**PREREQUISITE:**

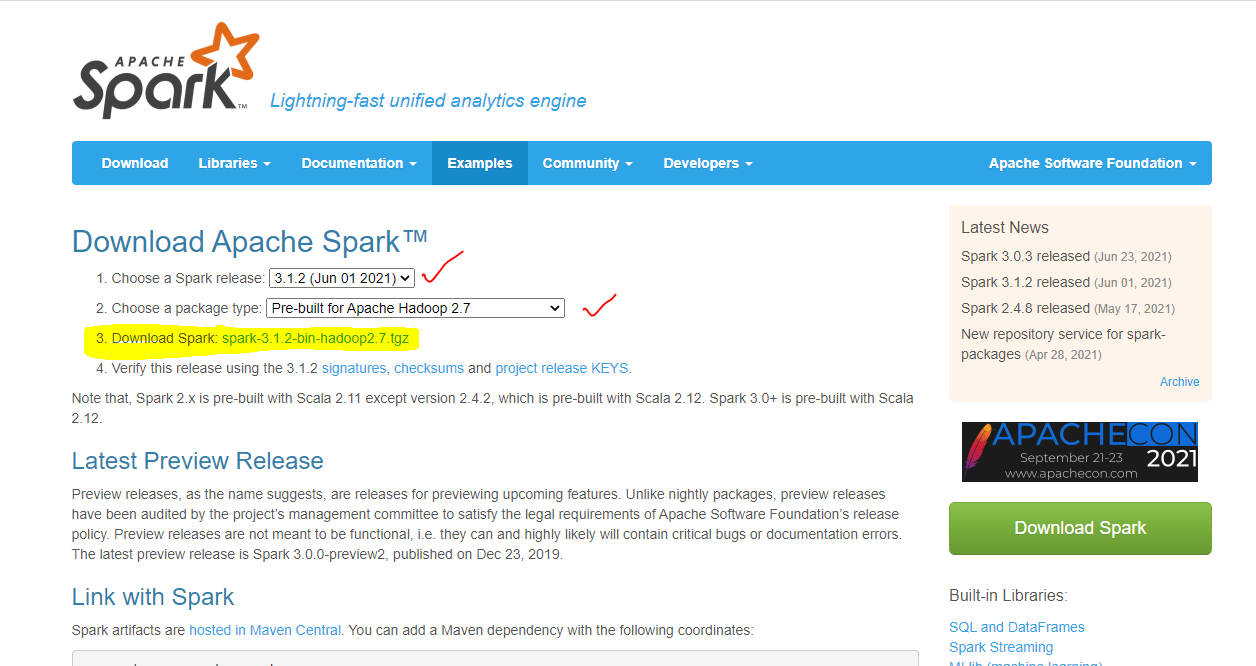
1. **Anaconda (with python 3.8) should be download and installed:** [**https://www.anaconda.com/products/individual**](https://www.anaconda.com/products/individual)
2. **Python 3.8.8 is already installed on your machine**[**https://www.python.org/downloads/release/python-388/**](https://www.python.org/downloads/release/python-388/)

**STEPS TO FOLLOW:**

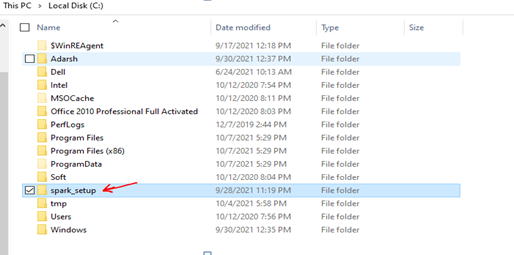
1. Using the below link download the spark:

