Follow the steps below to install PySpark on Windows.

Installing Prerequisites:

1. Java

PySpark requires Java 1.8.x
Check if Java is already available by opening the command prompt and giving the command as shown here.

```
C:\>java -version
java version "1.8.0_281"
Java(TM) SE Runtime Environment (build 1.8.0_281-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.281-b09, mixed mode)
```

If not available, then download JDK that is appropriate to your system from the link below and install it.
https://www.oracle.com/java/technologies/downloads/#java8-windows

If you are using Windows 64-bit operating system, then download and install *x64 Installer*. Else if you are using Windows 32-bit operating system, then download and install *x86 Installer*. Please note that you will need administrator privileges on the system.

After installing check as shown above and make sure the correct version is installed and available. Also, make a note of the folder where it is installed as we will need the path later.

2. Python

Python 3.x is required to run PySpark.
Check if Python is already from the command prompt as shown here.

```
C:\>python --version
Python 3.9.7
```

If not available, then install Python 3.x by downloading Anaconda Individual Edition from the link below.
https://www.anaconda.com/products/individual

You can also download the Python installer from the link below and install it.
https://www.python.org/downloads/windows/

However, Anaconda is preferred as it includes popular IDEs.

While installing make sure that Python installation folder is added to the path.

After installing check as shown above and make sure the correct version is installed and available. Also, make a note of the folder where it is installed as we will need the path later.

Installing Spark:

Download Spark from Apache Spark web site:
`https://spark.apache.org/downloads.html`

Choose Spark release: preferably `3.1.2`
Most importantly choose package type: `Pre-build for Apache Hadoop 2.7`

Copy the above file into a folder `C:\spark_setup` Make sure not to use spaces in the folder names.

Extract the files into a sub-folder say `spark-3.1.2-bin-hadoop2.7` using a utility like `7-Zip` or `WinRAR` that is available on your system.

Note that if you use 7-Zip, from the above `.tgz` file it first extracts `.tar` file. You need to use `7-Zip` on the `.tar` file one more time and extract the files and folders from it.

Make a note of the folder where it is installed as we will need the path later.

Adding `winutils.exe`:

To run Apache Spark on windows, you need `winutils.exe`

Download it from:
`https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe`

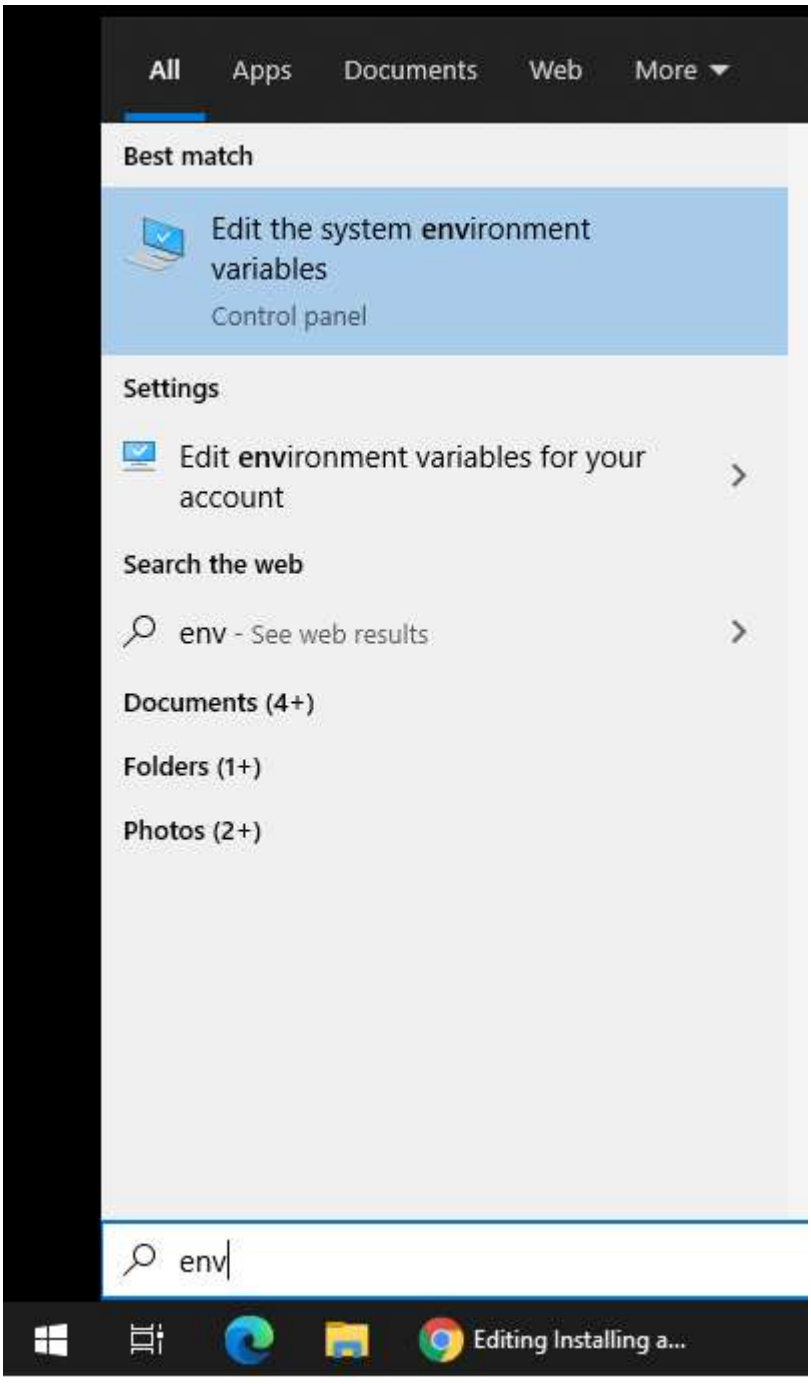
Make sure you download the `winutils.exe` file corresponding to the Spark and Hadoop version you are using.

Create a folder `C:\hadoop_utils` and a sub-folder named `bin` in it. Make sure you do not use spaces in the folder names. Now copy the file winutils.exe in the sub-folder `bin`.

Setting up Environment Variables:

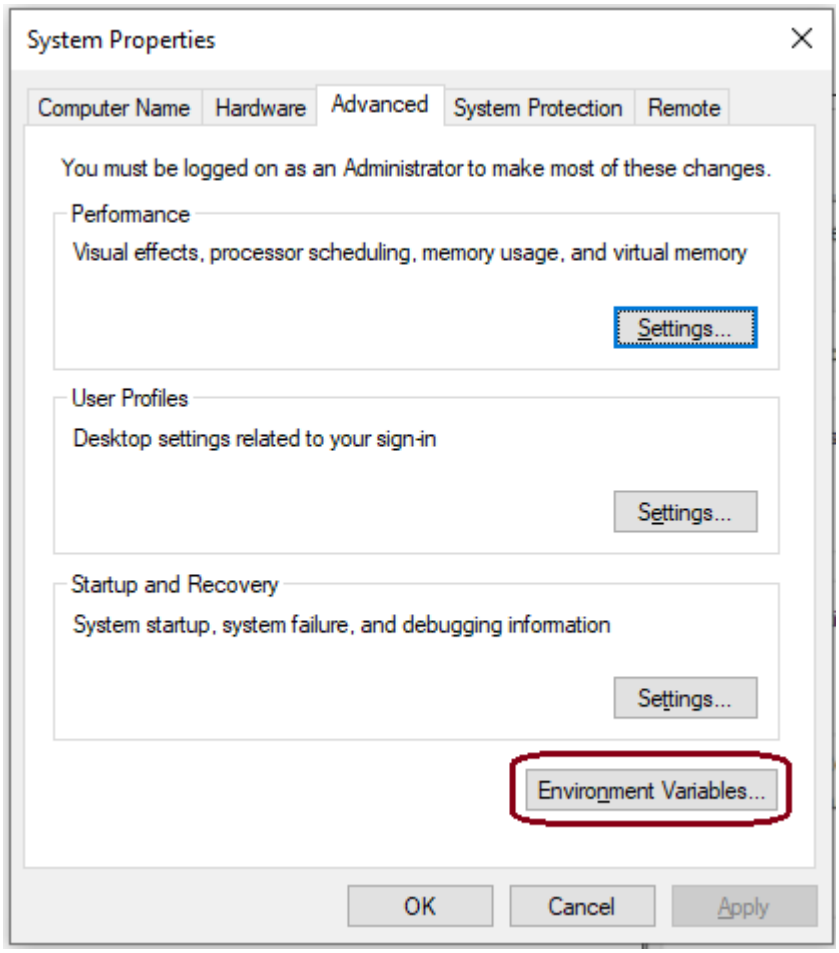
Environment variables need to be set so that PySpark can look up the correct folders for the required files.

