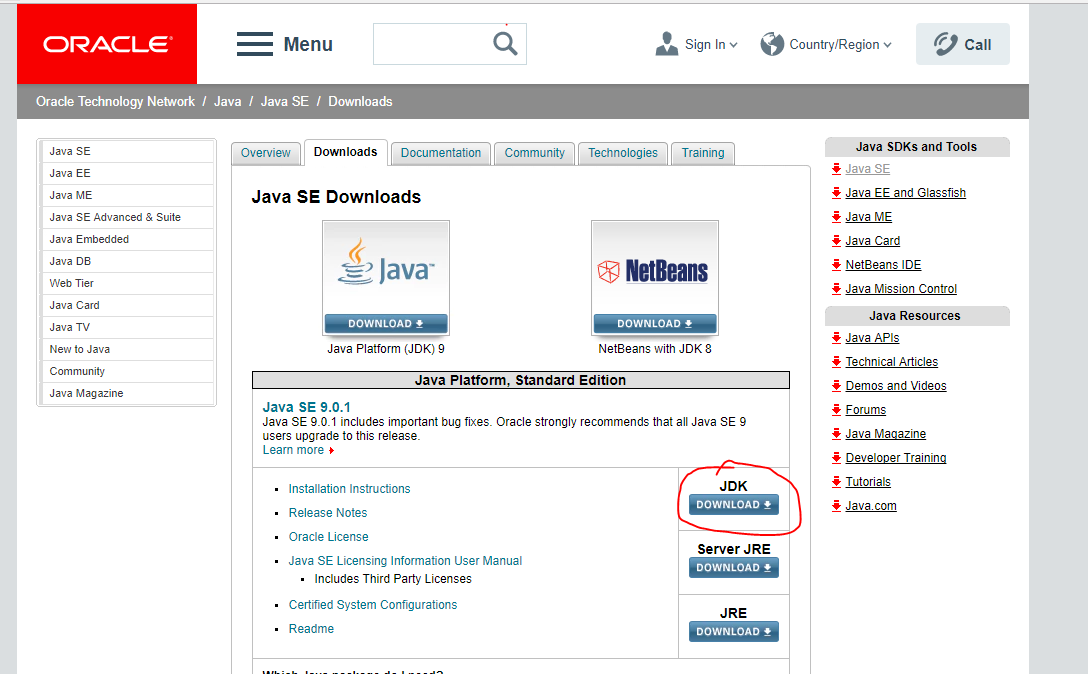
**Prerequisites:**

1. Anaconda
2. Java SDK
3. Spark

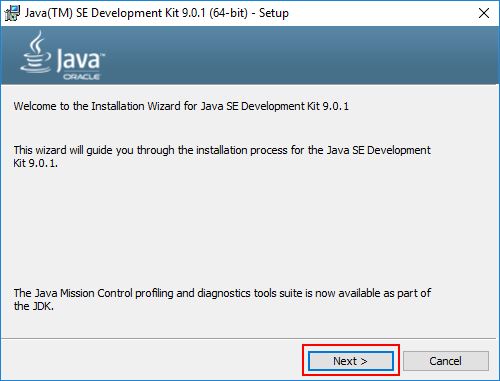
2. Installing Java:

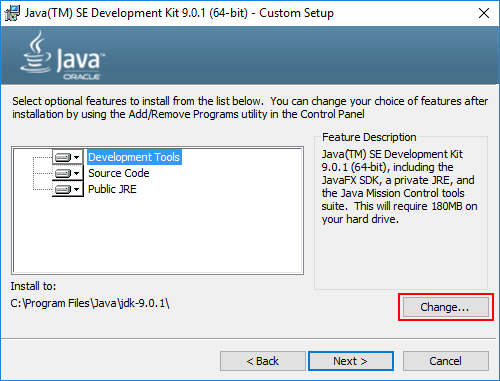
Download the java SDK setup file from following site.