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

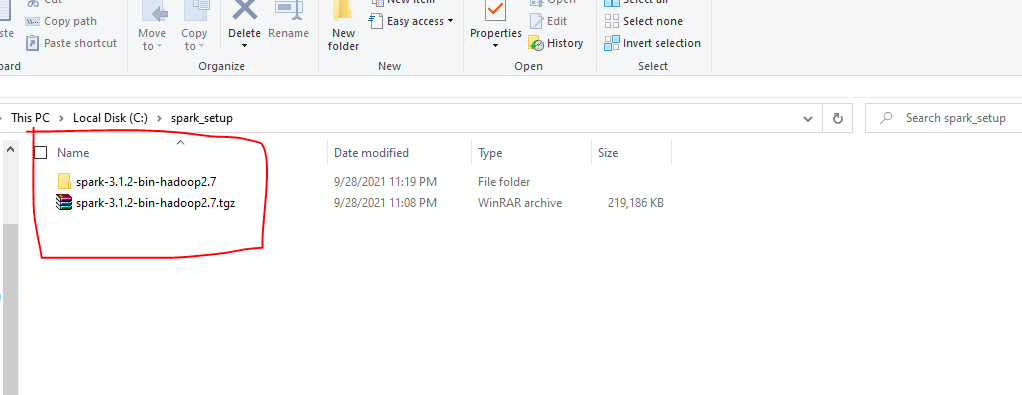
1. Choose the spark release and package type as mentioned in the screenshot below and then click on the download link.



1. Anywhere in you **(C:) Drive** , create a folder **spark\_setup** (not a rule, just for keeping stuff isolated)

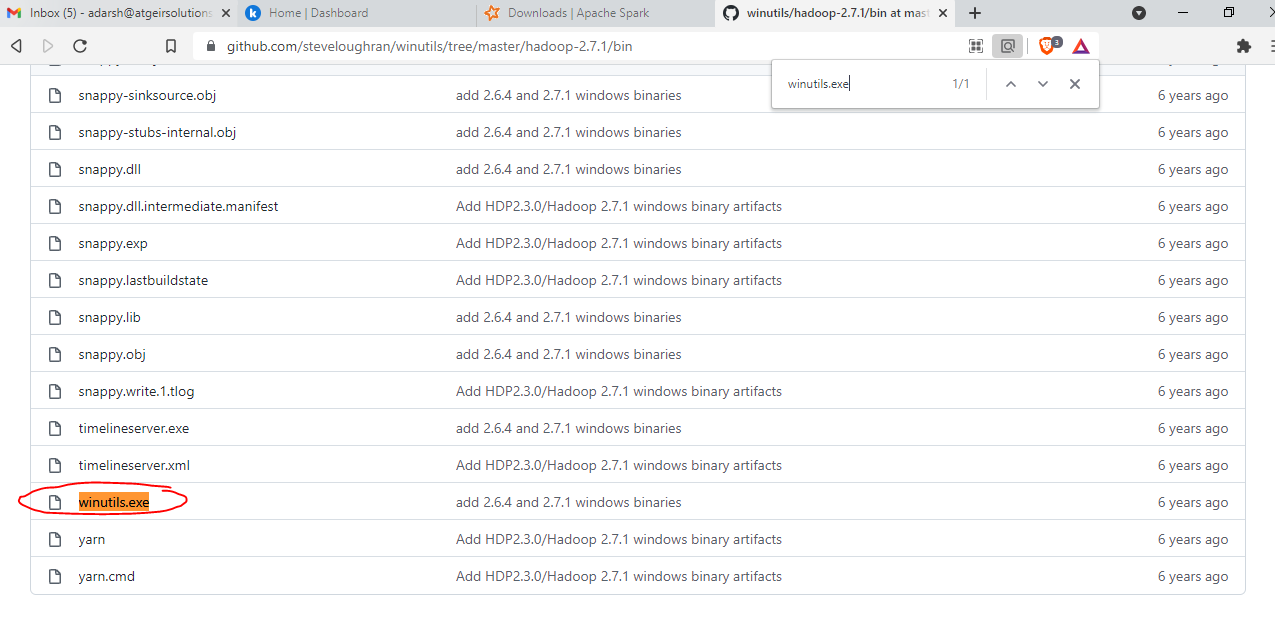


1. Put the downloaded compressed file inside this folder and upzip it. A new folder with the name **spark-3.1.2-bin-hadoop2.7** should start appearing as shown below..

