Follow the steps below to install and configure Spark on Windows.

Installing Prerequisites:

1. Java

Spark requires Java 1.8.x or above
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 official Oracle website from the link below and install it.

https://www.oracle.com/in/java/technologies/javase/javase8u211-later-archive-downloads.html

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.

Note: If you are going to use Spark only with Scala then Python installation is not required and you can skip the following step.

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
```

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

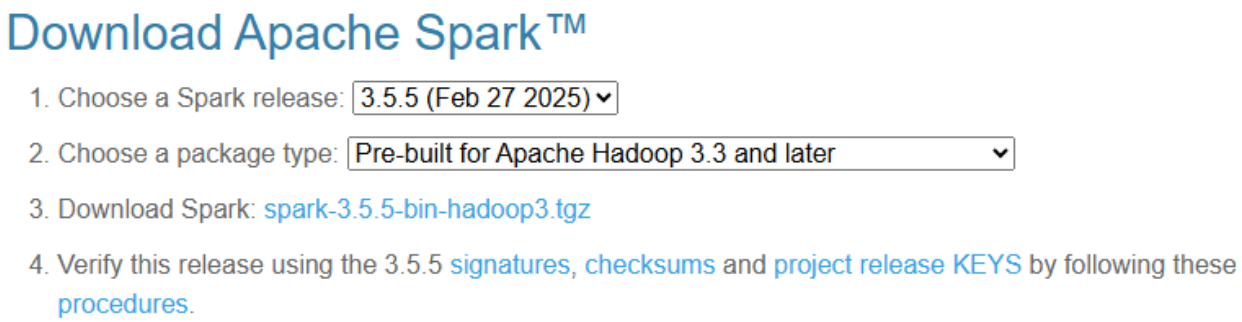
For an IDE, based on your requirement you can choose to install Jupyter or Anaconda or PyCharm community edition.
https://jupyter.org/install
https://www.jetbrains.com/pycharm/download/?section=windows **or**
https://www.jetbrains.com/pycharm/download/download-thanks.html?platform=windows&code=PCC
https://www.anaconda.com/products/individual

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

Installing Spark:

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

Choose Spark release: most recent release as now is `3.5.5`
Most importantly choose package type: `Pre-built for Apache Hadoop 3.3`