In the Windows search box (textbox next to Windows icon in the status bar) type `env` to add and edit environment variables.



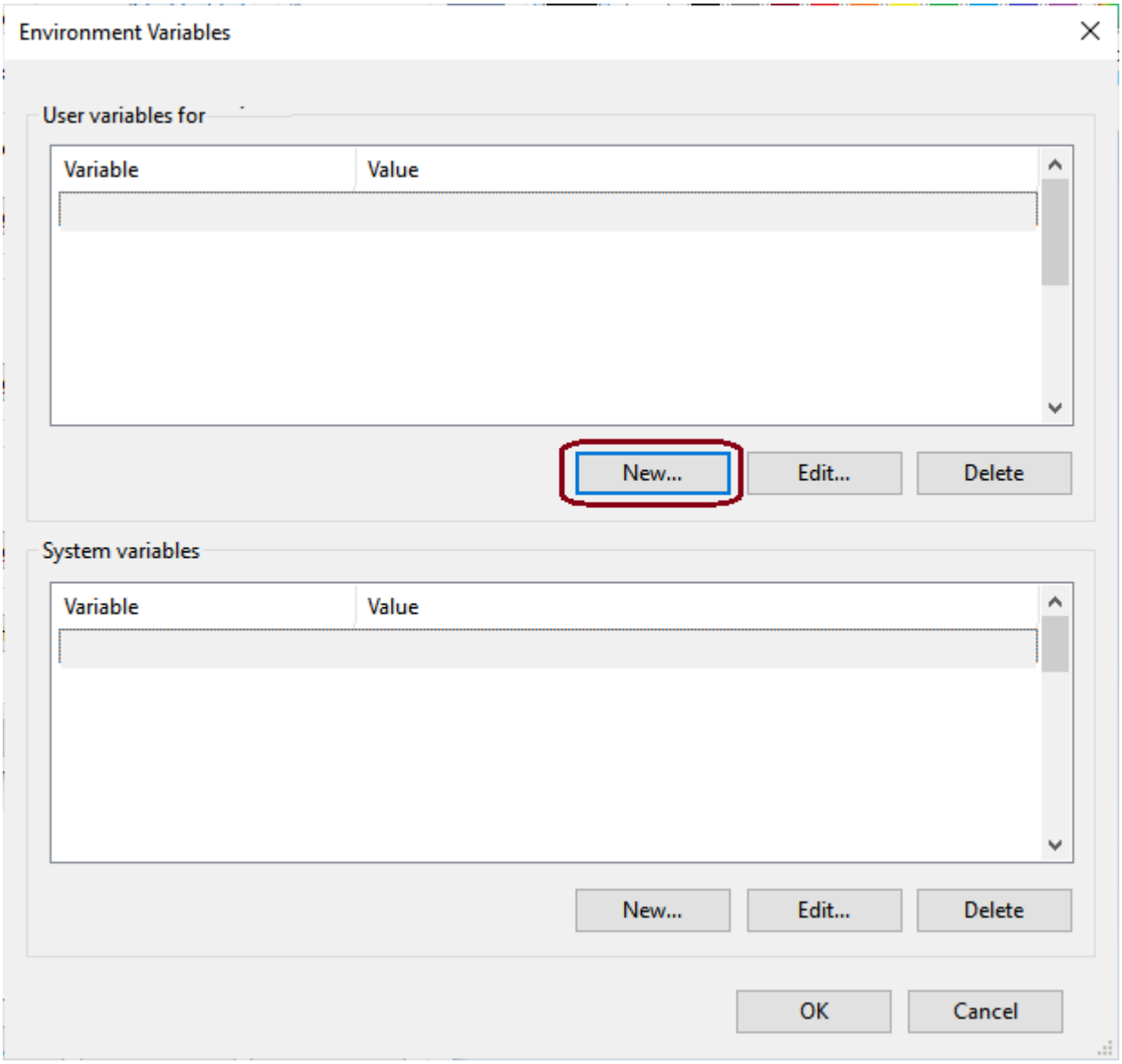
Click on the `Edit the system environment variables`