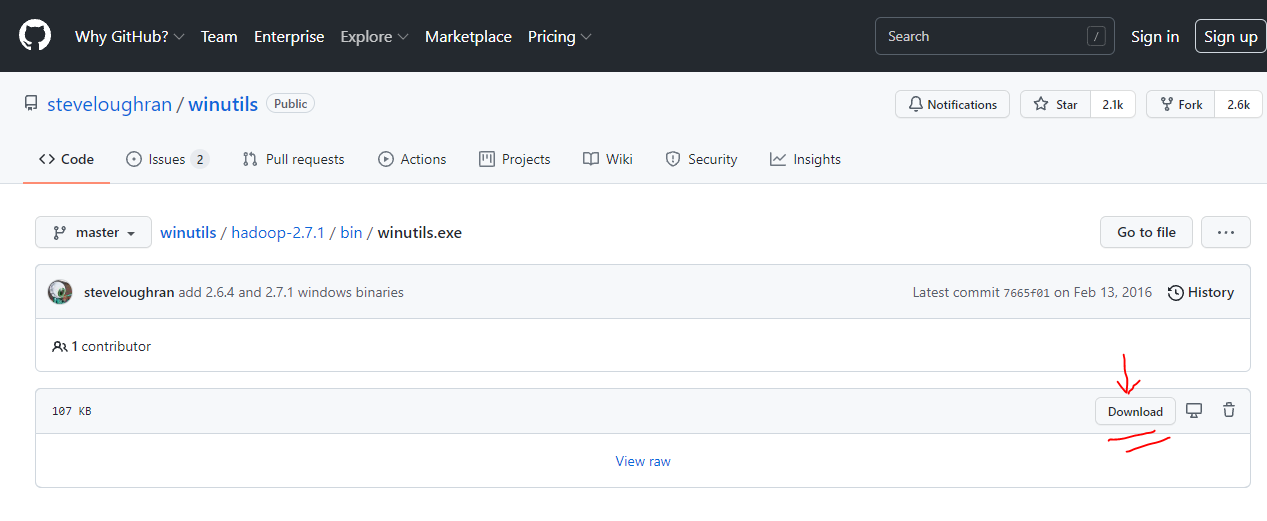


1. There is another file that needs to be downloaded, it’s called **winutils.exe,** and can be downloaded from here below git hub page:

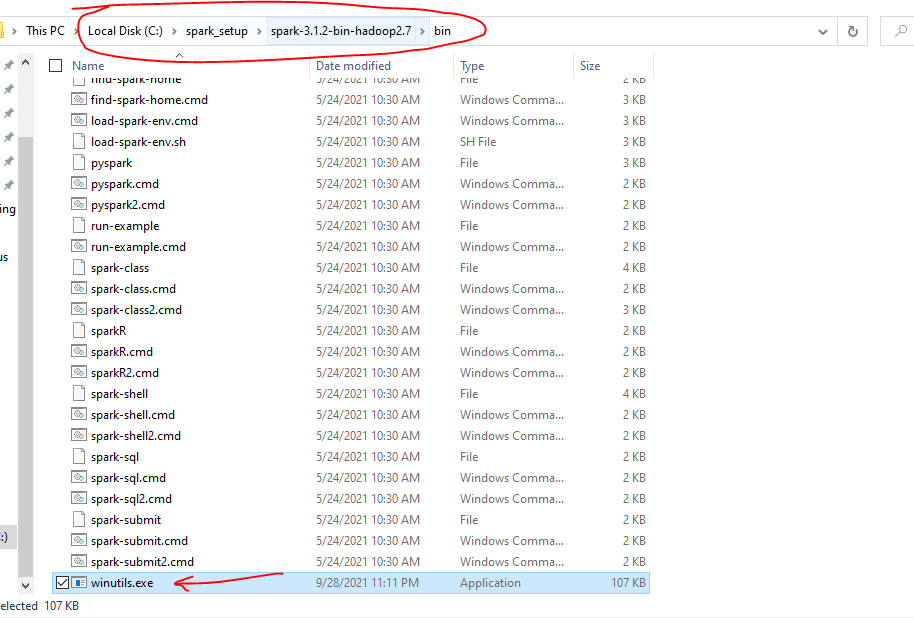
<https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin>