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

Extract the files into a sub-folder say `spark-3.5.5-bin-hadoop3` 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 Hadoop `winutils.exe`:

To run Apache Spark on windows, you need Hadoop utilities `winutils.exe` and `hadoop.dll`

Download them from:
`https://github.com/cdarlint/winutils/blob/master/hadoop-3.3.5/bin/winutils.exe`
`https://github.com/cdarlint/winutils/blob/master/hadoop-3.3.5/bin/hadoop.dll`

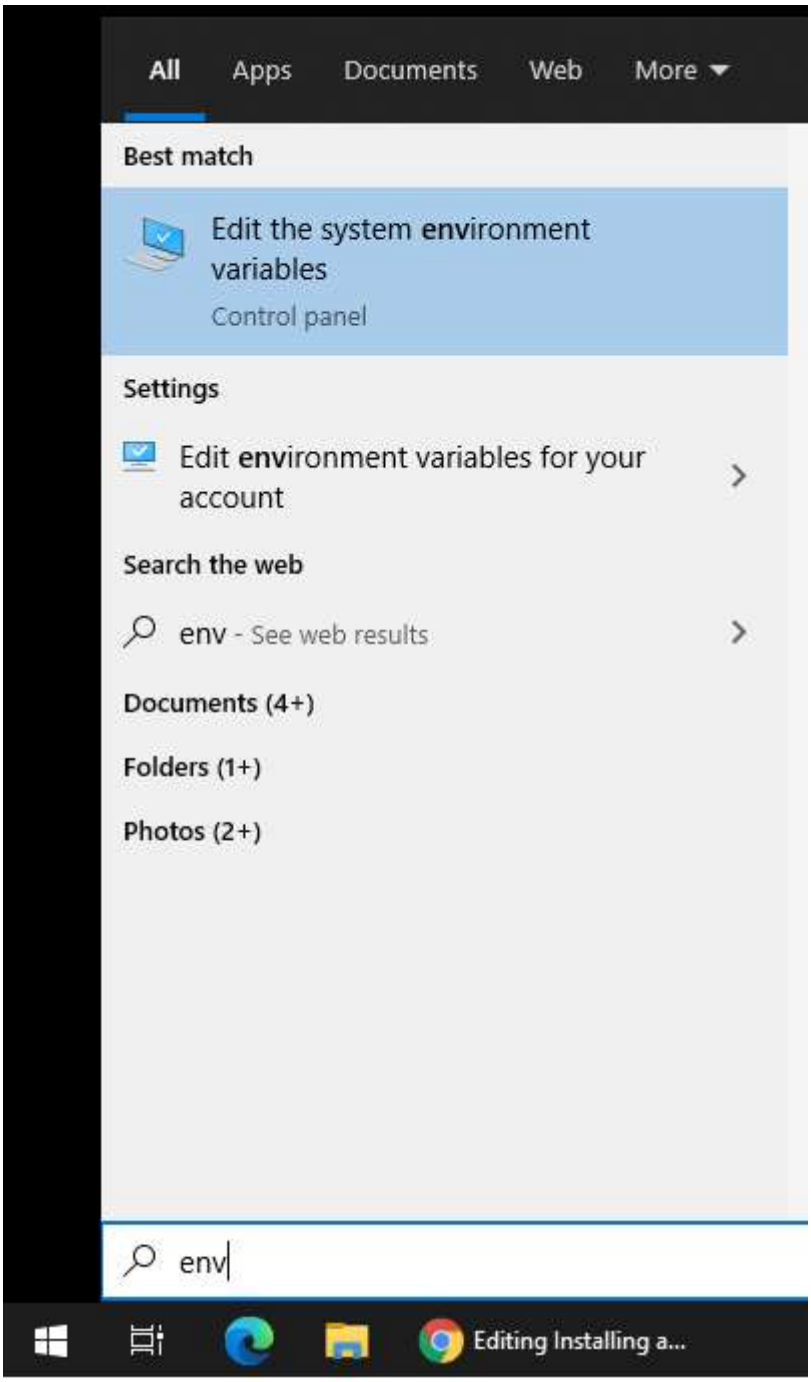
Make sure you download the `winutils.exe` file corresponding to the Spark and Hadoop version you are using.

