**Install Apache Spark**

**on**

**Windows 10**

**By**

**Priyanka Nirpase**

**Email:** [**pnirpase@scu.edu**](mailto:pnirpase@scu.edu)

**October 9, 2017**

* You will need to install below software in your system:

1. **Scala**: Latest version available here <https://scala-lang.org/download/>

It will be installed in C:\Program Files (x86)\sbt or C:\Program Files (x86)\Scala.

1. **Java**: Latest version available here <http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

It will be installed in C:\Program Files\Java\jdk1.8.0\_131

1. **Apache** **Spark**: Latest version available here <https://spark.apache.org/downloads.html>

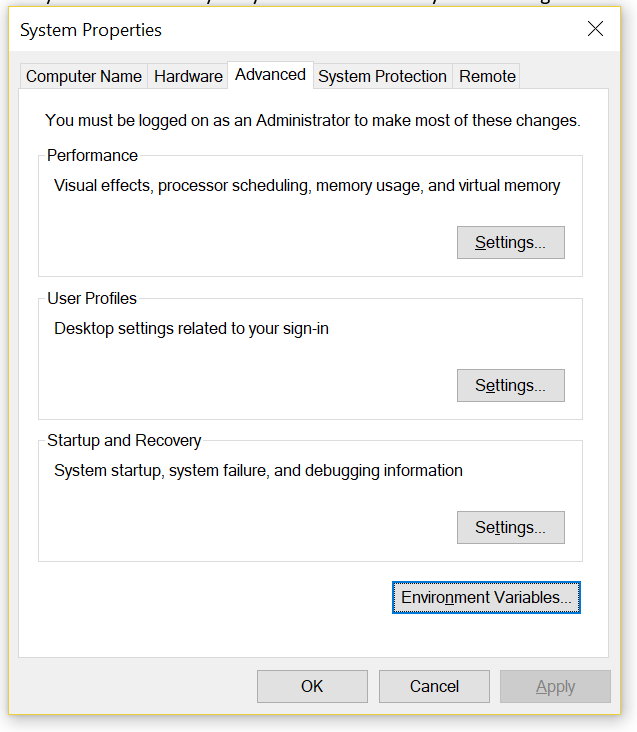
Extract the contents of*spark-2.0.0-bin-hadoop2.7.tgz at* at C:\Spark\

1. **Windows** **utility(winutils.exe):** <https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin>

Place the winutils.exe at C:\winutils\bin

1. **Anaconda** **and Jupyter:** Latest version available at <https://www.anaconda.com/download/>

* Once the software are installed, it’s time to change environment variables. Go to Control Panel🡪 System and Security🡪 System🡪Advanced System settings🡪Environment Variables



Add below user variables with their corresponding paths:

**SCALA\_HOME: : C:\Program Files (x86)\sbt\bin**

**JAVA\_HOME: C:\Program Files\Java\jdk1.8.0\_131**

**HADOOP\_HOME: C:\winutils**

**SPARK\_HOME: C:\spark\spark-2.2.0-bin-hadoop2.7\bin**

Make sure you add bin directory at the end for Scala and Spark paths.

In addition to above paths, we need to provide paths for Python and Jupyter.

