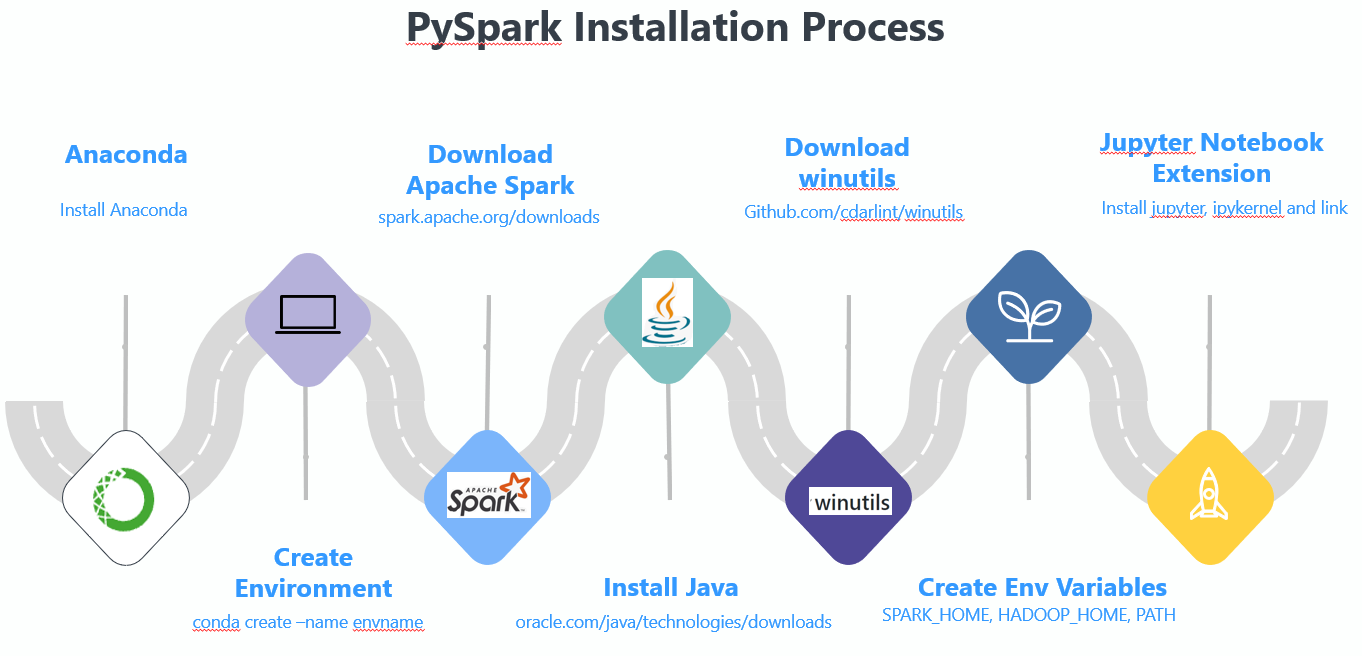
**Introduction**

A guide, on installing Apache Spark with PySpark, with anaconda and jupyter notebook, on windows machine.



Contents

[1. Install Anaconda. 2](#_Toc84928505)

[2. Create Conda environment 2](#_Toc84928506)

[3. Download and unpack apache spark 3](#_Toc84928507)

[4. Install Java 3](#_Toc84928508)

[5. Download winutils for window users 4](#_Toc84928509)

[6. Create environment variables 5](#_Toc84928510)

[7. Test Instalation of Pyspark 6](#_Toc84928511)

[8. Jupyter Notebook Extension 6](#_Toc84928512)