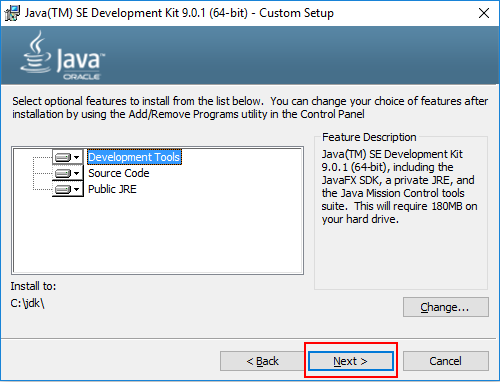
<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

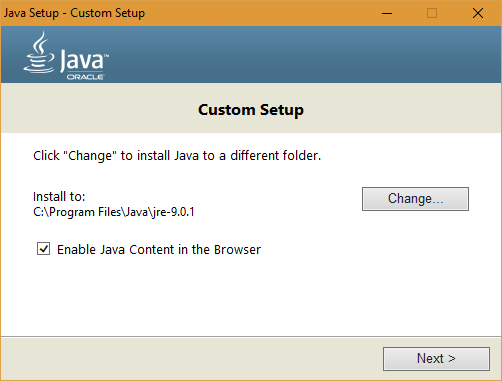


Now you will be navigated to download page as below, there you need to accept license agreement and download Windows version of installer. And Run the installer.