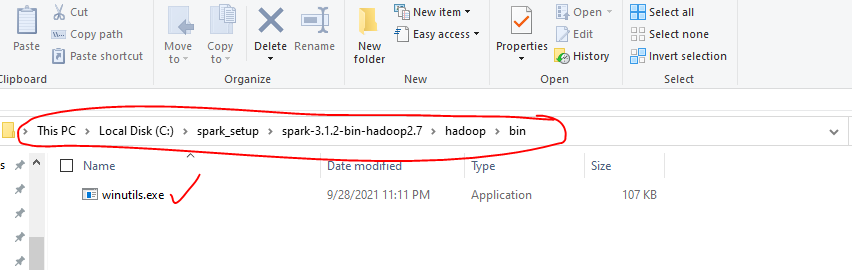
Click on this file and then click on Download:



1. Put this **winutils.exe** file inside the **bin** folder present inside the folder: **spark-3.1.2-bin-hadoop2.7** which appeared after unzipping.



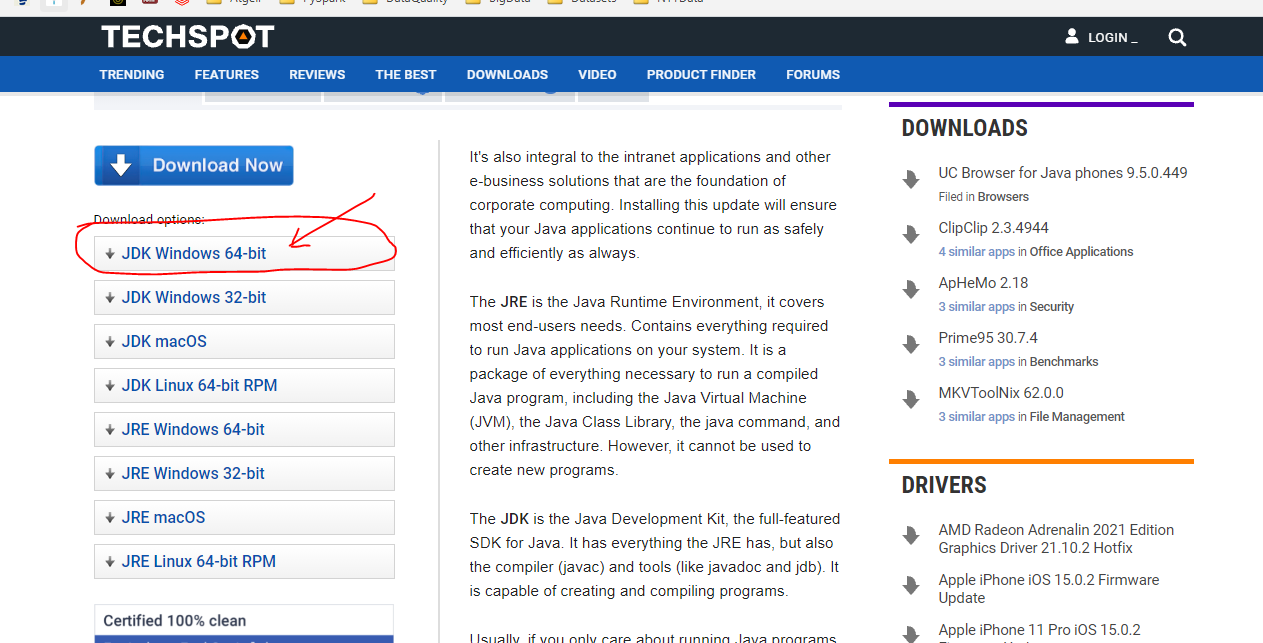
1. Under the folder **spark-3.1.2-bin-hadoop2.7** create another folder called **hadoop** and inside this folder create another folder called **bin.** Copy-Paste the file **winutils.exe** here as well.



