ey-advisory-technology-testing / **etl-step-definitions** Internal

Code Issues Pull requests Actions Projects Wiki Security

Installation

Edit New Page

Jump to bottom

Gajendra Varadhan edited this page on Dec 23, 2020 · 1 revision

Spark

For Mac OS, there is no further installations needed. For Windows 10 and above, please follow the below steps for Installing Apache Spark in your machine.

Prerequisites

- A system running Windows 10
- A user account with administrator privileges (required to install software, modify file permissions, and modify system PATH)
- Command Prompt or Powershell
- A tool to extract .tar files, such as 7-Zip

Step 1: Install Java 11 or above

Apache Spark requires Java 11 or above. I currently have Java 14 which works without issues. You can check to see if Java is installed using the command prompt.

Open the command line by clicking Start > type cmd > click Command Prompt.

Type the following command in the command prompt:

```
java -version
```

If you don’t have Java installed:

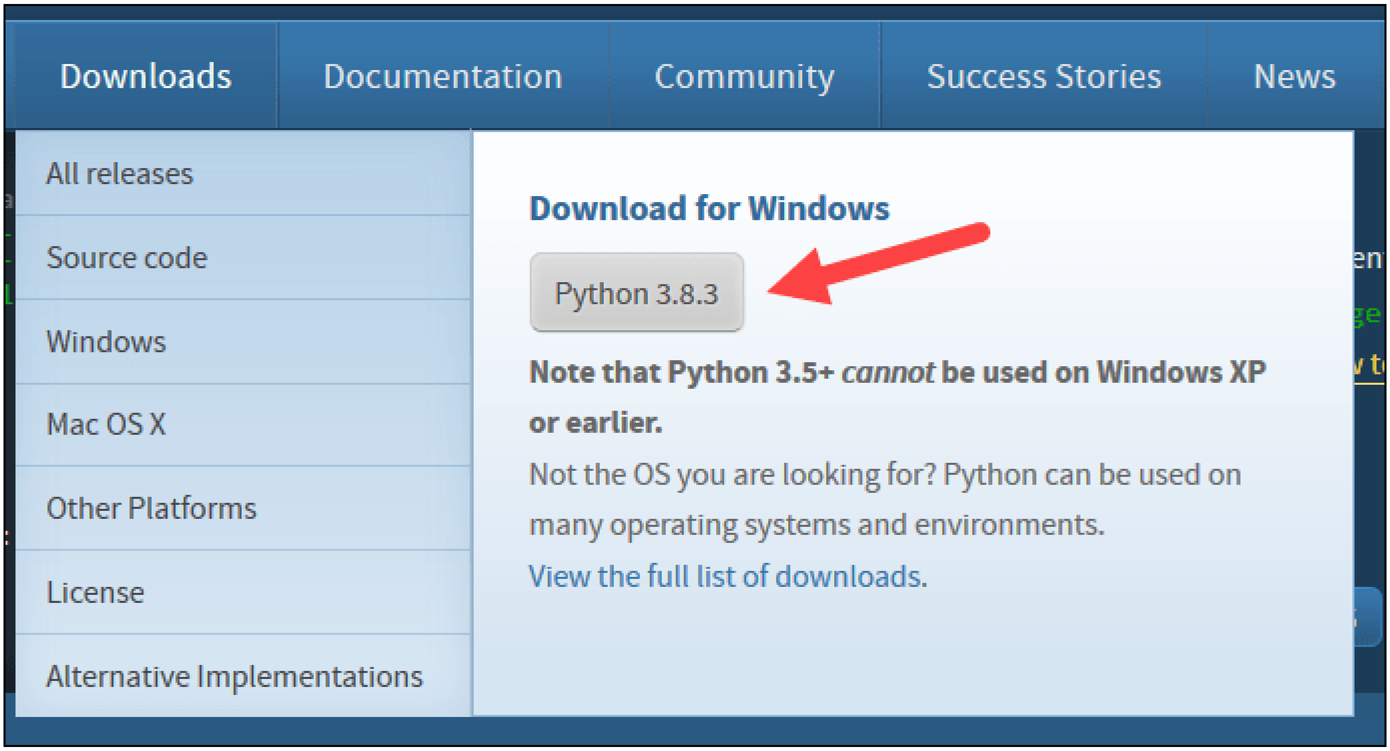
Download it from https://jdk.java.net/

Step 2: Install Python (Optional)

1. To install the Python package manager, navigate to https://www.python.org in your web browser.

2. Mouse over the Download menu option and click **Python 3.8.3**. **3.8.3** is the latest version at the time of writing this manual.

