

Installation Guide for Jupyter Notebook and Apache Spark, PySpark Setup - Windows

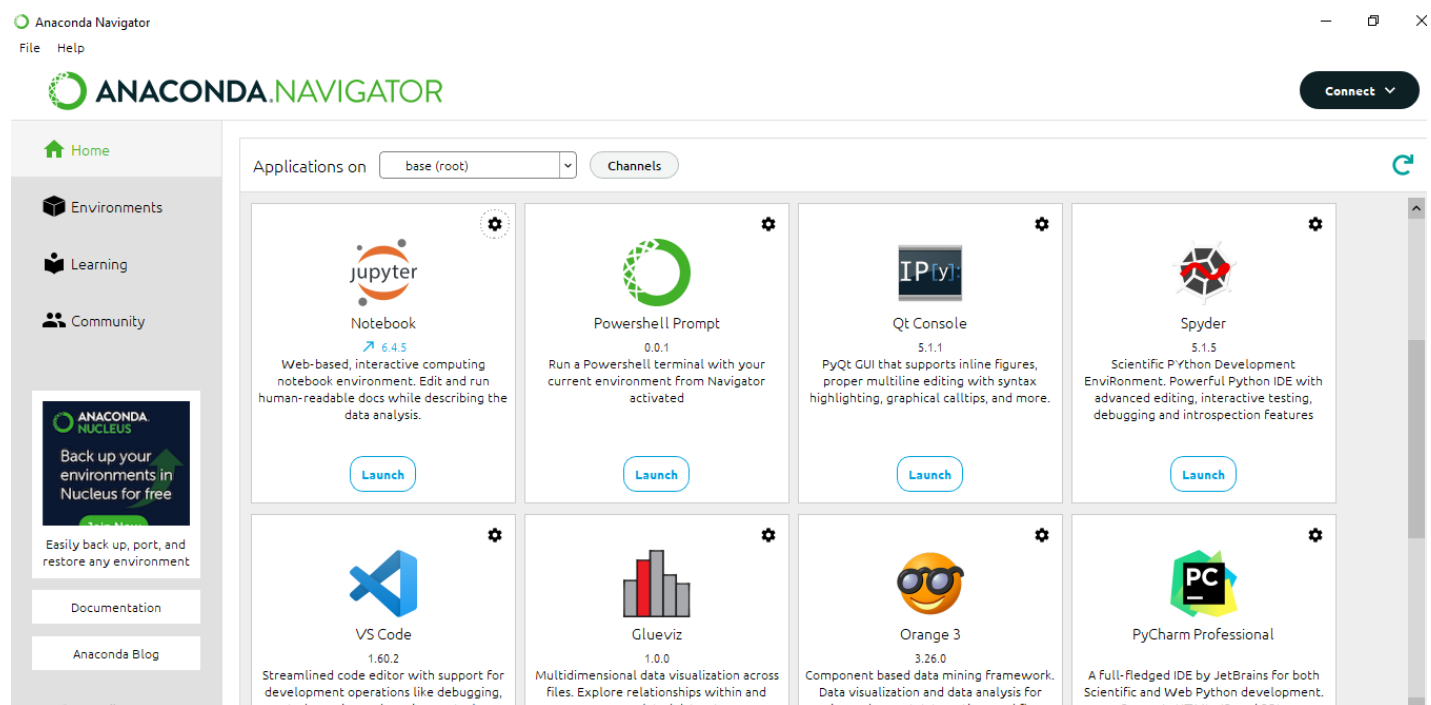
Install Python and Jupyter Notebook

Jupyter Notebook can be installed by using Anaconda:

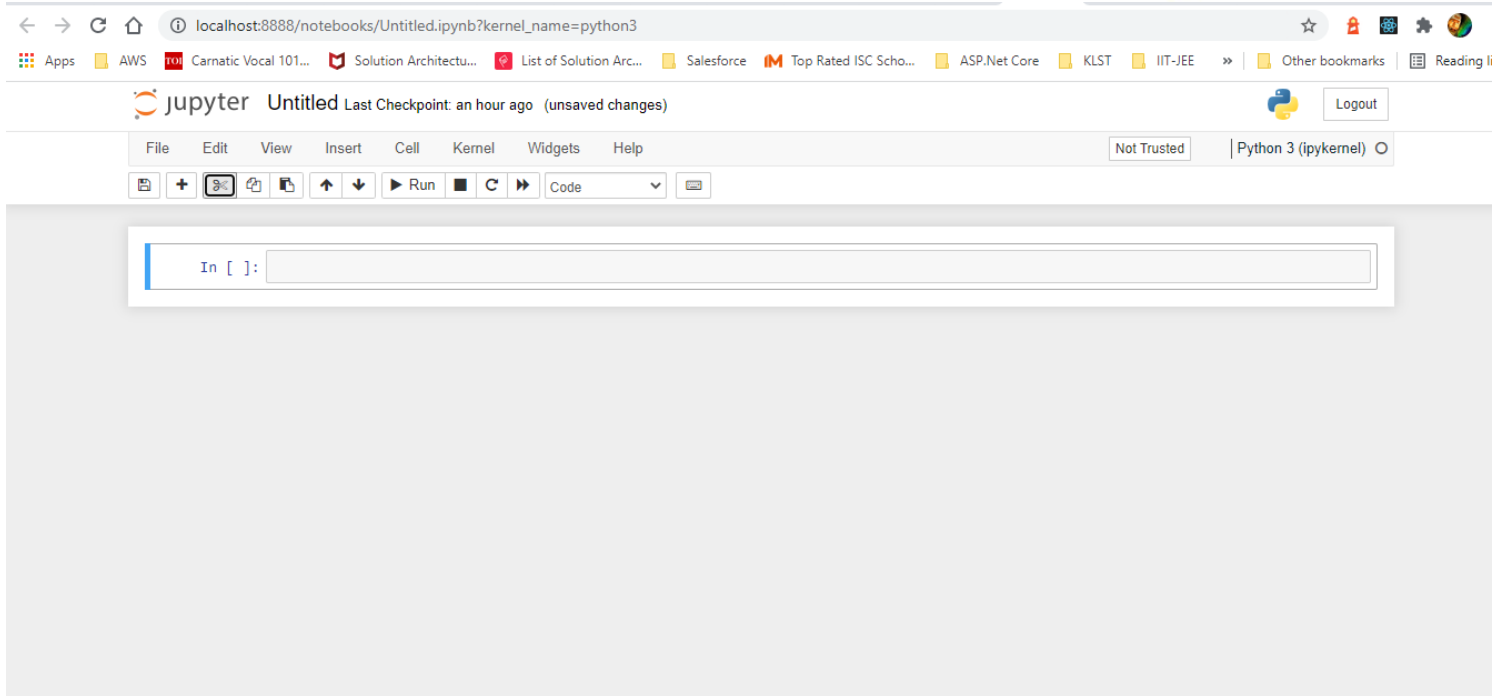
<https://www.anaconda.com/products/individual#windows>

Install Python and Jupyter using the Anaconda Distribution, which includes Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.

Post Anacondo installed go to **Anaconda Launcher** and choose Jupyter and launch



Once jupyter is successfully installed, you should see it automatically loads the jupyter notebook on your default browser at <http://localhost:8888>



Install Java 8

Before you start with spark and Hadoop, first make sure you have java 8 installed, or to install it.

Check if JAVA is installed

Open cmd (windows command prompt) from start menu and run:

java -version

If not available download and install Java 8 from the official site of Java

Post that verify if the respective Java environment variables are set, if not

Add the following environment variable:

JAVA_HOME = C:\Program Files\Java\jdk1.8.0_201

Add to PATH variable the following directory:

C:\Program Files\Java\jdk1.8.0_201\bin

Download and Install Spark

Go to Spark home page, and download the .tgz file from 3.2.0 version



COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"



Projects ▾

People ▾

Community ▾

License ▾

Sponsors ▾

We suggest the following site for your download:

<https://d1cdn.apache.org/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz>

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

HTTP

<https://d1cdn.apache.org/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz>

BACKUP SITE

<https://downloads.apache.org/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz>

After downloading the .tgz file extract the tgz file to your chosen directory and keep it inside a folder for e.g

C:\dell-dataengg\apachespark.

There is another compressed directory inside the tar, extract that file also to the same folder as mentioned above.