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

Setting up Environment Variables:

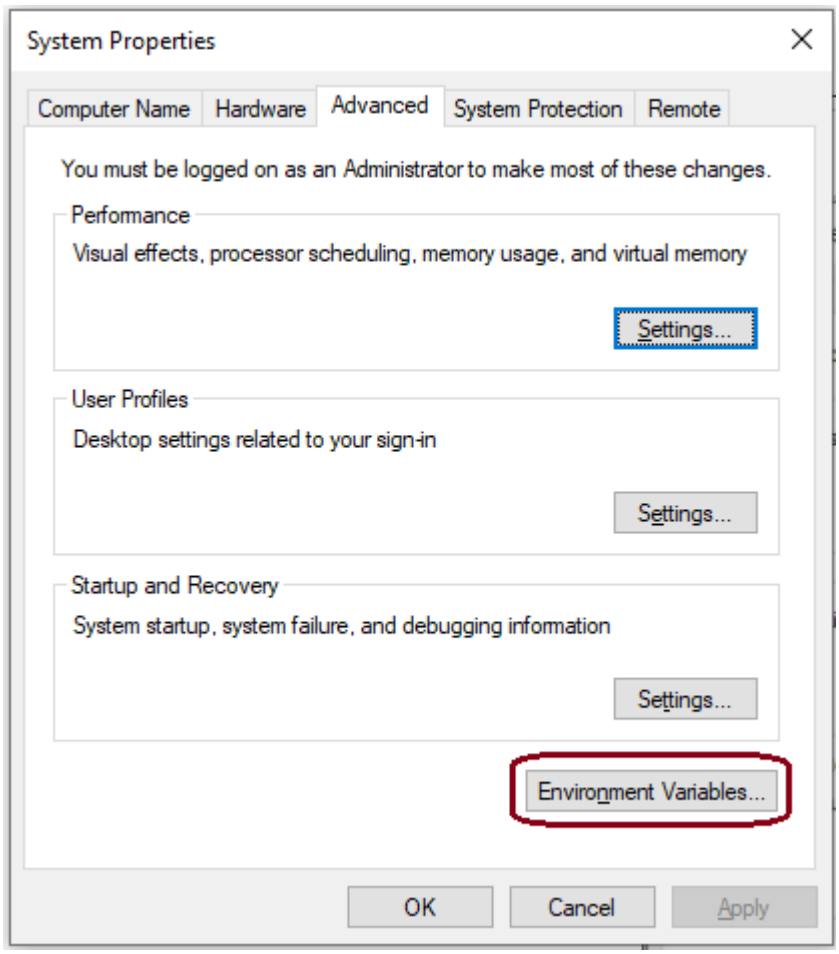
Environment variables need to be set so that Spark can look up the correct folders for the required files.

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



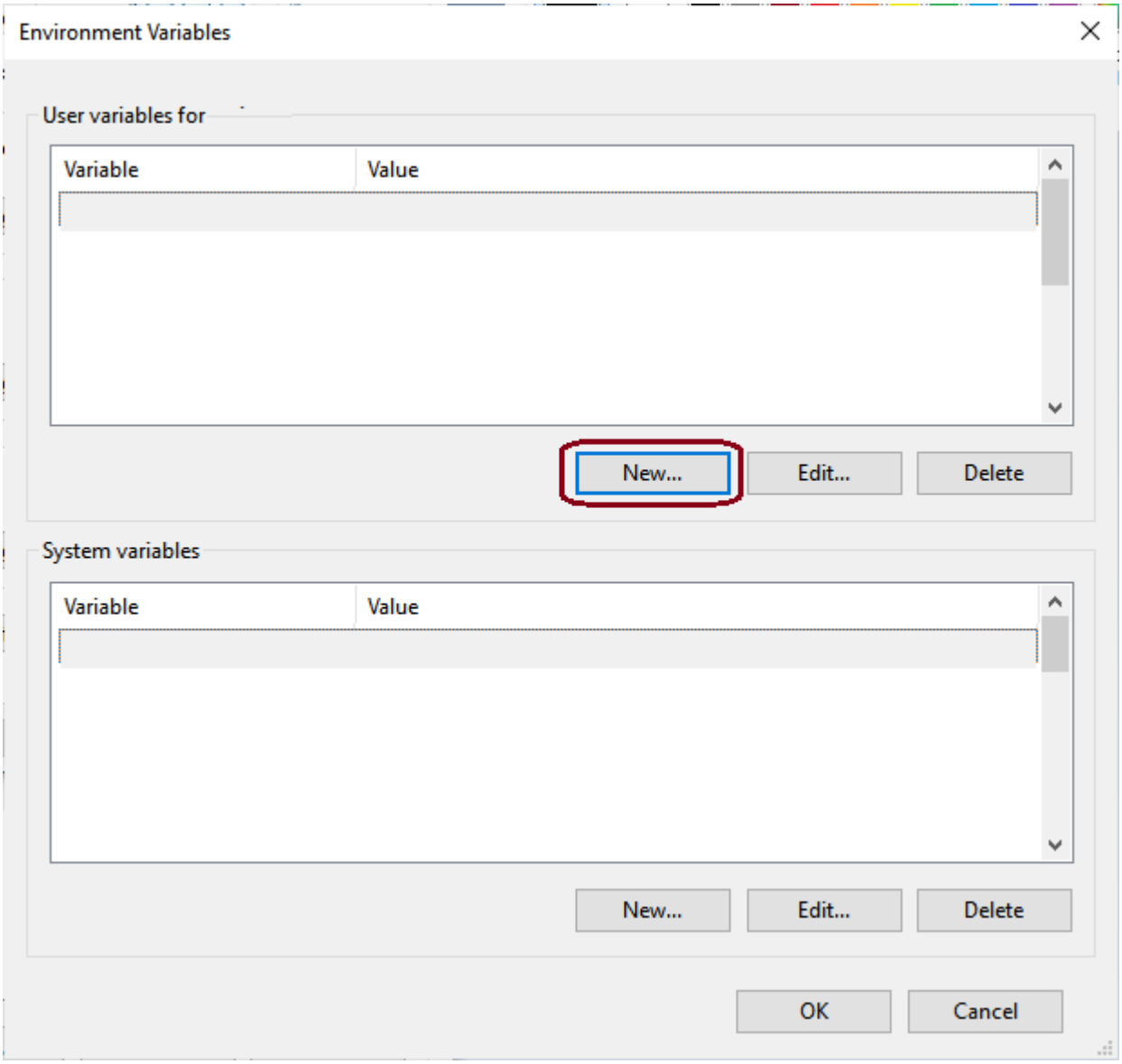
Click on the `Edit the system environment variables`

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

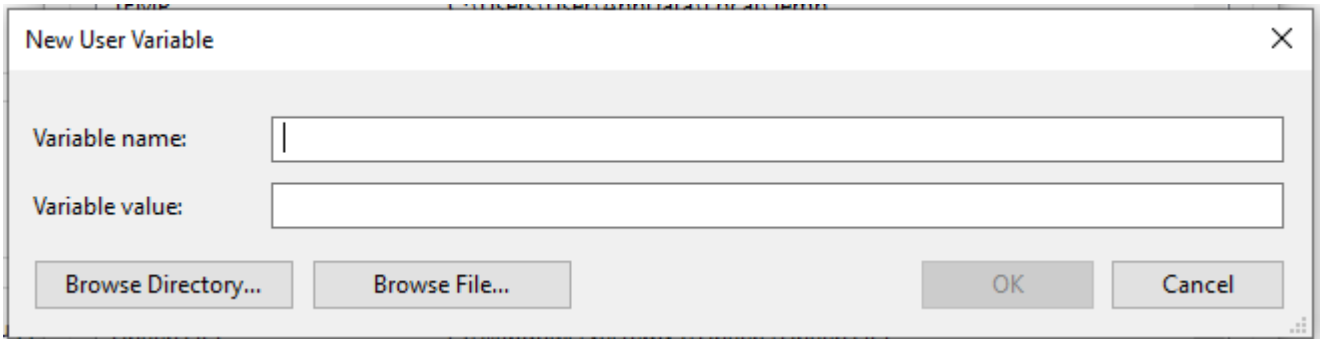


On the next screen (shown below) click on `New` button below `User variables for…` section.

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.