3. Once the download finishes, run the file.

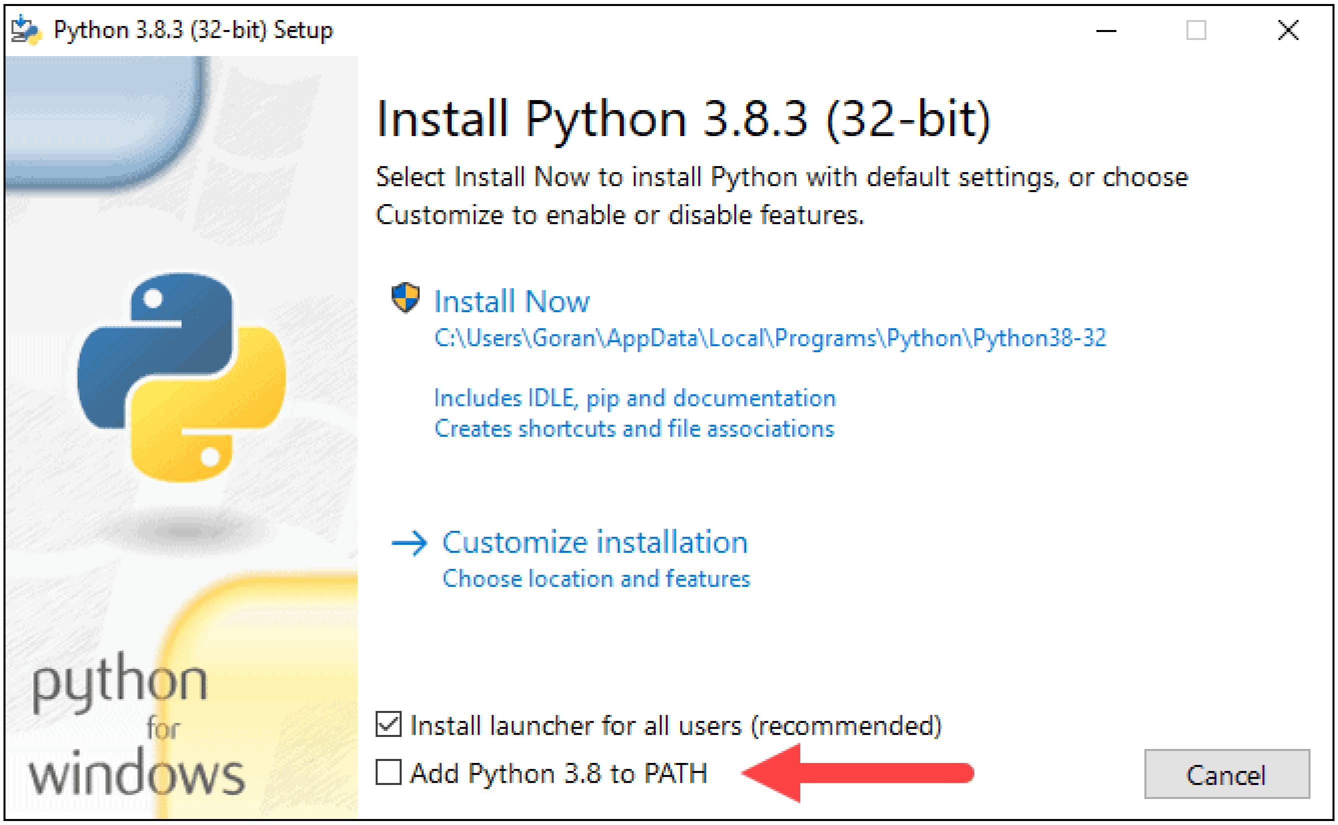


Note that Python 3.5+ *cannot* be used on Windows XP or earlier.

