**Installation of PySpark (Windows)**

**Pyspark = Python + Apache Spark**

Apache Spark is a new and open-source framework used in the big data industry for real-time processing and batch processing. It supports different languages, like Python, Scala, Java, and R.

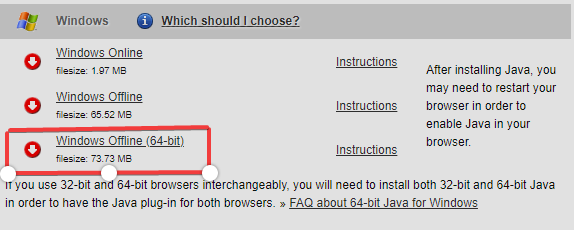
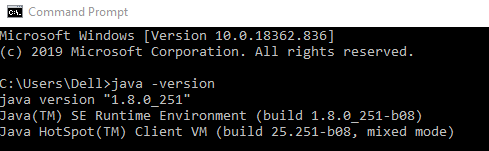
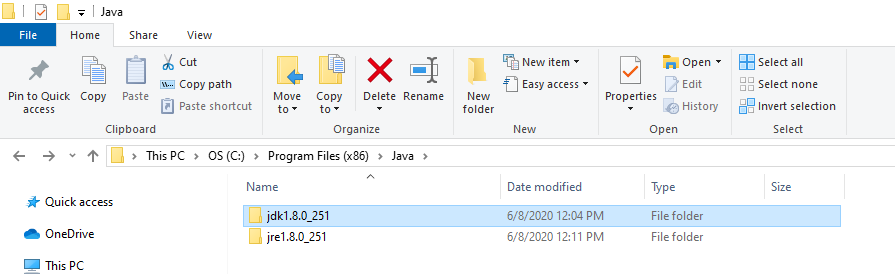
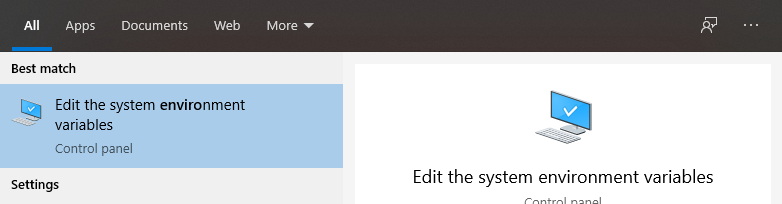
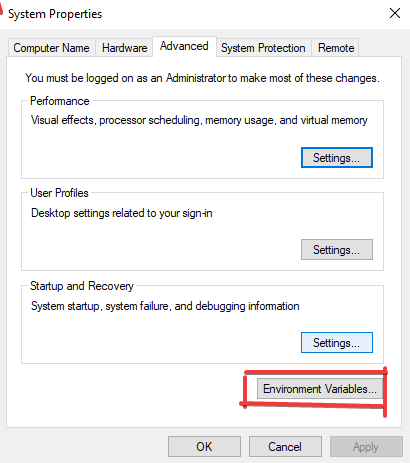
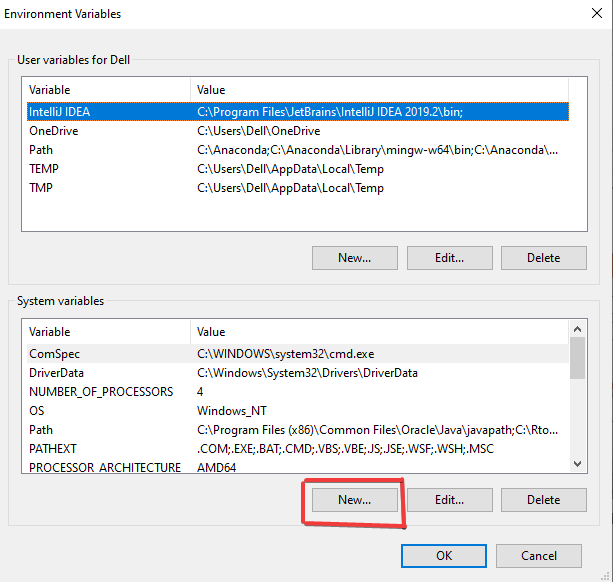
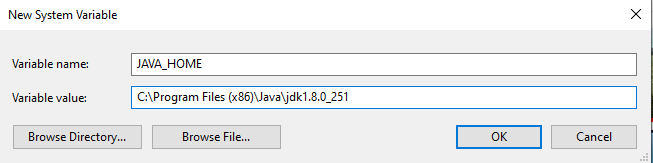
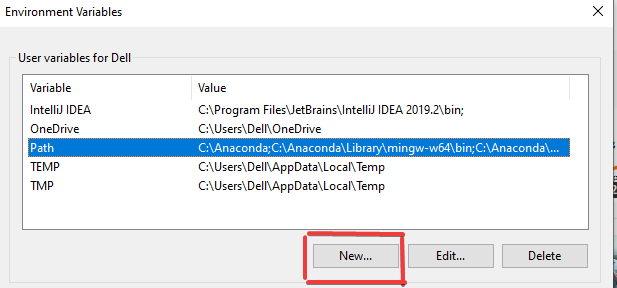
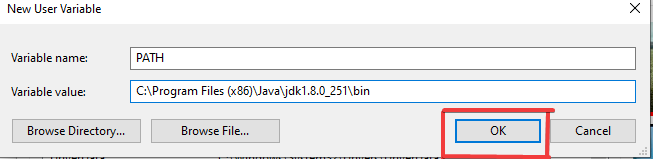
Apache Spark is initially written in a Java Virtual Machine(JVM) language called Scala, whereas Pyspark is like a Python API which contains a library called Py4J. This allows dynamic interaction with JVM objects.

