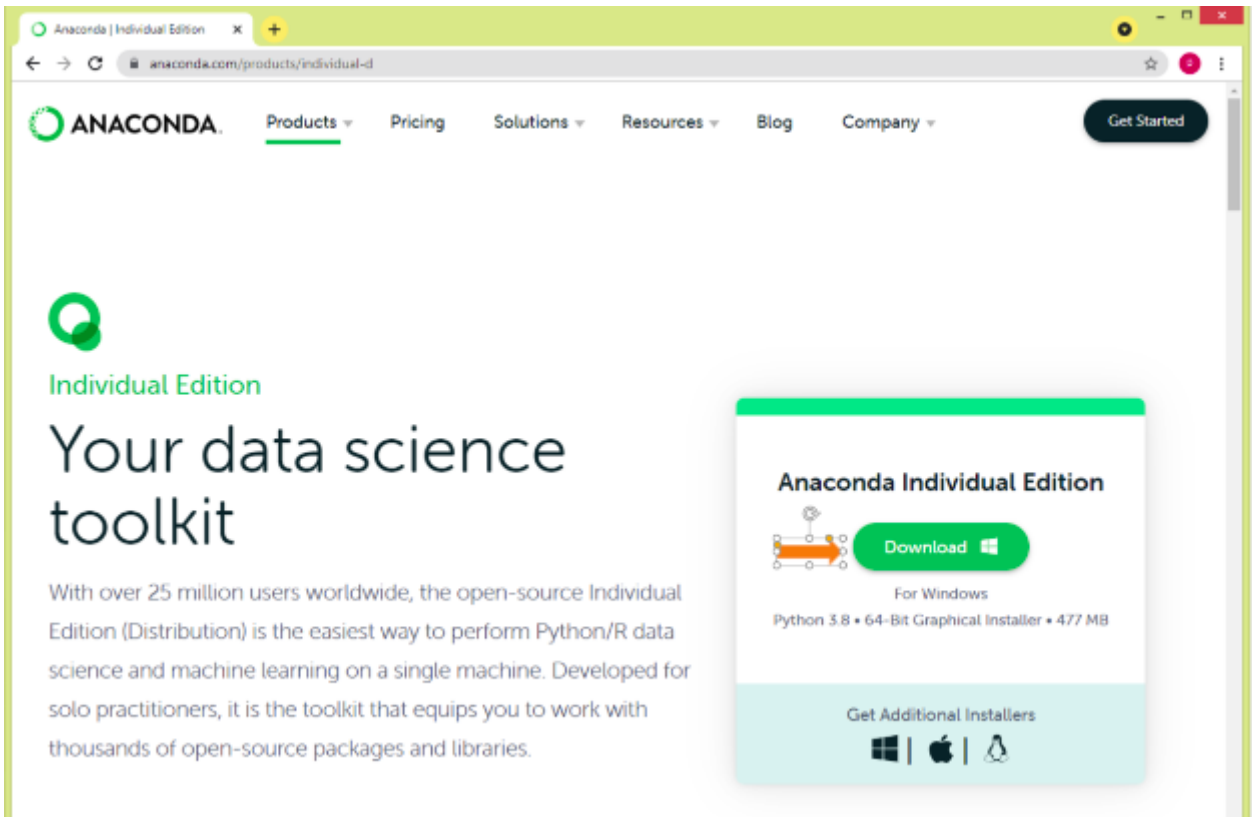
Install PySpark in Anaconda & Jupyter Notebook

Steps to Install PySpark in Anaconda & Jupyter notebook

- Step 1. Download & Install Anaconda Distribution
- Step 2. Install Java
- Step 3. Install PySpark
- Step 4. Install FindSpark
- Step 5. Validate PySpark Installation from pyspark shell
- Step 6. PySpark in Jupyter notebook
- Step 7. Run PySpark from IDE

1. Download & Install Anaconda Distribution

Go to https://anaconda.com/ and select **Anaconda Individual Edition** to download the Anaconda and install, for windows you download the `.exe` file and for Mac download the `.pkg` file.



.

2. Install Java

PySpark uses Java underlying hence you need to have Java on your Windows or Mac. Since Java is a third party, you can install it using the Homebrew command `brew`. Since Oracle Java is not open source anymore, I am using the OpenJDK version 11. Open Terminal from Mac or command prompt from Windows and run the below command to install Java.'

```
# Install OpenJDK 11
conda install openjdk
```

The following Java version will be downloaded and installed. Depending on OS and version you are using the installation directory would be different.

3. Install PySpark

To install PySpark on Anaconda I will use the `conda` command. conda is the package manager that the Anaconda distribution is built upon. It is a package manager that is both cross-platform and language agnostic.

```
# Install PySpark using Conda
conda install pyspark
```

.

4. Install FindSpark

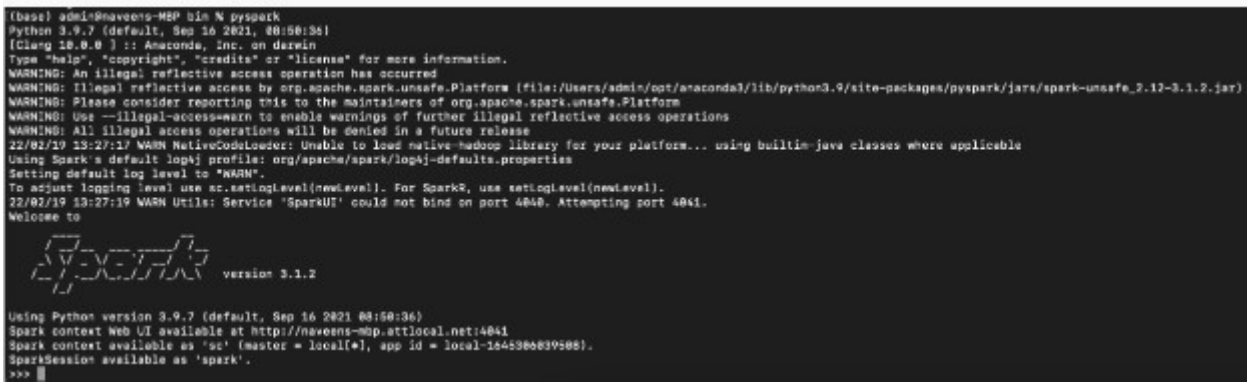
In order to run PySpark in Jupyter notebook first, you need to find the PySpark Install, I will be using `findspark` package to do so. Since this is a third-party package we need to install it before using it.

```
conda install -c conda-forge findspark
```