Not the OS you are looking for? Python can be used on many operating systems and environments.

4. Near the bottom of the first setup dialog box, check off Add Python 3.8 to PATH. Leave the other box checked.

5. Next, click Customize installation.



Python 3.8.3 (32-bit) Setup

Install Python 3.8.3 (32-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

Install Now
C:\Users\Goran\AppData\Local\Programs\Python\Python38-32

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

☑ Install launcher for all users (recommended)

□ Add Python 3.8 to PATH

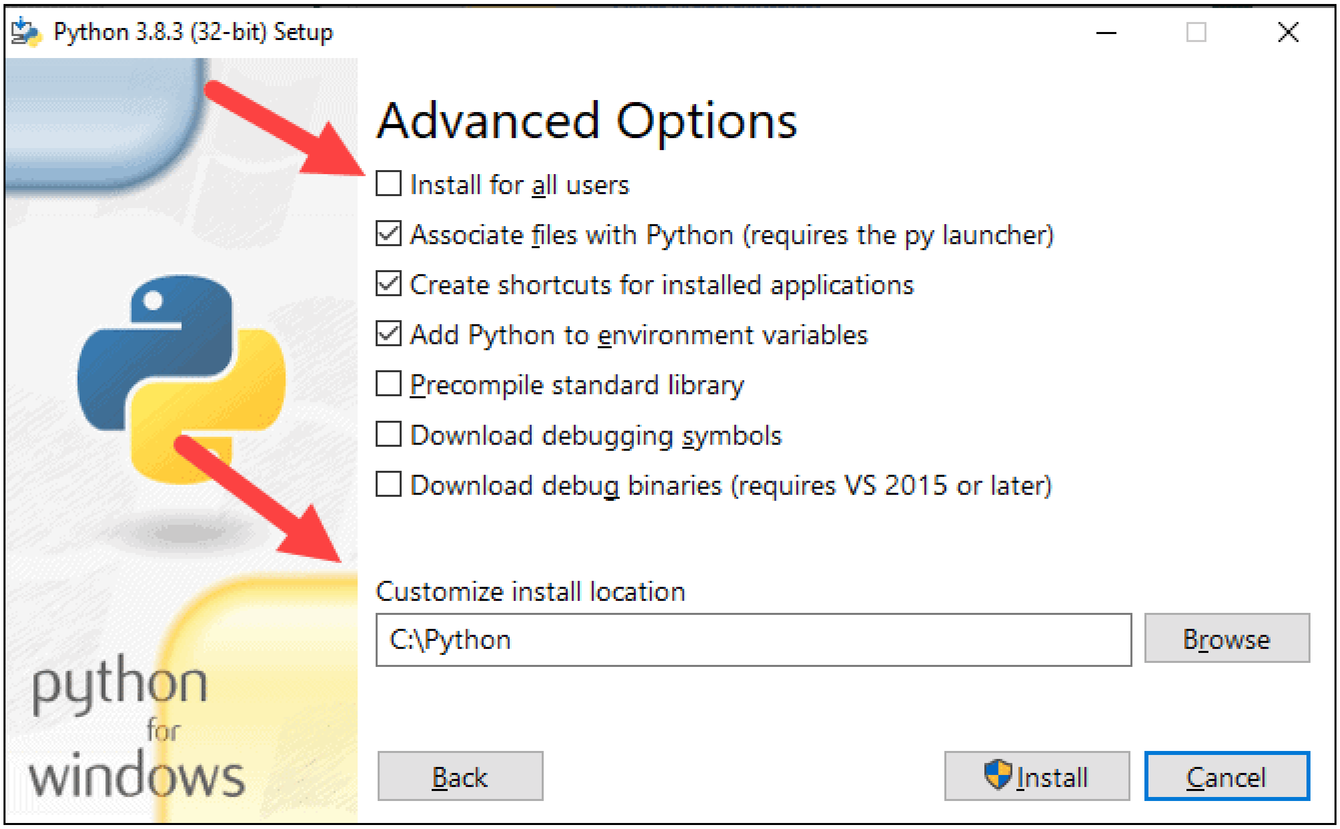
6. You can leave all boxes checked at this step, or you can uncheck the options you do not want.