A window as shown below pops up. Click on `Environment Variables…` button.

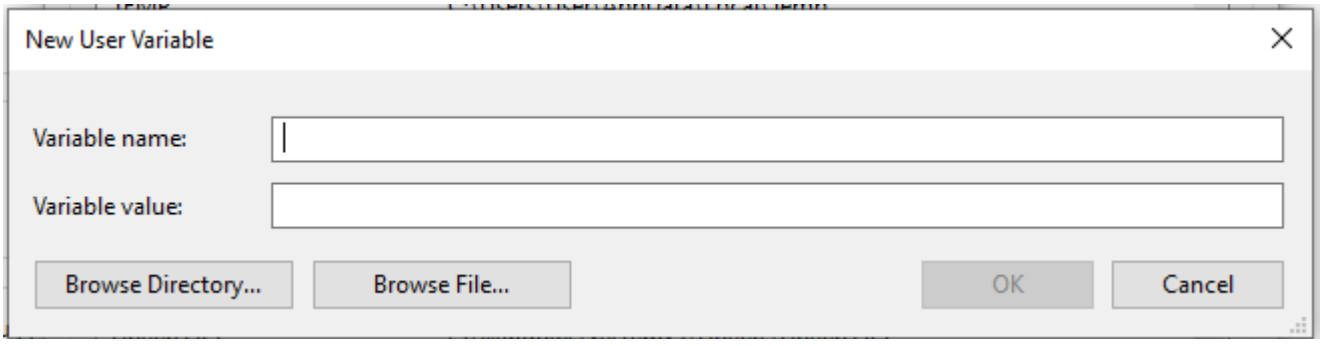


On the next screen (shown below) click on `New` button below `User variables for…` section.

On the next screen (shown below) click on `New` button below `User variables for…` section.



And add a new variable in the next screen as shown below.