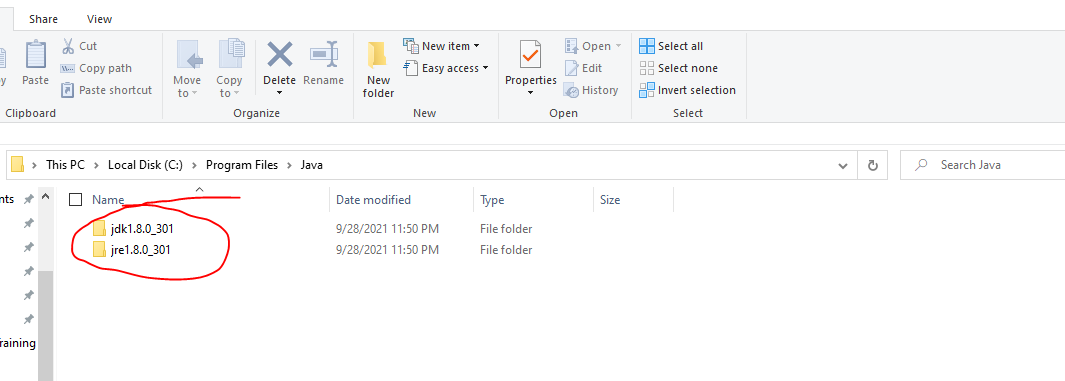
1. **This is the most important part: Install Java version 8 ONLY!!! As other versions will create issues when running pyspark later.**

**Download here:** [**https://www.techspot.com/downloads/5198-java-jre.html#google\_vignette**](https://www.techspot.com/downloads/5198-java-jre.html#google_vignette)

Click on the location pointed by the arrow in the below screenshot, and in few seconds download will begin…