Add the environment variables to windows Environment Variables under system variables

SPARK_HOME = C:\dell-dataengg\apachespark\spark-3.2.0-bin-hadoop3.2

HADOOP_HOME = C:\ dell-dataengg \apachespark\spark-3.2.0-bin-hadoop3.2

Add the below path to **PATH** environment variable to windows Environment Variables under system variables:

C:\ dell-dataengg \apachespark\spark-3.2.0-bin-hadoop3.2\bin

Download and setup winutils.exe

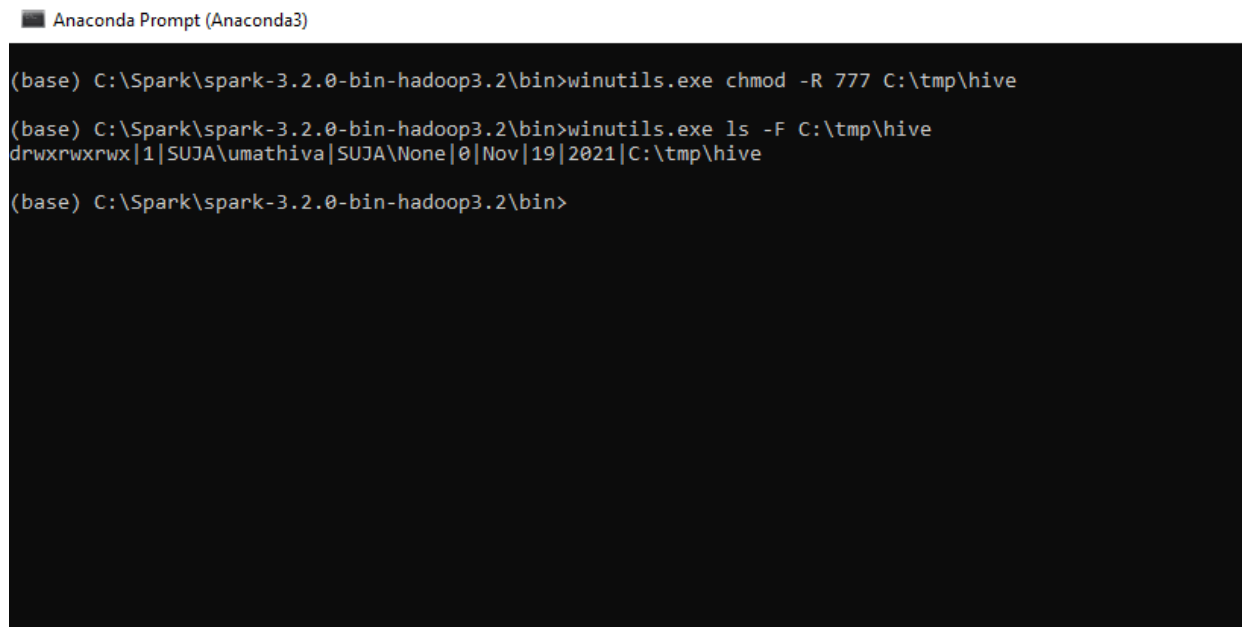
In hadoop binaries repository, <https://github.com/steveloughran/winutils> choose hadoop version 3.0.0, then goto bin, and download the winutils.exe file.

Save winutils.exe in to bin directory of your spark installation, **SPARK_HOME\bin** directory

Create the folder **C:\tmp\hive**

1. Execute the following command in **cmd** started using the option **Run as administrator**.

```
winutils.exe chmod -R 777 C:\tmp\hive  
winutils.exe ls -F C:\tmp\hive
```



Anaconda Prompt (Anaconda3)

```
(base) C:\Spark\spark-3.2.0-bin-hadoop3.2\bin>winutils.exe chmod -R 777 C:\tmp\hive  
(base) C:\Spark\spark-3.2.0-bin-hadoop3.2\bin>winutils.exe ls -F C:\tmp\hive  
drwxrwxrwx|1|SUJA\umathiva|SUJA\None|0|Nov|19|2021|C:\tmp\hive  
(base) C:\Spark\spark-3.2.0-bin-hadoop3.2\bin>
```

After executing the command and setting the Path variable now from anaconda prompt type **pyspark**, to enter pyspark shell. To be prepared, best to check it in the python environment from which you run jupyter notebook.