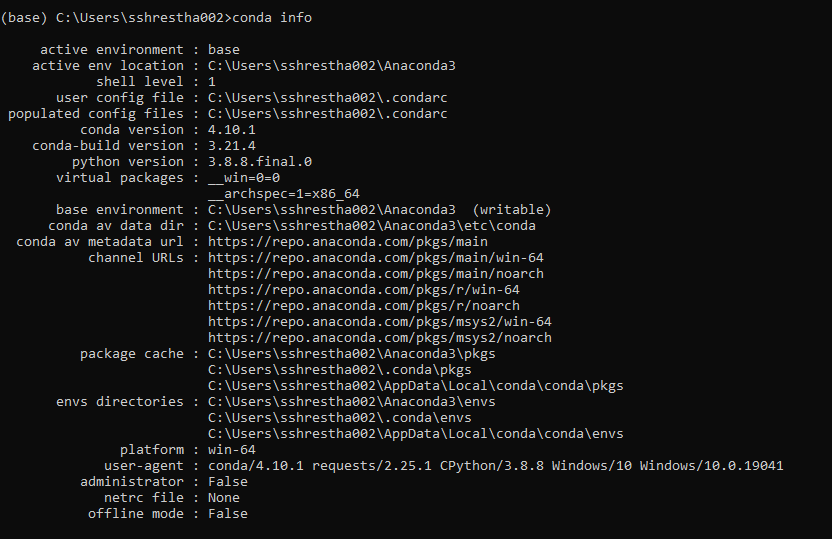
[9. Test script 7](#_Toc84928513)

1. Install Anaconda.

<https://www.anaconda.com/products/individual>

* Once installed use following command in **Anaconda Prompt** to ensure the installation is done correctly.
* Open anaconda Prompt

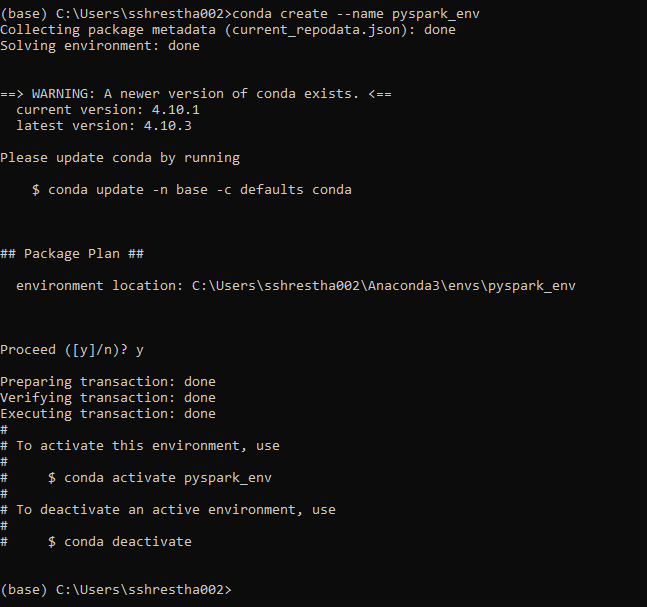
conda info



1. Create Conda environment

* Below, pyspark\_env is your environment name.

conda create –name pyspark\_env

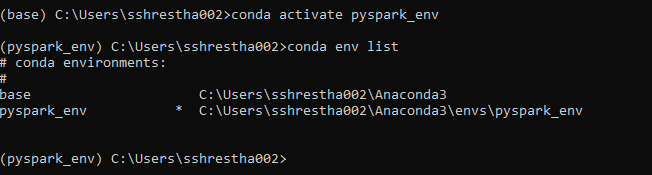


* Check the created conda env using

conda env list

* Activate the newly created env

conda activate pyspark\_env



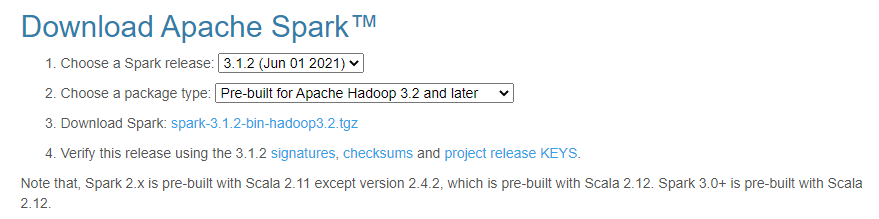
Activated environment

1. Download and unpack apache spark

* Download latest version of apache spark from

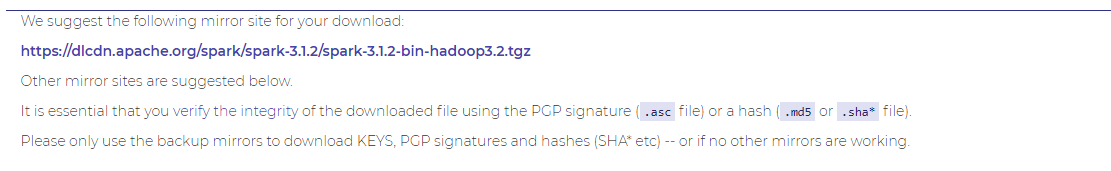
<https://spark.apache.org/downloads.html>