7. Click Next.

8. Select the box Install for all users and leave other boxes as they are.

9. Under Customize install location, click Browse and navigate to the C drive. Add a new folder and name it Python.

10. Select that folder and click OK.



Advanced Options

☐ Install for all users
☑ Associate files with Python (requires the py launcher)
☑ Create shortcuts for installed applications
☑ Add Python to environment variables
☐ Precompile standard library
☐ Download debugging symbols
☐ Download debug binaries (requires VS 2015 or later)

Customize install location

C:\Python

11. Click Install, and let the installation complete.
12. When the installation completes, click the Disable path length limit option at the bottom and then click Close.
13. If you have a command prompt open, restart it. Verify the installation by checking the version of Python:

```
python --version
```

The output should print `Python 3.8.3` .

Step 3: Download Apache Spark

1. Open a browser and navigate to https://spark.apache.org/downloads.html.
2. Under the Download Apache Spark heading, there are two drop-down menus. Use the current non-preview version.

It can be any latest version. In our case, let's choose a Spark release from the drop-down menu `3.0.1 (Sep 02, 2020)` . In the second drop-down Choose a package type, select `Pre-built for Apache Hadoop 3.2 and later` . 3. Click the **spark-3.0.1-bin-hadoop3.2.tgz** link. 4. A page with a list of mirrors loads where you can see different servers to download from. Pick any from the list and save the file to your Downloads folder.

Step 5: Install Apache Spark

Installing Apache Spark involves extracting the downloaded file to the desired location.

1. Extract spark-3.0.1-bin-hadoop3.2.tgz to the desired location (say **C:**). Usually the content is again an archive file. So, extract again and you can see the bin folder.

Step 6: Add winutils.exe and hadoop.dll files

Download the winutils.exe and hadoop.dll files for the Hadoop version selected above.

1. Navigate to this URL https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.1/bin, download the winutils.exe, and hadoop.dll files by clicking on them and clicking the download button on the top right.
2. Paste both the files under the bin folder of the spark extracted above.