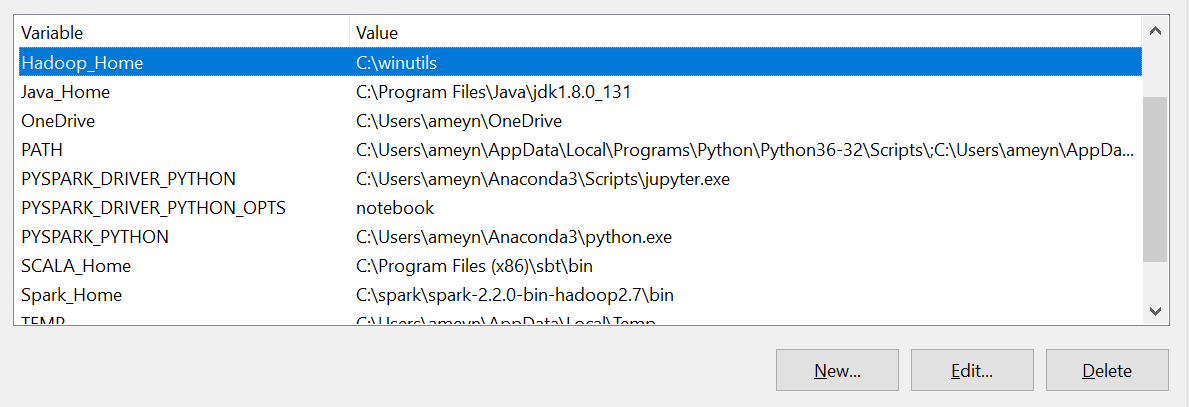
Locate Jupyter.exe and python.exe in your system. Mostly, Jupyter.exe will bel located at C:\Users\<>user\_name>\Anaconda3\Scripts and Python.exe at C:\Users\<user\_name>\Anaconda3.

Add PYSPARK\_DRIVER\_PYTHON user variable under environment variables and provide the path for jupyter.exe

Similarly, PYSPARK\_PYTHON and provide path for python.exe

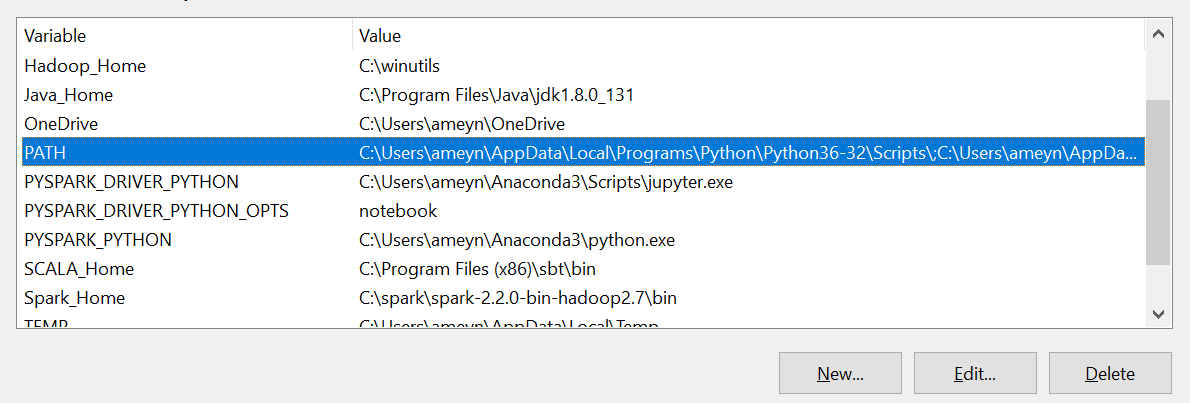
Add PYSPARK\_DRIVER\_PYTHON\_OPTS and provide “notebook” as path.

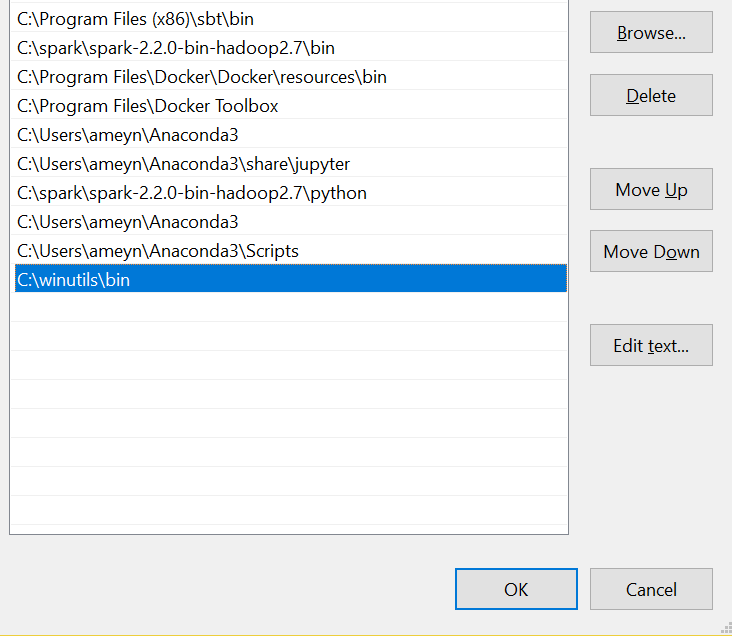
Refer the screenshot.