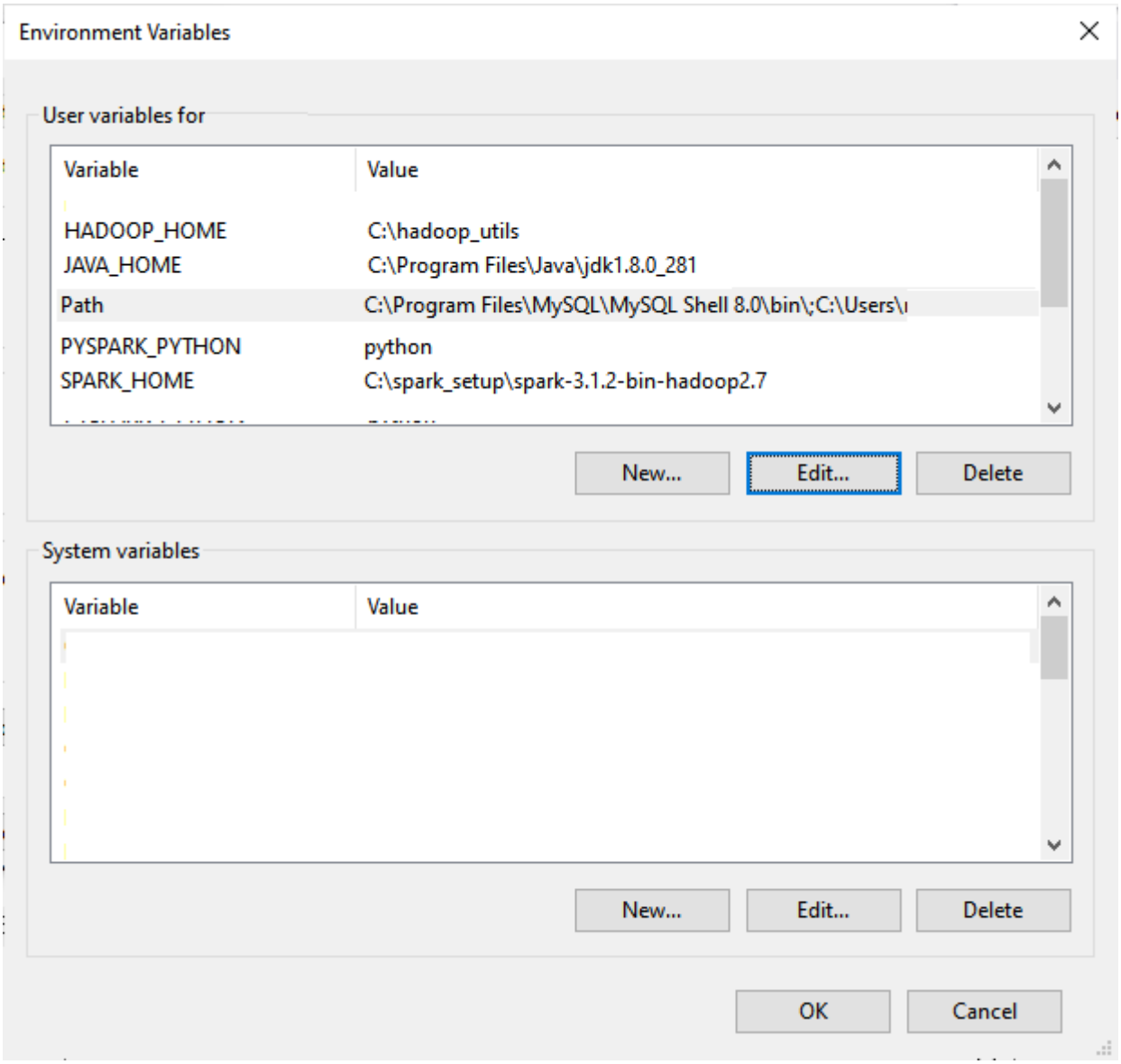


Add each of the new variables listed below.

