Guide to install Spark and use PySpark from Jupyter in Linux

Jupyter is one of the powerful tools for development. However, it doesn't support Spark development implicitly. This article aims to simplify that and enable the users to use the Jupyter itself for developing Spark codes with the help of PySpark. Kindly follow the below steps to get this implemented and enjoy the power of Spark from the comfort of Jupyter.

1- Install Python and Jupyter Notebook

The following links provide a step-by-step tutorial for setting up a virtual environment and installing Python and Jupyter Notebook:

- Linux
- Windows
- macOS

After installing Jupyter, make a new directory to put all your notebooks. Move into the directory and start the Jupyter notebook using the following command at the Terminal (Mac/Linux) or Command Prompt (Windows):

```
jupyter notebook
```

Once it is started you can see the logging.

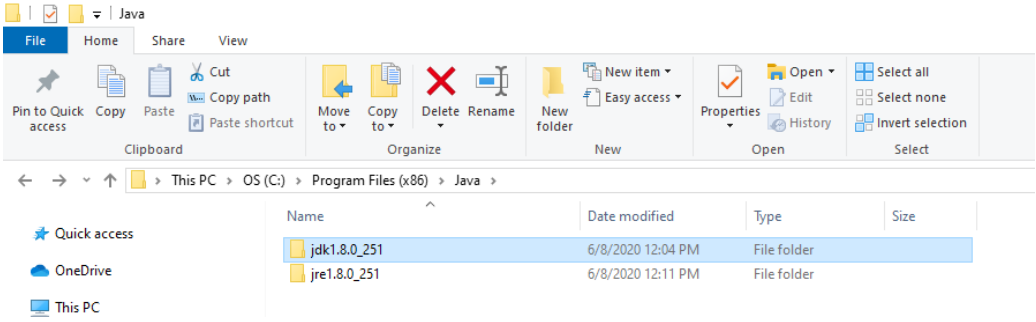
2. Install JDK v8

To run PySpark application, you would need Java 8 or later version (JDK v8 is recommended). Download the Java from Oracle and install it on your system.

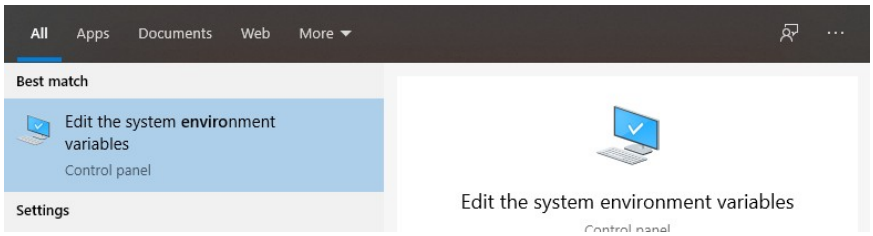
Remember to set the environment variable `JAVA_HOME` and `PATH` to use JDK v8 if your installation does not do it automatically for you:

- **Windows:**

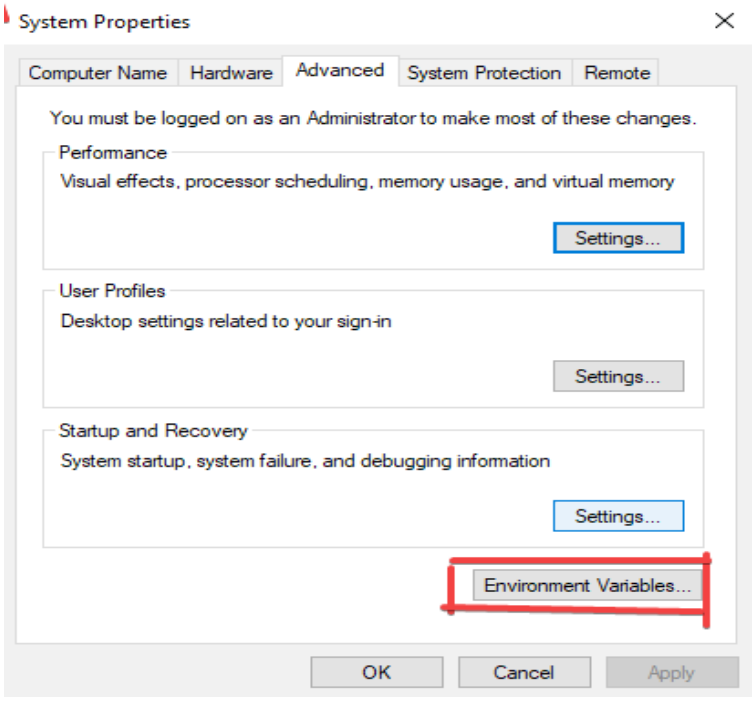
Find the Java path:



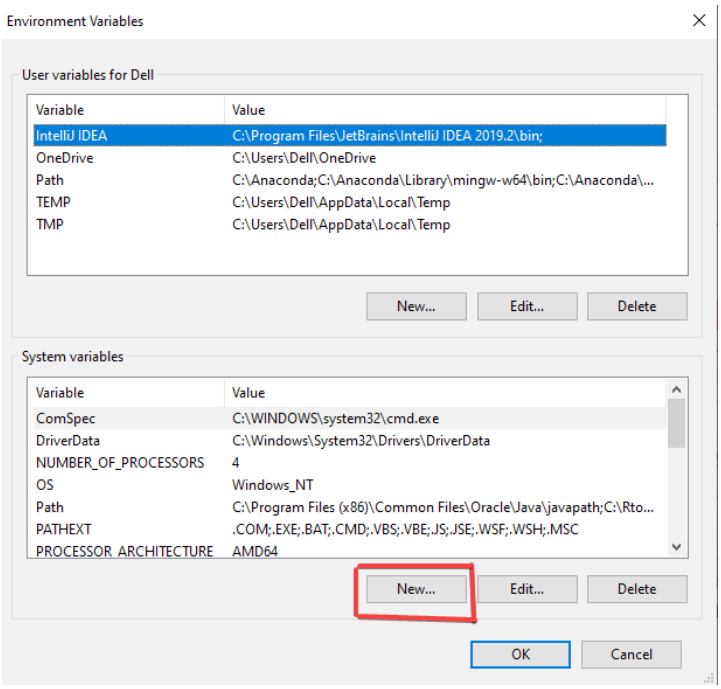
Go to the search bar and "EDIT THE ENVIRONMENT VARIABLES:



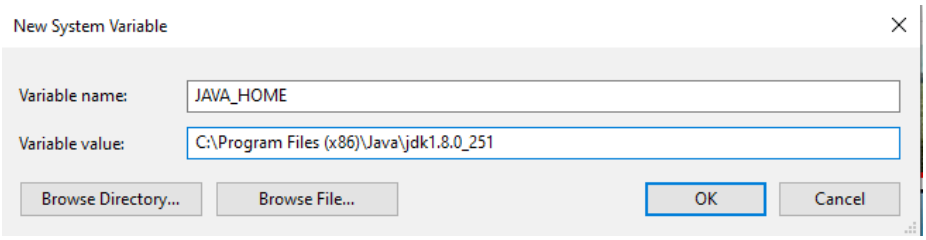
Click into the "Environment Variables":



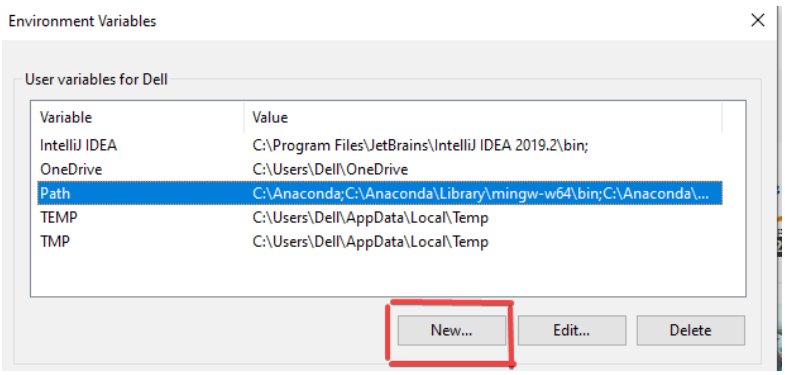
Click into "New" to create your new Environment variable:



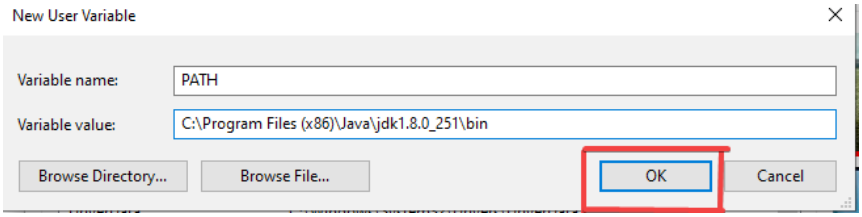
Use Variable Name as "JAVA_HOME' and your Variable Value as 'C:\Program Files (x86)\Java\jdk1.8.0_251'. This is your location of the Java file. Click 'OK' after you've finished the process:



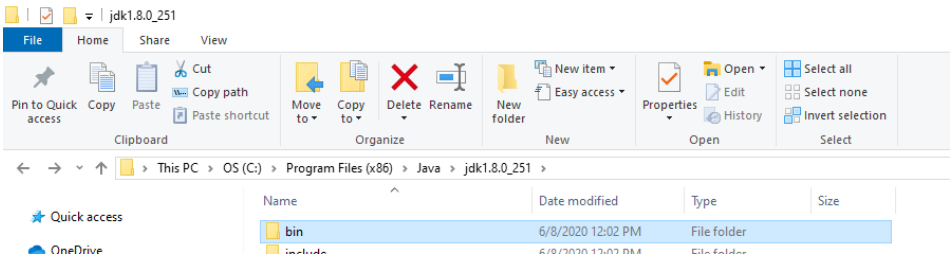
Let's add the User variable and select 'Path' and click 'New' to create it:



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.



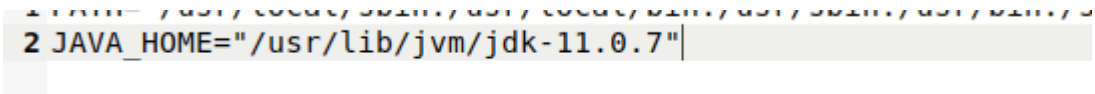
- **Linux:**

you can check your java version using 'java --version' command. For configuring environment variables, let's open the '**gedit**' text editor using the following command:

```
sudo gedit /etc/environment
```

Locate the java folder in your machine. It might be /usr/lib/jvm/java-8-openjdk-amd64. Let's make the change by providing the following information where the 'Java' path is specified:

```
JAVA_HOME=”/usr/lib/jvm/java-8-openjdk-amd64”
```



To make a final change, let's type the following command:

```
source /etc/environment
```

3. install PySpark

We will install PySpark using PyPi. To install just run the following command from Instruction_pyspark_localinside the virtual environment:

```
$ pip install pyspark
```

Now you can use pyspark in jupyter notebook.

Reference:

[1] https://www.javacodemonk.com/installing-pyspark-with-jupyter-notebook-on-ubuntu-18-04-lts-31cd3781