You will find a user variable called “Path”, click on it and select edit. Add paths for winutils.exe, Scala, Spark, Java, Python and Jupyter by selecting “New” option.

Please ensure that you have added bin folder at the end of the path for winutils(Refer screenshot)

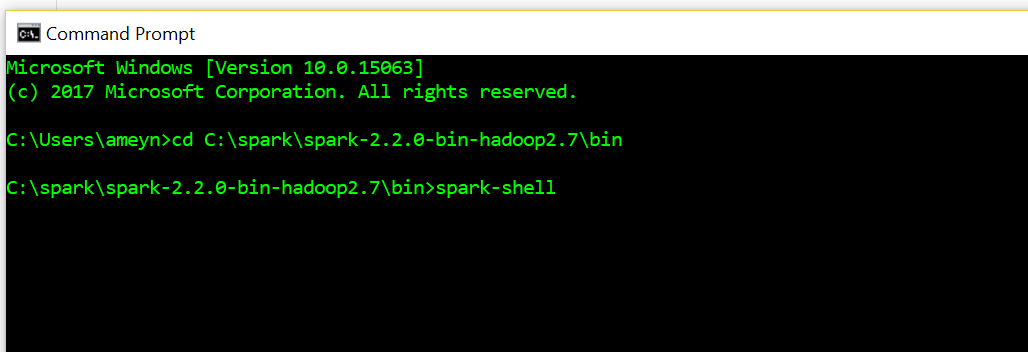




* **Execute Spark**

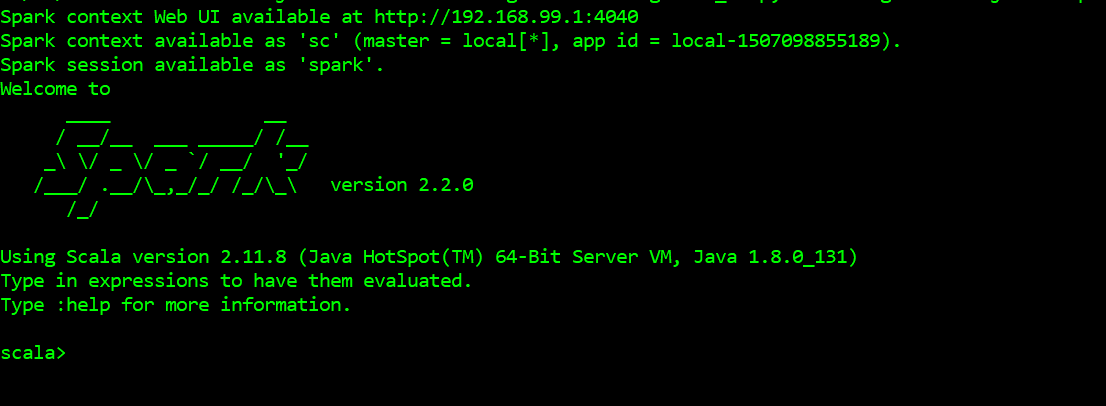
Next step is to execute Spark. Go to command prompt, change directory to C:\spark\spark-2.2.0-bin-hadoop2.7\bin and execute below command.