* Select option 3



Select

* click on the recommended link to download



* From anacoda prompt, cd to location where the downloaded file (spark-3.1.2-bin-hadoop3.2.tgz ) is.

cd C:\

* Unzip file in the same locaiton

tar -xzf spark-3.1.2-bin-hadoop3.2.tgz



* Use below command to move file to new location and folder

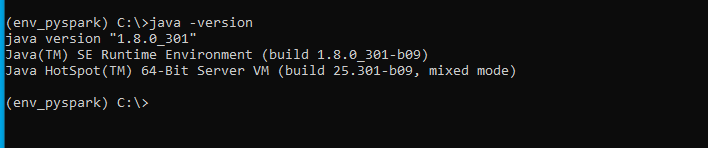
move spark-3.1.2-bin-hadoop3.2 C:\Users\sshrestha002\spark-3.1.2



1. Install Java

* Requires Java 8 or later.
* Check java with

java -version



* If java is not available, go to link below and download

<https://www.oracle.com/java/technologies/downloads/#java8-windows>

* Probably need to create an account.

1. Download winutils for window users

* This step includes downloading winutils, creating Hadoop and bin directories inside the folder spark-3.1.2
* Go to spark-3.1.2 folder you created in Step 3.

cd C:\Users\sshrestha002\spark-3.1.2

* Make directory hadoop and bin

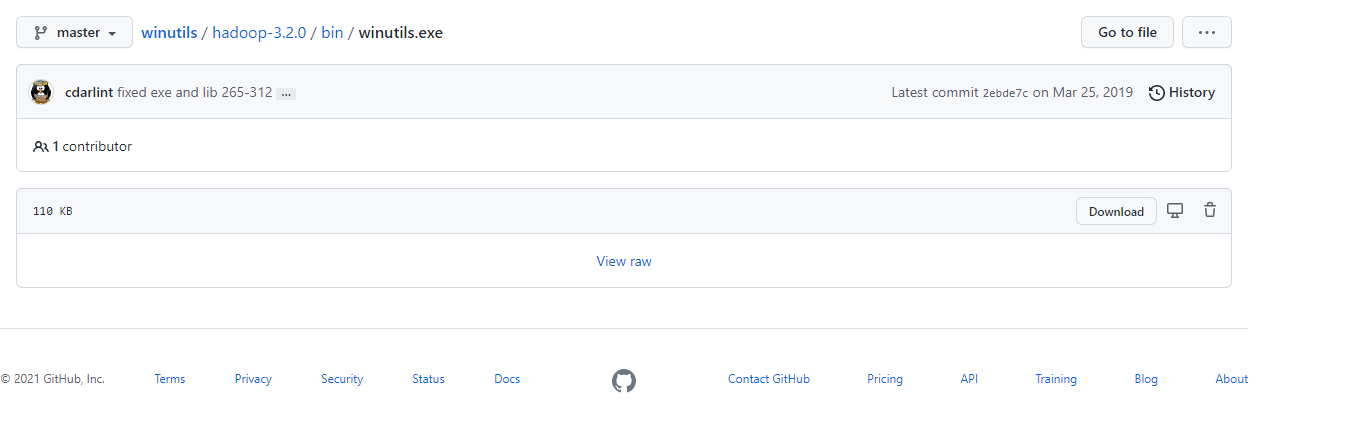
mkdir hadoop\bin

* Once the directory is created we need to download winutils.exe and copy and paste into this bin folder.
* You need to download winutils relevant to the Hadoop version you installed in Step 3. In my case it is 3.2.
* To get the winutils go to :

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

* Based on the Hadoop version you can choose one here. In my case it is 3.2. go to Hadoop-3.2.0/bin

Snapshot:



download

* Place the downloaded winutils.exe inside hadoop\bin you created earlier. You can manually copy and paste the file into the location.