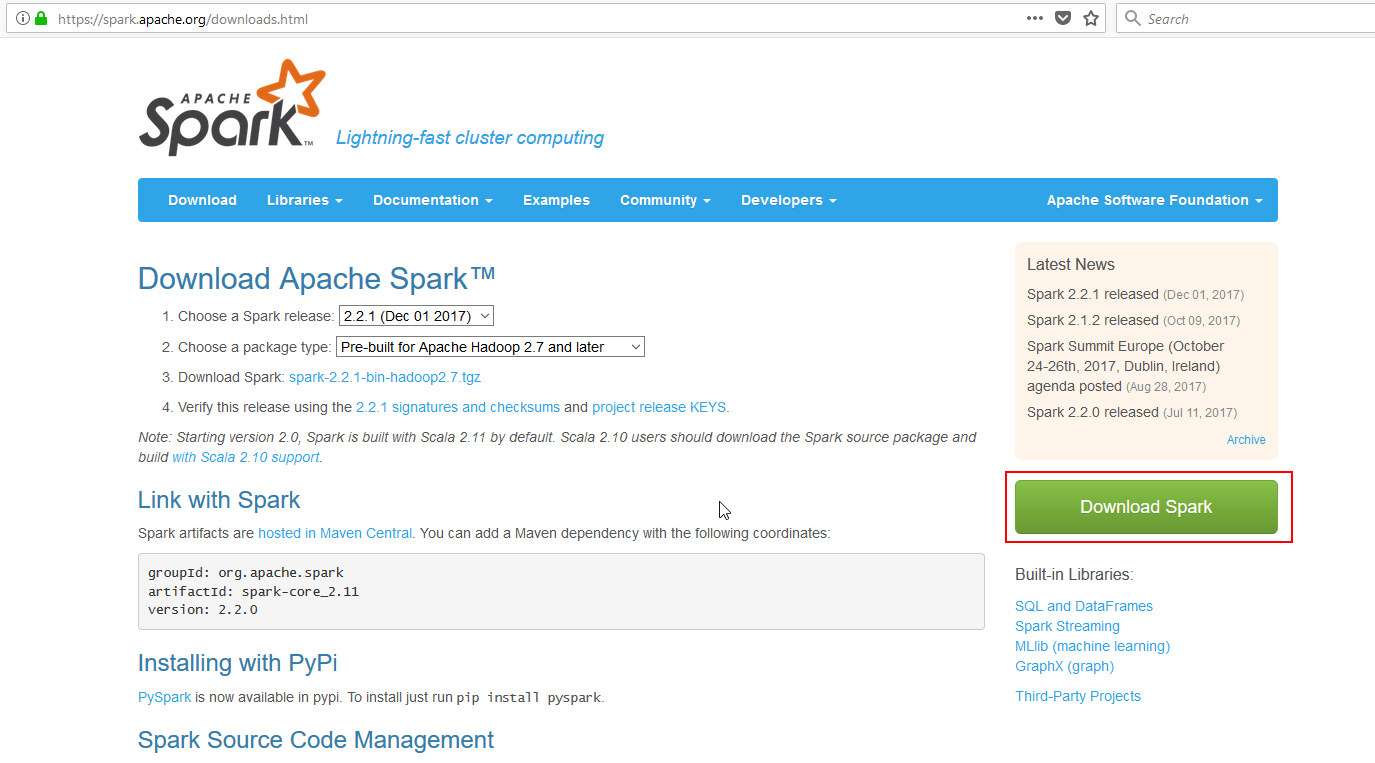


2. Installing Java:

Download a **pre-built** version of Apache Spark from

<https://spark.apache.org/downloads.html>

Note: If required, download and install WinRAR so you can extract the .tgz file you downloaded. <http://www.rarlab.com/download.htm>



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.

Download “hadoop-2.7.1” from below link

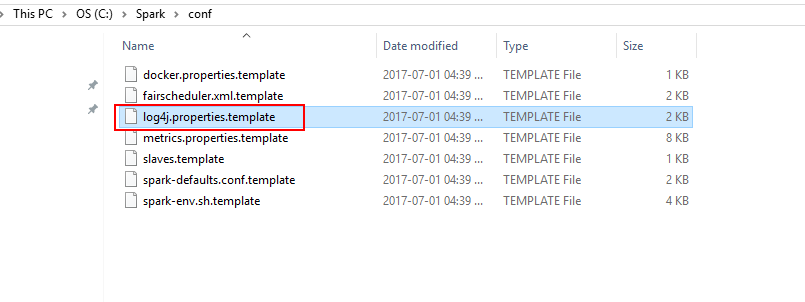
<https://github.com/steveloughran/winutils>

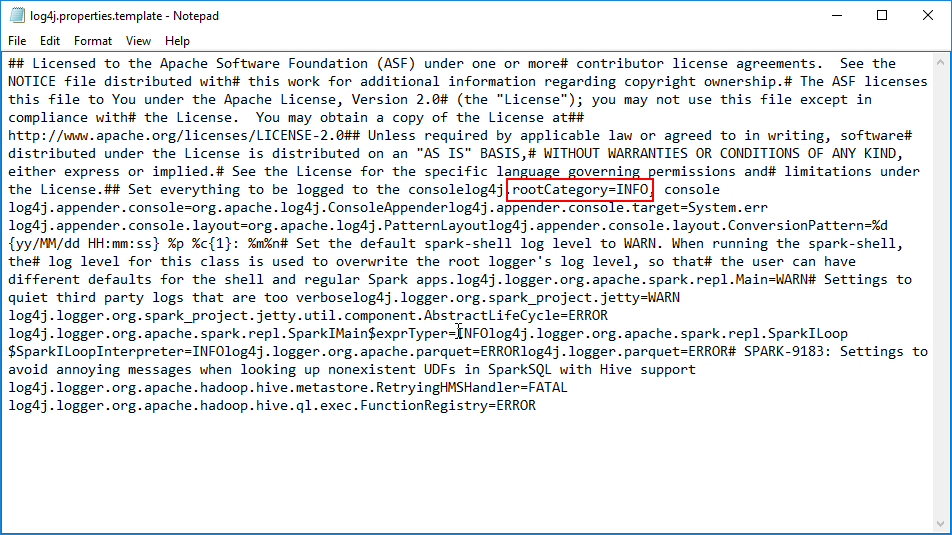
And move folder contents into a **C:\winutils** folder that you’ve created.