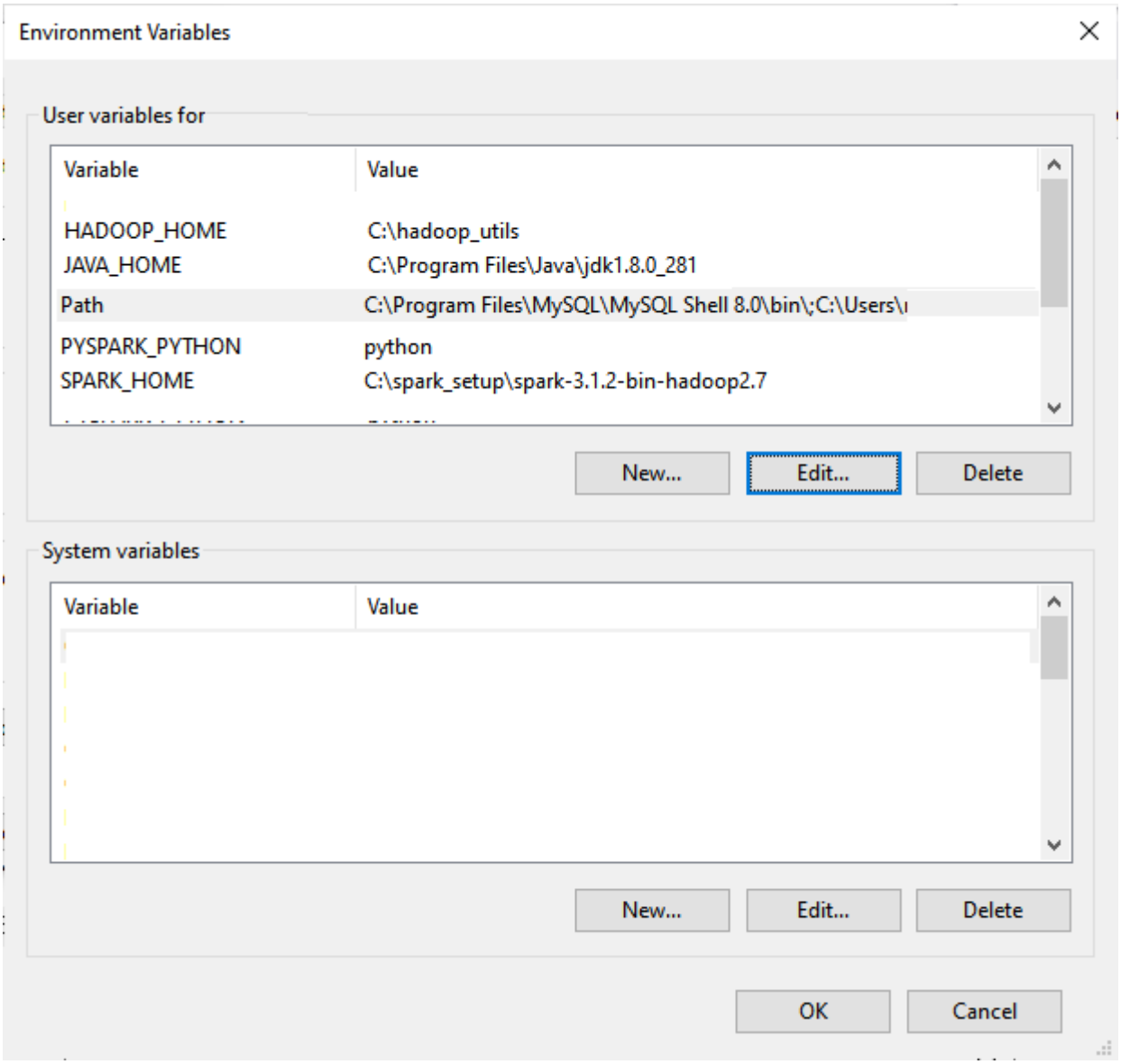


Add each of the new variables listed below.