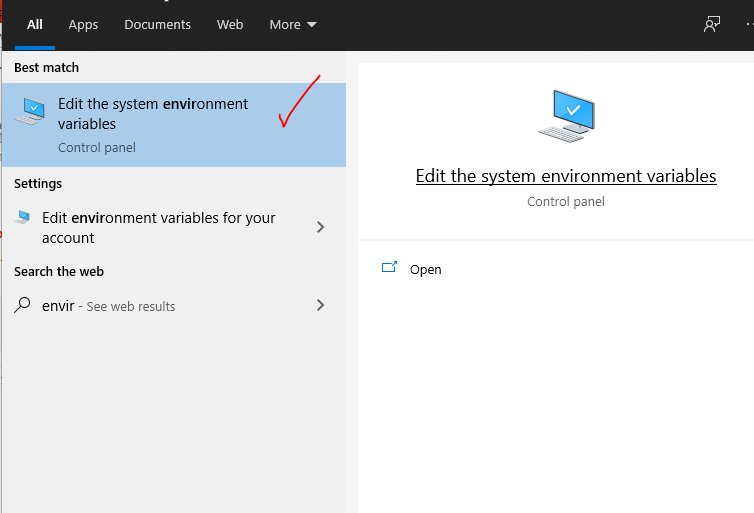


After installation of JDK, below shown folder should appear in your machine too…

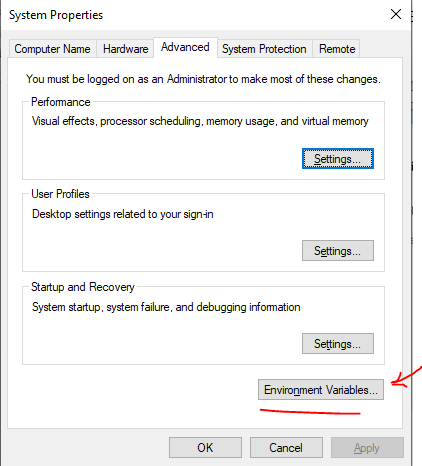


1. Setting up the environment variables…

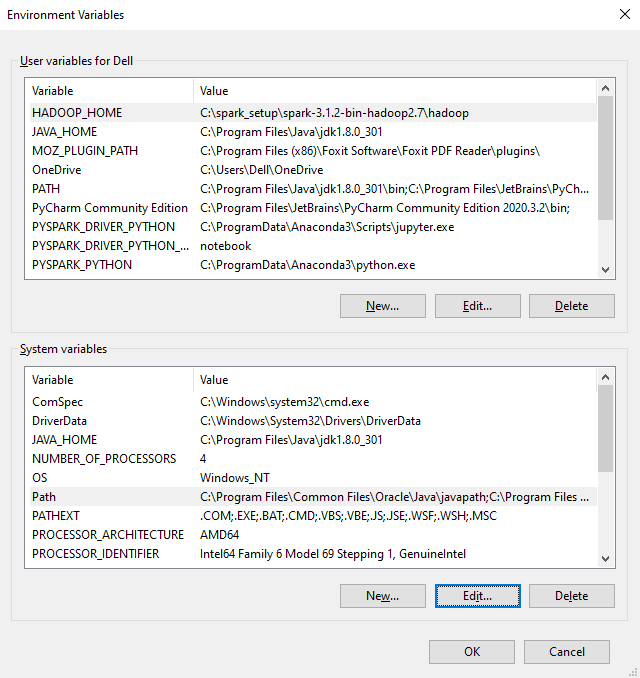
Press the window button and search: **Environment Variable.** Click open it.



Click on the button **Environment Variable.**



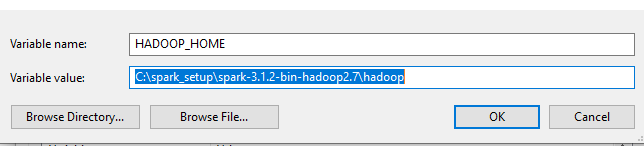
You will now see the below screen:



Under the **“User Variables for Dell”**

1. click on button **New** and put as below:

**Variable Name:** HADOOP\_HOME **Variable Value:** C:\spark\_setup\spark-3.1.2-bin-hadoop2.7\hadoop

and click on the **OK** button.

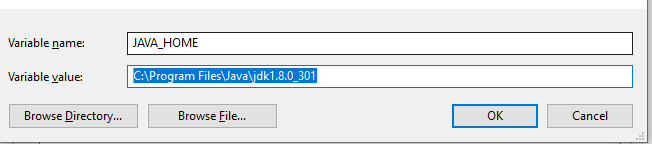
***Note:*** *Make sure the* ***path*** *mentioned against the field* ***’variable value’*** *is according to the folder location where you unzipped the spark file.*

1. Again click on button **New** and put as below:

**Variable Name:** JAVA\_HOME

**Variable Value:** C:\Program Files\Java\jdk1.8.0\_301

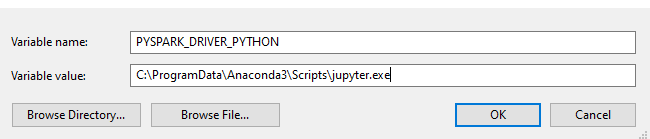
and click on the **OK** button.



1. Again click on button **New** and put as below (In-order to run PySpark on Jupyter notebook):

**Variable Name:** PYSPARK\_DRIVER\_PYTHON

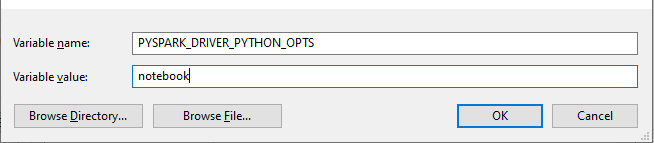
**Variable Value:** C:\ProgramData\Anaconda3\Scripts\jupyter.exe

and click on the **OK** button.

1. Again click on button **New** and put as below (In-order to run PySpark on Jupyter notebook):

**Variable Name:** PYSPARK\_DRIVER\_PYTHON\_OPTS

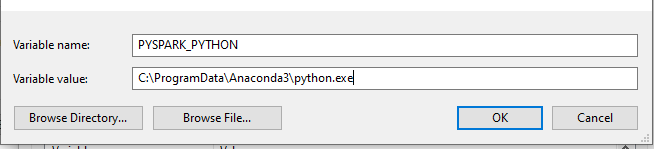
**Variable Value:** notebook

and click on the **OK** button.