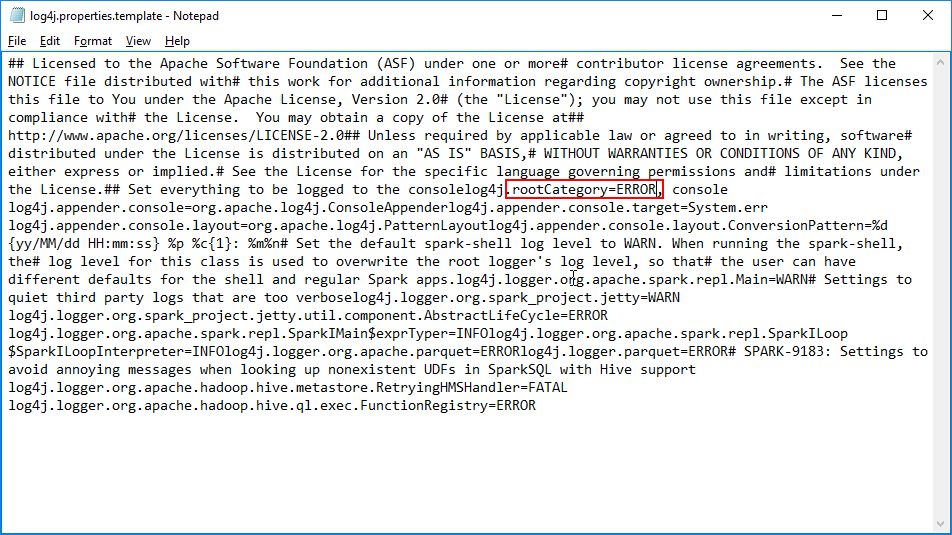
Open the the **c:\spark\conf** folder, and make sure “File Name Extensions” is checked in the “view” tab of Windows Explorer. Rename the log4j.properties.template file to log4j.properties.

Edit this file (using Wordpad or something similar) and change the error level from INFO to