spark-shell

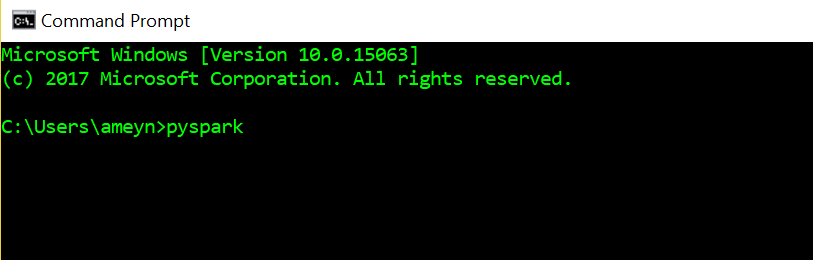


You should get below window with the message as spark session available as ‘spark’.

This means you have successfully installed Spark.



1. Open another command prompt session. Type in pyspark and press enter. You should be connected to Jupyter.



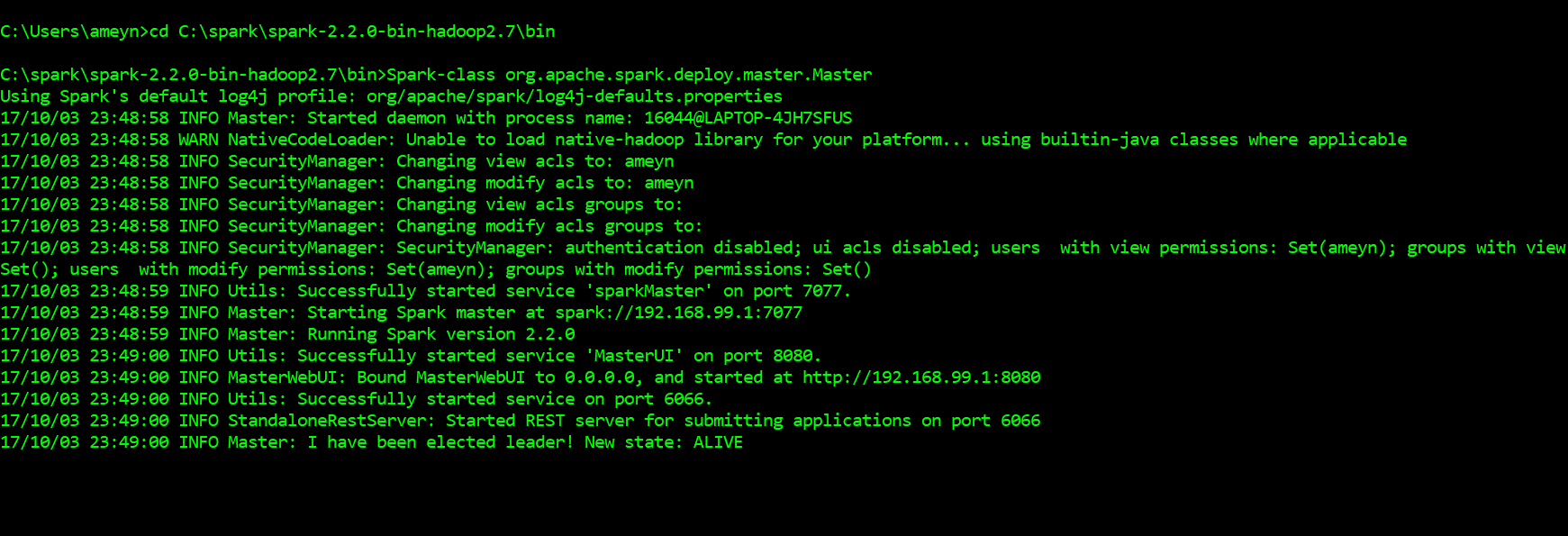


You have successfully connected pyspark to Jupyter.

* Installing cluster on Windows

Go to command prompt, change directory to C:\spark\spark-2.2.0-bin-hadoop2.7\bin and execute below command to initiate a master.

Spark-class org.apache.spark.deploy.master.Master



Master will give address for starting application.

Now go to browser and type <http://localhost:8080/>

You should be able to see below window which shows that spark is running.