1. Again click on button **New** and put as below (In-order to run PySpark on Jupyter notebook):

**Variable Name:** PYSPARK\_PYTHON

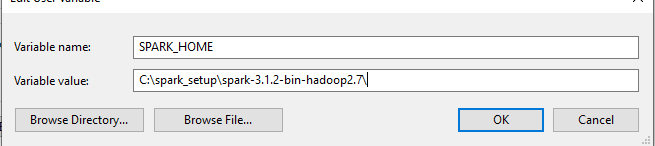
**Variable Value:** C:\ProgramData\Anaconda3\python.exe

and click on the **OK** button.

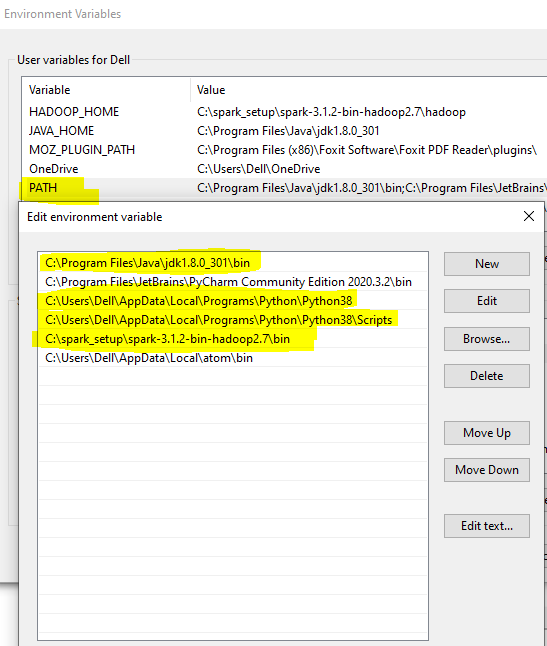
1. Again click on button **New** and put as below:

**Variable Name:** SPARK\_HOME

**Variable Value:** C:\spark\_setup\spark-3.1.2-bin-hadoop2.7\

and click on the **OK** button.  


1. Double click on the variable present in the list; called **PATH** and check the highlighted paths are present in it….

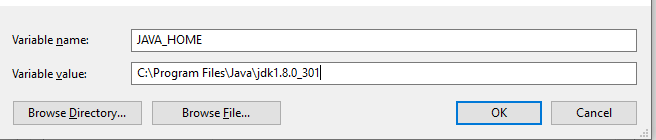


Under the **“System Variables”**

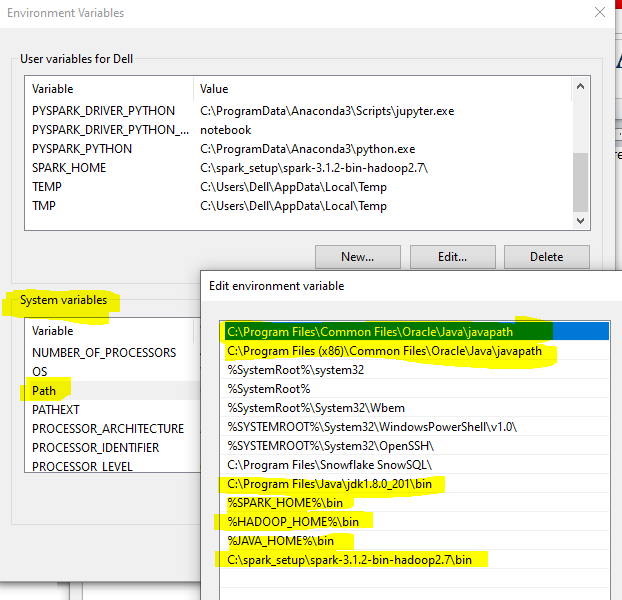
1. Click on **New** button and put data as below:

**Variable Name:** JAVA\_HOME

**Variable Value:** C:\Program Files\Java\jdk1.8.0\_301

and click on the **OK** button.

1. Click on **Path** variable and put highlighted data according to the location in your machine:  
   If the paths are present then no need to put again….