ERROR for log4j.rootCategory

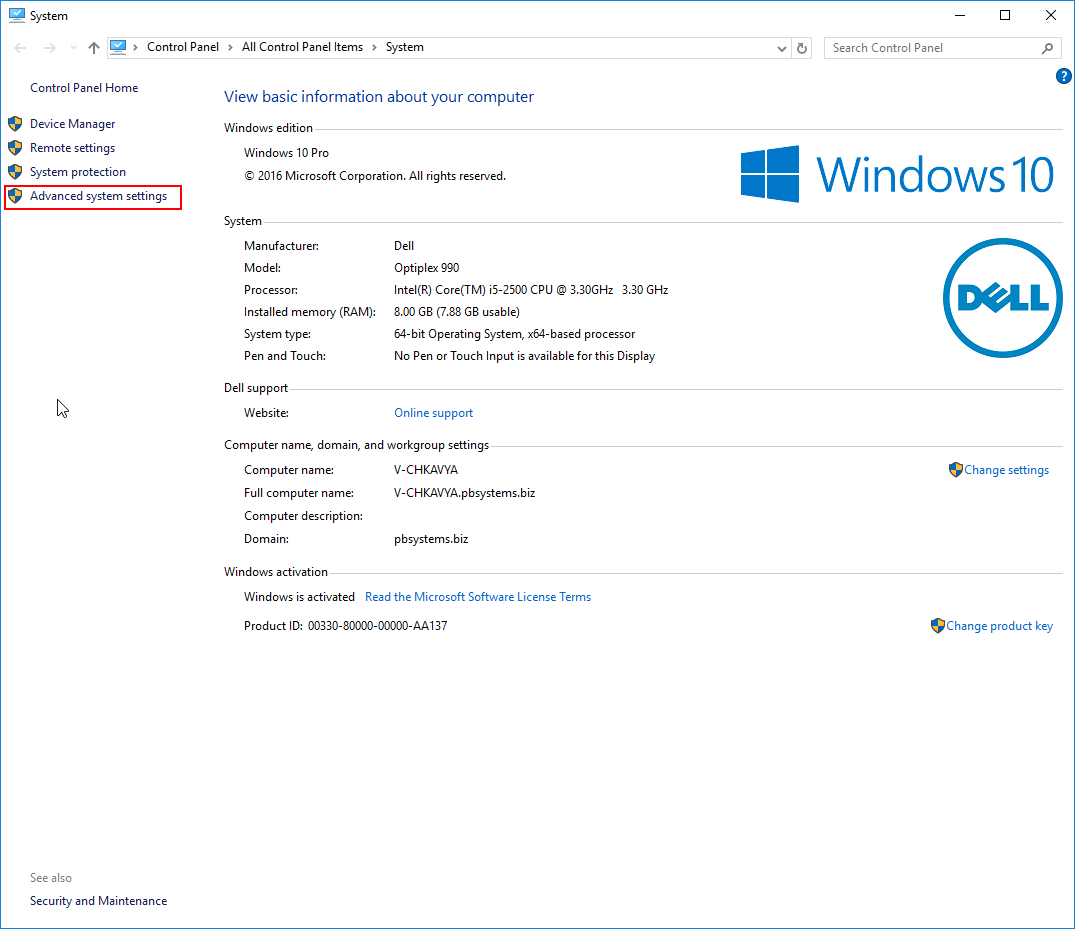




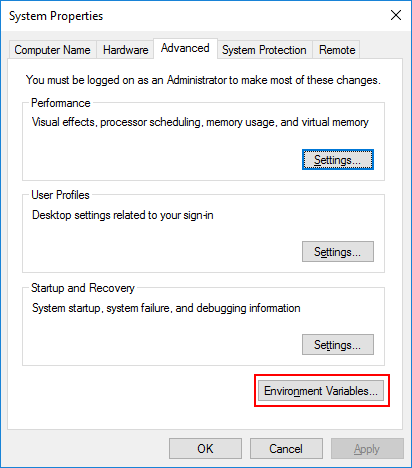


Setting Environment Variables:

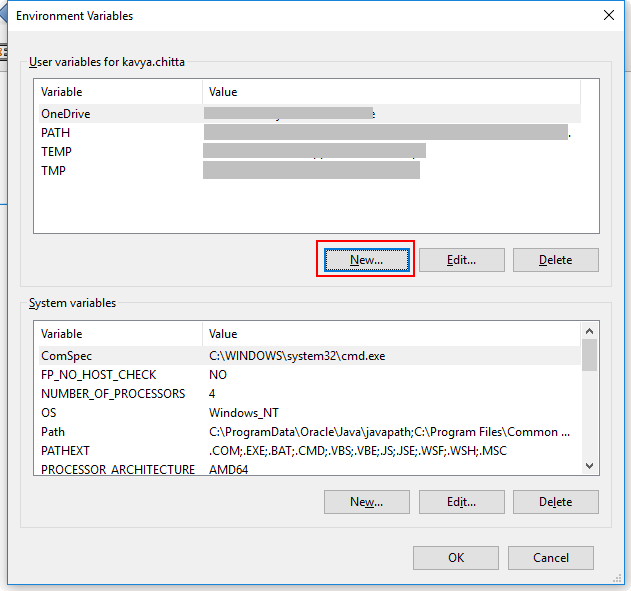
1. Right-click your Windows menu, select System.
2. Click on “Advanced System Settings” and then the “Environment Variables” button.



1. Click on “New” and then the “Environment Variables” button.

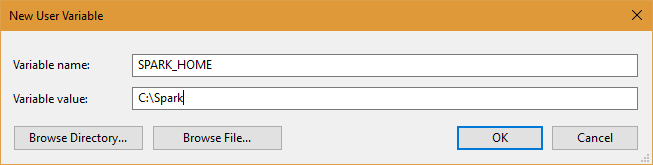


1. Click on “New” and then the “Environment Variables” button.



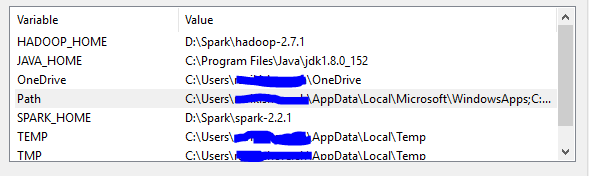
Enter name: SPARK\_HOME

Value: <<Your spark extracted path>>



1. Similarly repeat step 4 for following variables,

|  |  |
| --- | --- |
| **Name** | **Value** |
| JAVA\_HOME | <<JDK Path>> |
| HADOOP\_HOME | <<Win Utils Path>> |
| PYSPARK\_DRIVER\_PYTHON | ipython |
| PYSPARK\_DRIVER\_PYTHON\_OPTS | notebook |
|  |  |



1. Now double click on “Path” under system Variables and click “Add” button and enter following

%SPARK\_HOME%\bin

1. Again click “Add” button and enter following

%JAVA\_HOME%\bin