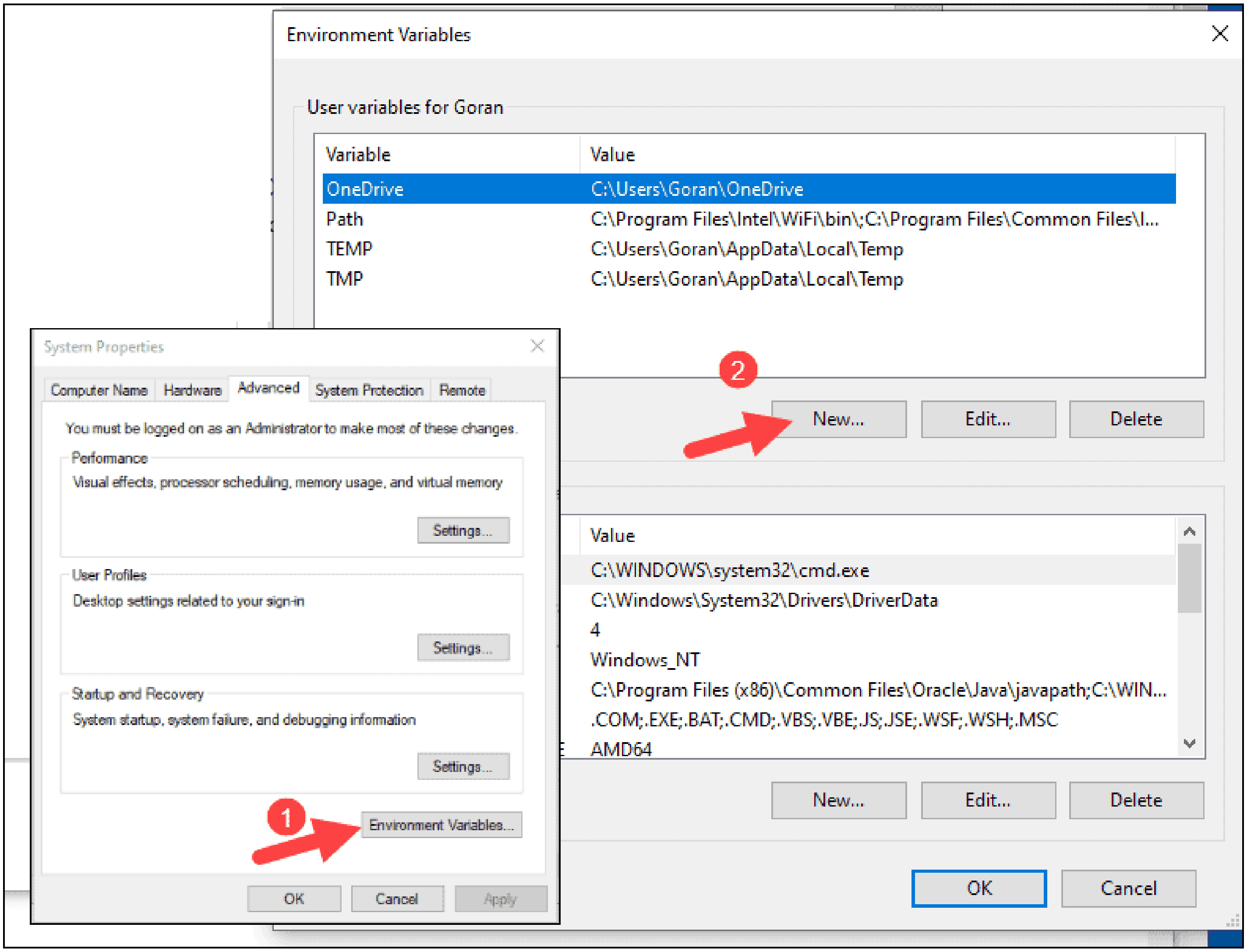
Step 7: Configure Environment Variables

This step adds the Spark and Hadoop locations to your system PATH. It allows you to run the Spark shell directly from a command prompt window.

1. Click Start and type environment.

2. Select the result labeled Edit the system environment variables.

3. A System Properties dialog box appears. In the lower-right corner, click Environment Variables and then click New in the next window.



Environment Variables

User variables for Goran

Variable	Value
OneDrive	C:\Users\Goran\OneDrive
Path	C:\Program Files\Intel\WiFi\bin\;C:\Program Files\Common Files\I...
TEMP	C:\Users\Goran\AppData\Local\Temp
TMP	C:\Users\Goran\AppData\Local\Temp

Value
C:\WINDOWS\system32\cmd.exe
C:\Windows\System32\Drivers\DriverData
4
Windows_NT
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WIN...
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
AMD64

System Properties

You must be logged on as an Administrator to make most of these changes.

Performance
Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles
Desktop settings related to your sign-in