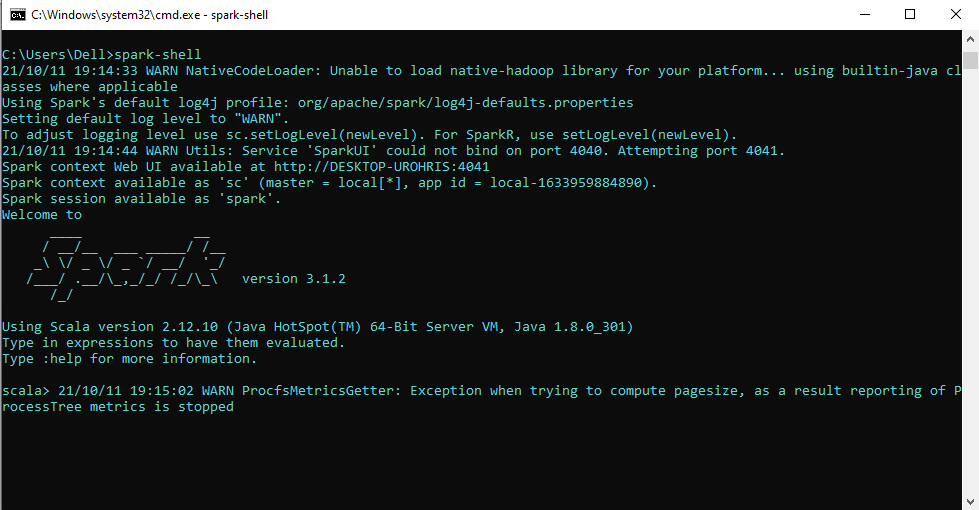


Finally click on the **OK** button on the **Environment Variable** dialog box and then **OK** on the **System Properties** dialog box

1. Launch the Command Prompt and run the command –

**Spark-shell**

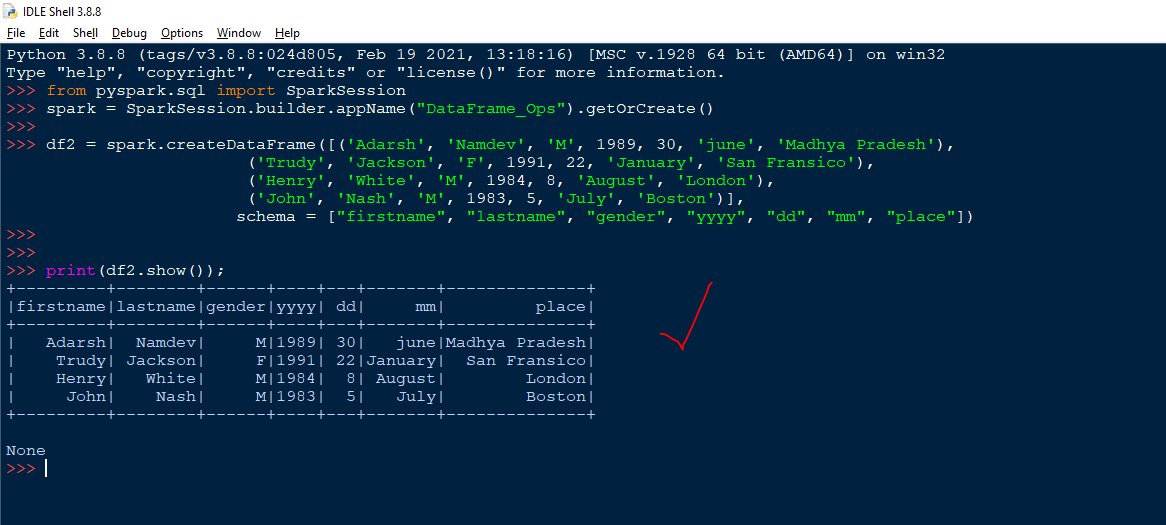
You should see the below upon running this command.



1. Launch another instance of command prompt and run the below command….

**pip install pyspark**

1. Once pyspark is installed….. Launch the IDLE or any development IDE like pycharm / spyder and try running some command, you may try this command for creating a spark DataFrame…



**------------------------ --------------------------- END OF DOCUMENT ----------------------------------------------------**