5. Validate PySpark Installation

Now let's validate the PySpark installation by running `pyspark` shell. This launches the PySpark shell where you can write PySpark programs interactively.

.



6. Install Jupyter notebook & run PySpark

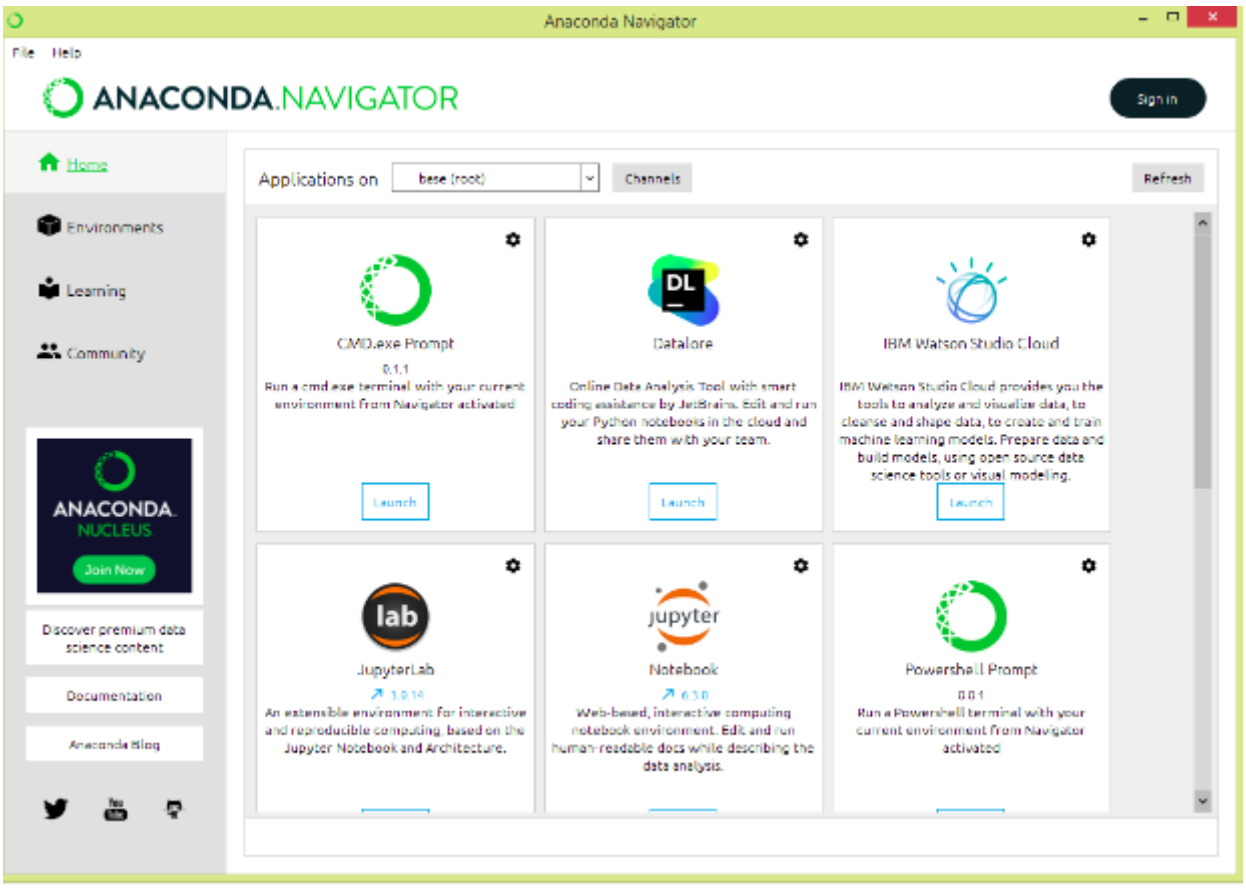
With the last step, PySpark install is completed in Anaconda and validated the installation by launching PySpark shell and running the sample program now, let's see how to run a similar PySpark example in Jupyter notebook.

Post-install, Open Jupyter by selecting `Launch` button.

7. Run PySpark from Spyder IDE

Here I will use Spyder IDE.

If you don't have Spyder on Anaconda, just install it by selecting `Install` option from navigator.
post install, write the below program and run it by pressing F5 or by selecting a run button from the menu.



.

7. Run PySpark from Spyder IDE

Here I will use Spyder IDE.

If you don't have Spyder on Anaconda, just install it by selecting `Install` option from navigator.
post install, write the below program and run it by pressing F5 or by selecting a run button from the menu.

```
# Import PySpark


from pyspark.sql import SparkSession


#Create SparkSession

spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
```

```
# Data
data = [("Java", "20000"), ("Python", "100000"), ("Scala", "3000")]

# Columns
columns = ["language","users_count"]

# Create DataFrame
df = spark.createDataFrame(data).toDF(*columns)

# Print DataFrame
df.show()
```