Startup and Recovery
System startup, system failure, and debugging information

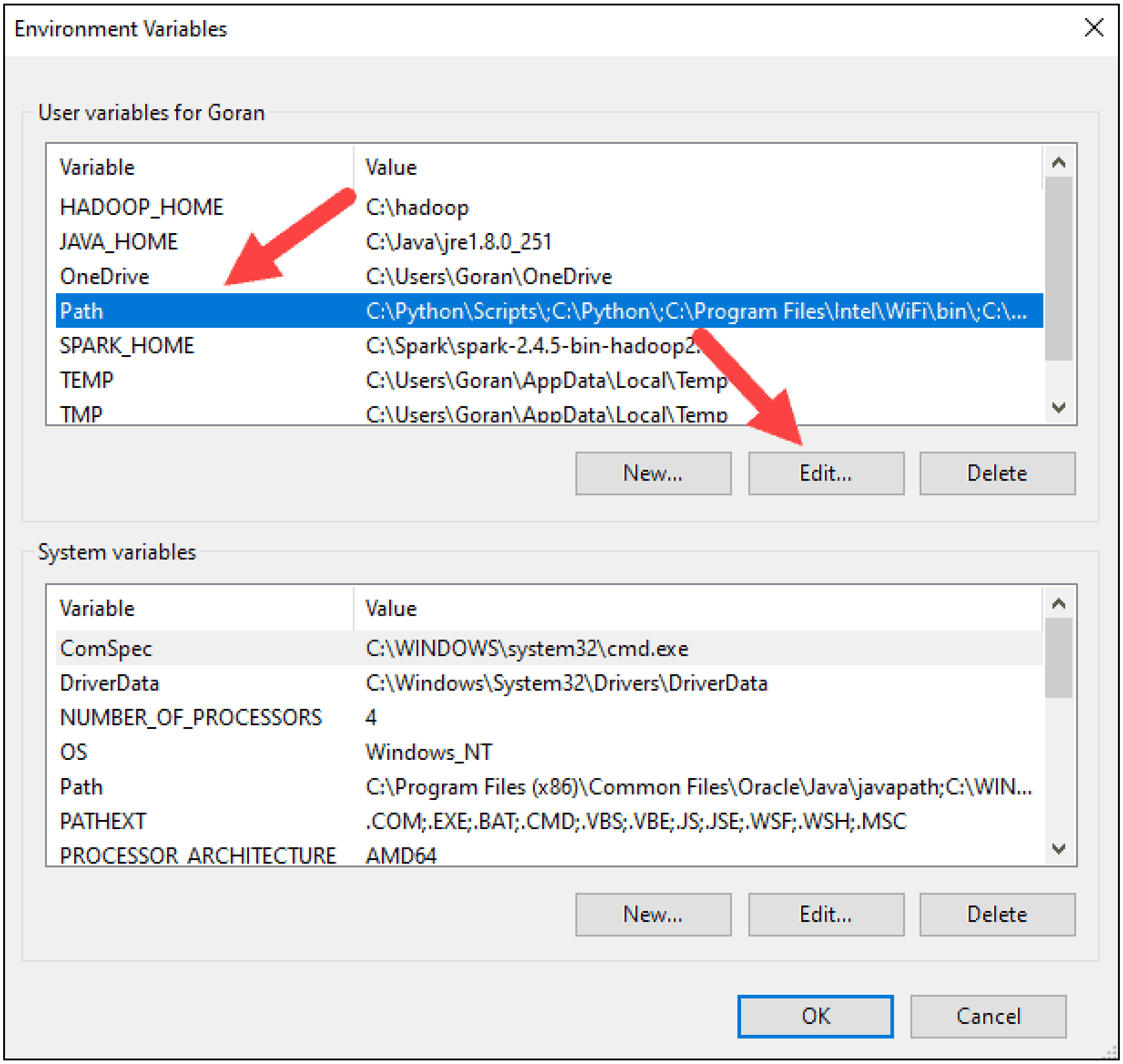
4. Create a new Variable with **Name **as **SPARK_HOME**.

5. For Variable **Value** type the folder where the spark bin is located, and click OK. If you changed the folder path, use that one instead.