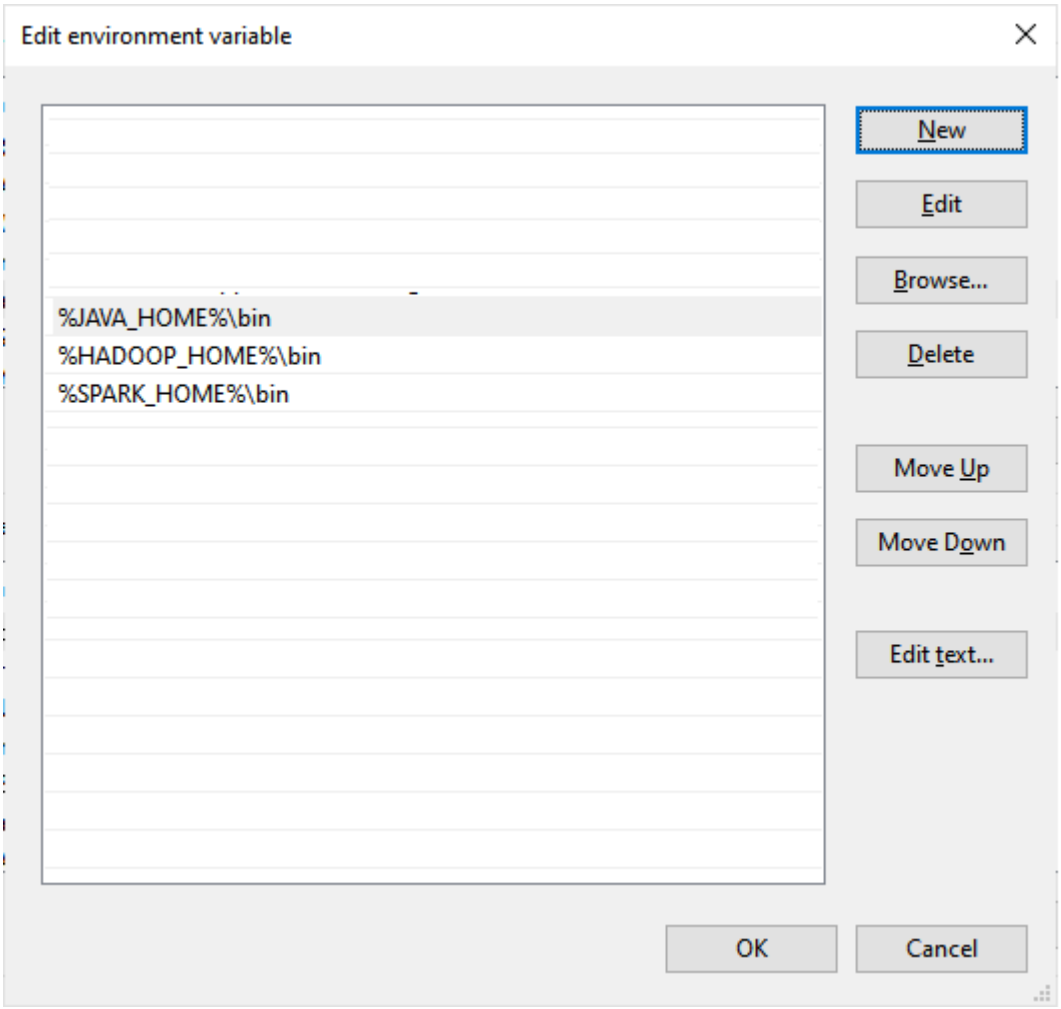
1. Variable name: `JAVA_HOME` (if not already present)
   Variable value: `<Name of the folder where Java is installed>`

(Usually it will be C:\Program Files\Java\jdk1.8.0_281 or C:\Program Files (x86)\ Java\jdk1.8.0_281)

2. Variable name: `SPARK_HOME`
   Variable value: `C:\spark_setup\spark-3.1.2-bin-hadoop2.7`
3. Variable name: `HADOOP_HOME`
   Variable value: `C:\hadoop_utils`
4. Variable name: `PYSPARK_PYTHON`
   Variable value: `python`
5. Set the `Path` variable.
   On `Environment Variables` window choose/highlight `Path` variable in `User variables for…` section and click on `Edit` button below this section.



In the next screen, click on the `New` button and add the variables as shown below.



1. `%JAVA_HOME%\bin` (if not present already)
2. `%HADOOP_HOME%\bin`
3. `%SPARK_HOME%\bin`

This completes the set up of Spark on the Windows.

You can run PySpark shell from the command prompt by giving the command `pyspark`.