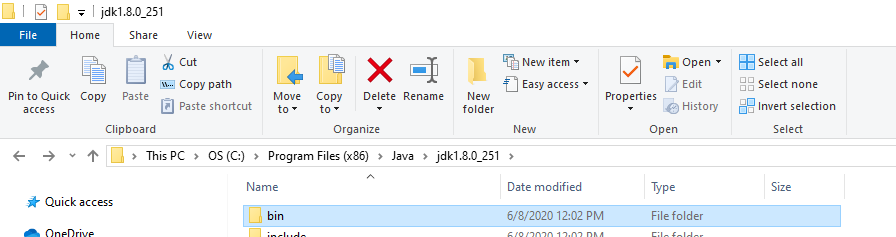
**Windows Installation**

The installation which is going to be shown is for the Windows Operating System. It consists of the installation of Java with the environment variable and Apache Spark with the environment variable.

The recommended pre-requisite installation is Python, which is done from **here**.

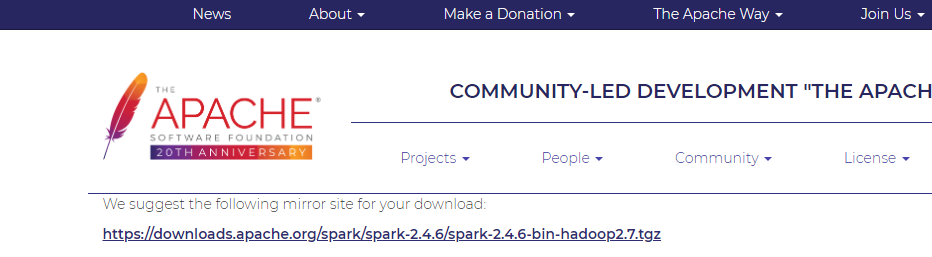
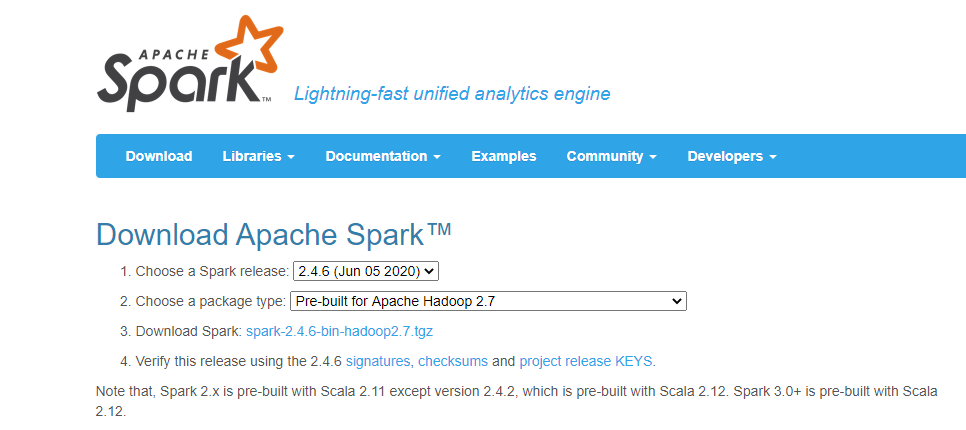
**Java installation**