6. Create another variable with **Name **as **HADOOP_HOME**

7. For the value, type **%SPARK_HOME%**

8. Edit the variable named Path.



Environment Variables

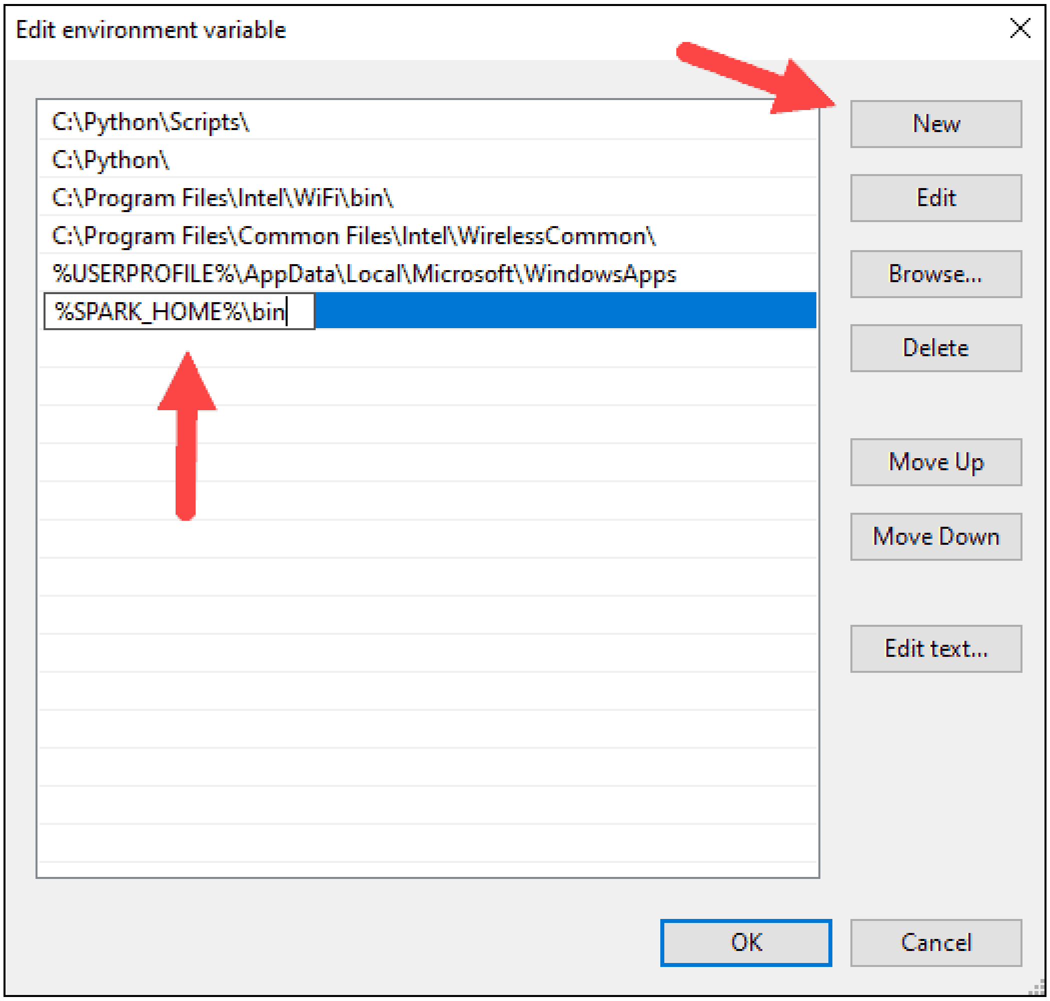
User variables for Goran

Variable	Value
HADOOP_HOME	C:\hadoop
JAVA_HOME	C:\Java\jre1.8.0_251
OneDrive	C:\Users\Goran\OneDrive
Path	C:\Python\Scripts\;C:\Python\;C:\Program Files\Intel\WiFi\bin\;C:\...
SPARK_HOME	C:\Spark\spark-2.4.5-bin-hadoop2.
TEMP	C:\Users\Goran\AppData\Local\Temp
TMP	C:\Users\Goran\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

9. Add an entry like **%SPARK_HOME%\bin**



```
C:\Python\Scripts\
C:\Python\
C:\Program Files\Intel\WiFi\bin\
C:\Program Files\Common Files\Intel\WirelessCommon\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
%SPARK_HOME%\bin
```

10. Click **OK** to close all open windows.

11. Reboot your machine for changes to take effect.

12. Open command prompt/terminal and type `spark-shell` . It will start Spark 3.0.1 successfully

+ Add a custom footer

Pages 3

Find a Page...

- Home
- Installation
- Using a GitHub hosted dependency

+ Add a custom sidebar

Clone this wiki locally

```
https://github.com/ey-advisory-technology-testing/etl-step-definitions.wiki.git
```