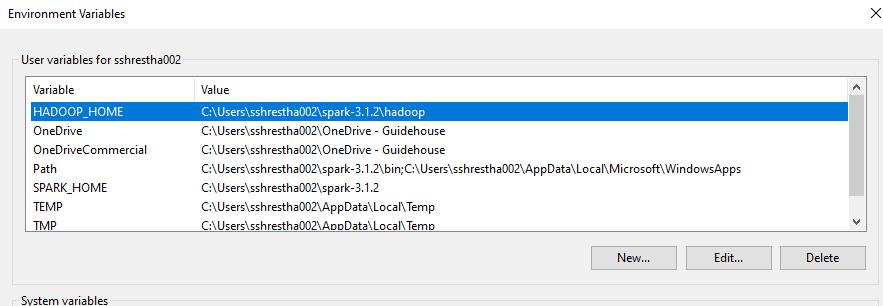
1. Create environment variables

* SPARK\_HOME, HADOOP\_HOME, PATH
* Got to search bar (near the start bottom) and search for “Edit environment variables for your account”.
* Now, create three environment variables:

Click New:

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Value | Note |
| HADOOP\_HOME | C:\Users\yourusername\spark-3.1.2\hadoop | Provide your location of spark/hadoop that you installed earlier |
| SPARK\_HOME | C:\Users\ yourusername \spark-3.1.2 |  |
| Path | C:\Users\ yourusername \spark-3.1.2\bin | If you already have a path here. Use ; to add additional path. |

* Snapsthot below:



* Once done, Click OK

1. Test Instalation of Pyspark

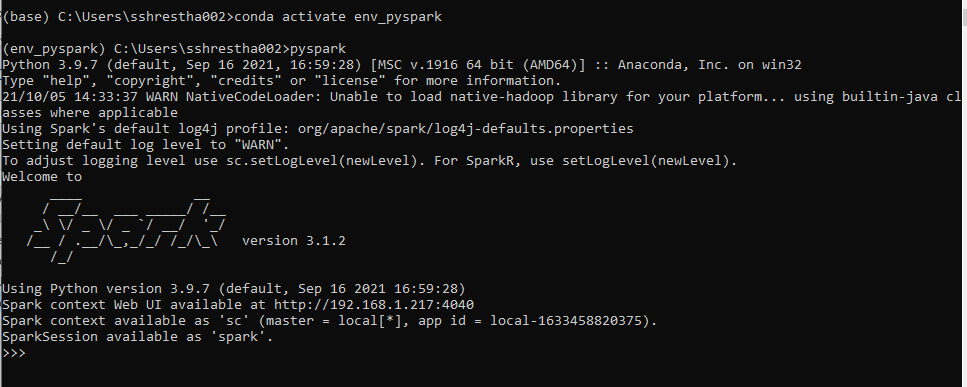
* Install python in this environement.
* Go to anaconda prompt and activate environemnt, install python and run pyspark command to test.

conda activate env\_pyspark

conda install python=3.9

pyspark

* This will open up as below.



1. Jupyter Notebook Extension

* To start jupyter notebook, you need to , Install jupyter, install ipykernel and link your environment to kernel in jupyter.

conda install jupyter)