1. Go to **Download Java JDK**.  
   Visit Oracle's website for the download of the Java Development Kit(JDK).
2. Move to download section consisting of operating system Windows, and in my case, it's Windows Offline(64-bit). The installer file will be downloaded.
3. Open the installer file, and the download begins.
4. Go to "Command Prompt" and type "java -version" to know the version and know whether it is installed or not.
5. Add the Java path
6. Go to the search bar and "EDIT THE ENVIRONMENT VARIABLES.
7. Click into the "Environment Variables'
8. Click into "New" to create your new Environment variable.
9. Use Variable Name as "JAVA\_HOME' and your Variable Value as 'C:\Program Files (x86)\Java\jdk1.8.0\_251'. This is your location of the Java file. Click 'OK' after you've finished the process.
10. Let's add the User variable and select 'Path' and click 'New' to create it.
11. Add the Variable name as 'PATH' and path value as 'C:\Program Files (x86)\Java\jdk1.8.0\_251\bin', which is your location of Java bin file. Click 'OK' after you've finished the process.

**Note:** You can locate your Java file by going to C drive, which is C:\Program Files (x86)\Java\jdk1.8.0\_251' if you've not changed location during the download.

**Installing Pyspark**

1. Head over to the **Spark homepage**.
2. Select the Spark release and package type as following and download the .tgz file.



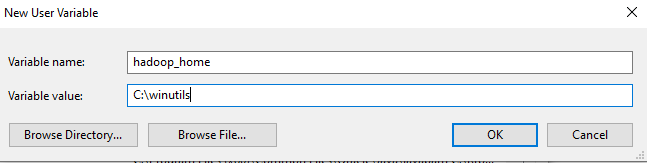
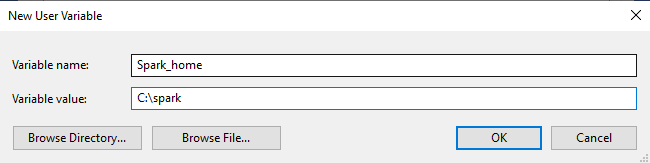
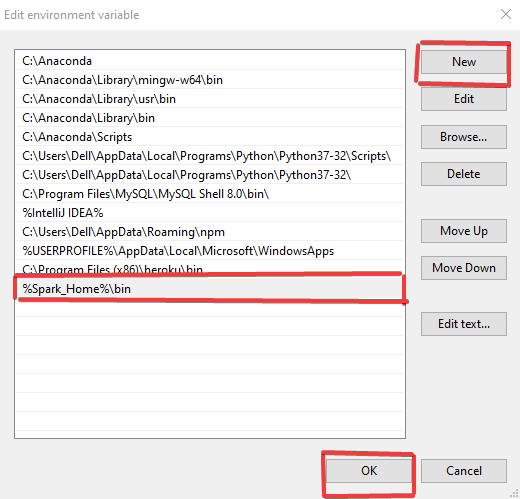
You can make a new folder called 'spark' in the C directory and extract the given file by using 'Winrar', which will be helpful afterward.

**Download and setup winutils.exe**

Go to **Winutils** choose your previously downloaded Hadoop version, then download the winutils.exe file by going inside 'bin'. The link to my Hadoop version is: **https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe**

Make a new folder called 'winutils' and inside of it create again a new folder called 'bin'.Then put the file recently download 'winutils' inside it.

**Environment variables**

1. Let's create a new environment where variable name as "hadoop\_home" and variable value to be the location of winutils, which is "C:\winutils" and click "OK".  
   
2. For spark, also let's create a new environment where the variable name is "Spark\_home" and the variable value to be the location of spark, which is "C:\spark" and click "OK".  
   
3. Finally, double click the 'path' and change the following as done below where a new path is created "%Spark\_Home%\bin' is added and click "OK".  
   

**Finalizing Pyspark Installation**

1. Open Command Prompt and type the following command.  
   Finalizing Pyspark Installation
2. Once everything is successfully done, the following message is obtained.  
   