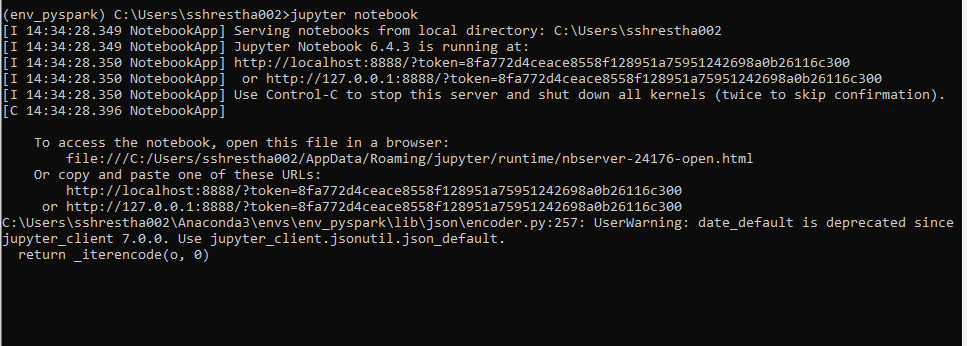
conda install ipykernel

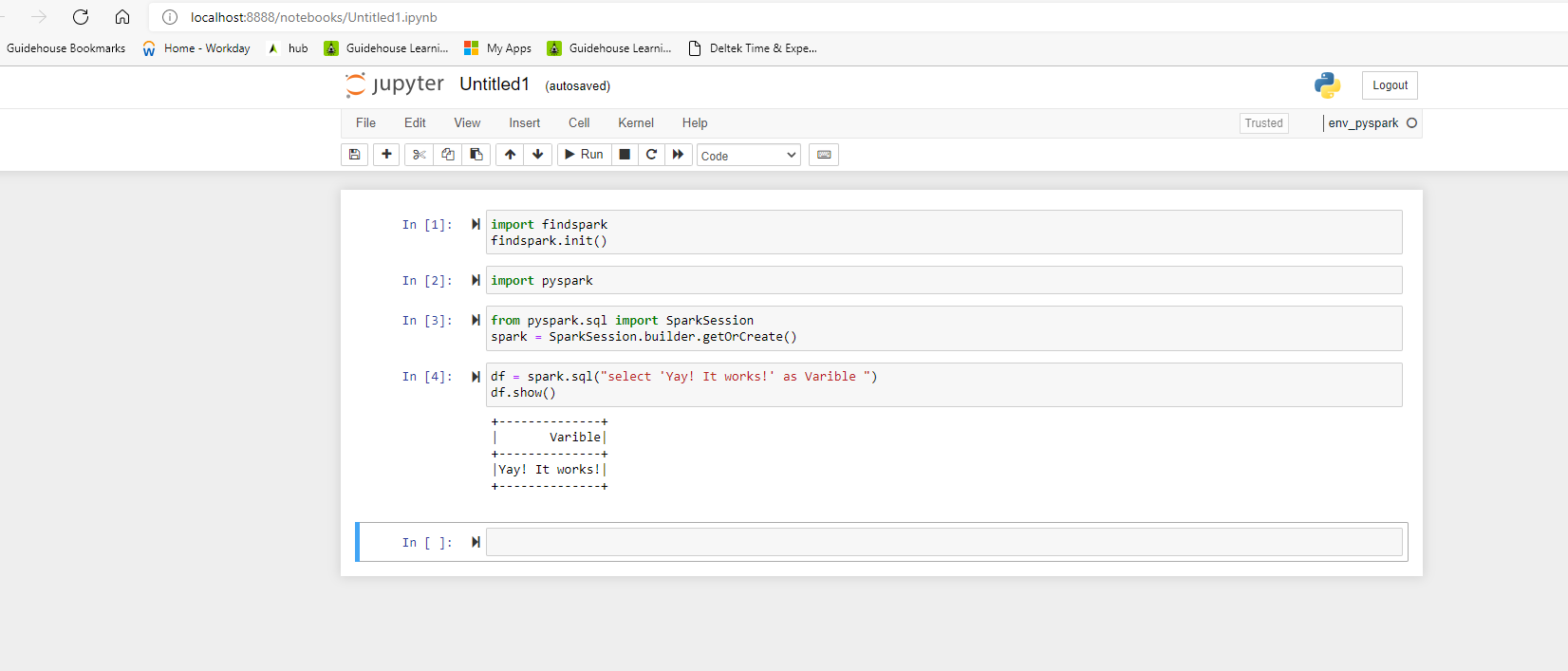
python -m ipykernel install –user –name env\_pyspark

* Check: Now in your pyspark environment run

jupyter notebook

* This will open up jupyter notebook and you can select kernel “env\_pyspark” to run pyspark .





Change your kernel here

1. Test script

* Use below script to test your pyspark:

Import pyspark

From pyspark.sql import SparkSession

Spark=SparkSession.builder.getOrCreate()

df = spark.sql(“select ‘Yay! It works!’ as Variable”)

df.show()

* Once everything is set up. Whenever you start a new session to run pyspark in jupyter, follow below steps:
* Open anaconda prompt and run below command to activate environment and jupyter notebook

conda activate env\_pyspark

jupyter notebook

End of the Process