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

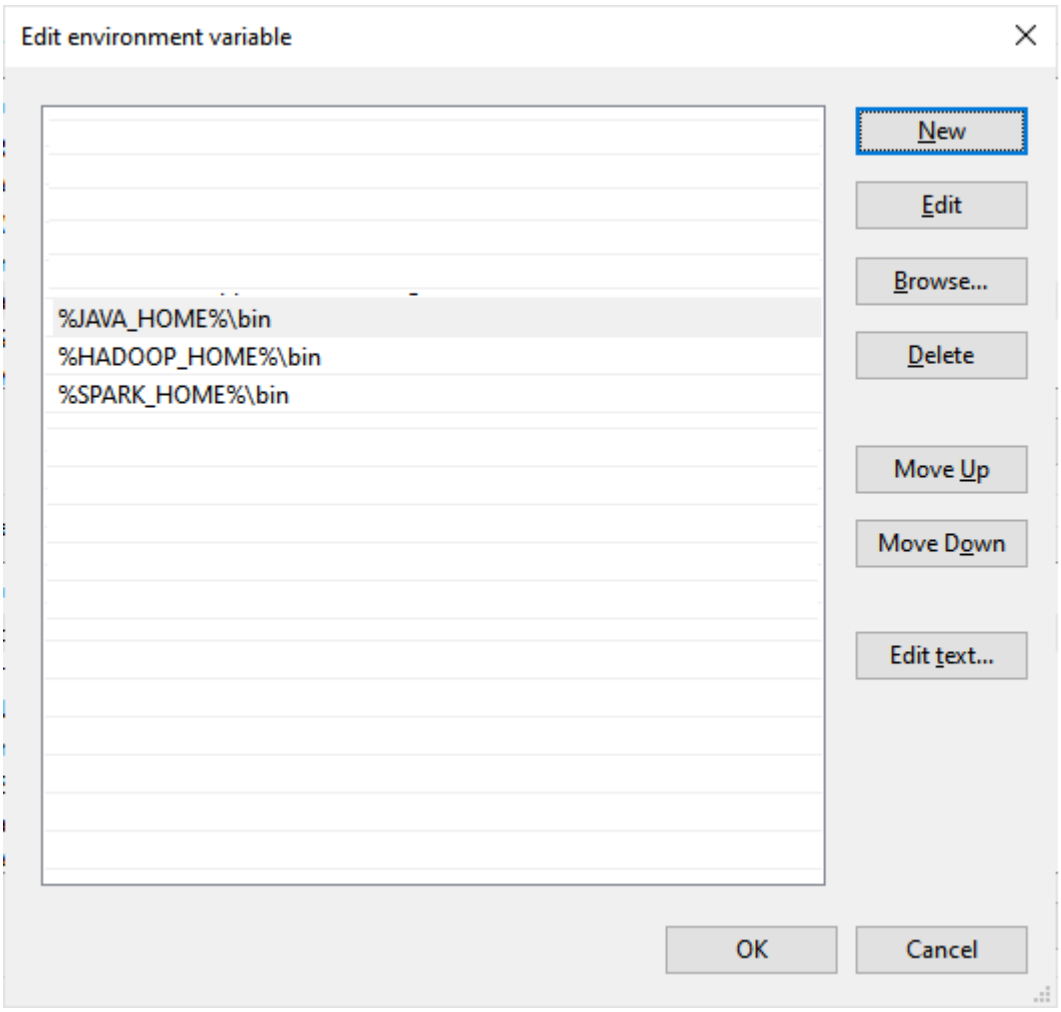
(Usually it will be C:\Program Files\Java\jdk1.8.0_281 or C:\Program Files (x86)\ Java\jdk1.8.0_281)

2. Variable name: `SPARK_HOME`
   Variable value: `C:\Spark\spark-3.5.5-bin-hadoop3`
3. Variable name: `HADOOP_HOME`
   Variable value: `C:\Hadoop`
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.



Note that the screen-shot above is given as an example; the folder names could be different; please use the names as given on your system.

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 installation of Spark on the Windows.

You can run Spark shell from the command prompt by giving the command `spark-shell` and verify.

For PySpark you can run PySpark shell from the command prompt by giving the command `pyspark` and verify.

Installing IntelliJ IDEA:

One of the preferred IDEs for Spark applications with Scala is IntelliJ IDEA of JetBrains.
You can download the free community edition from the link below.

`https://www.jetbrains.com/idea/download/?section=windows`

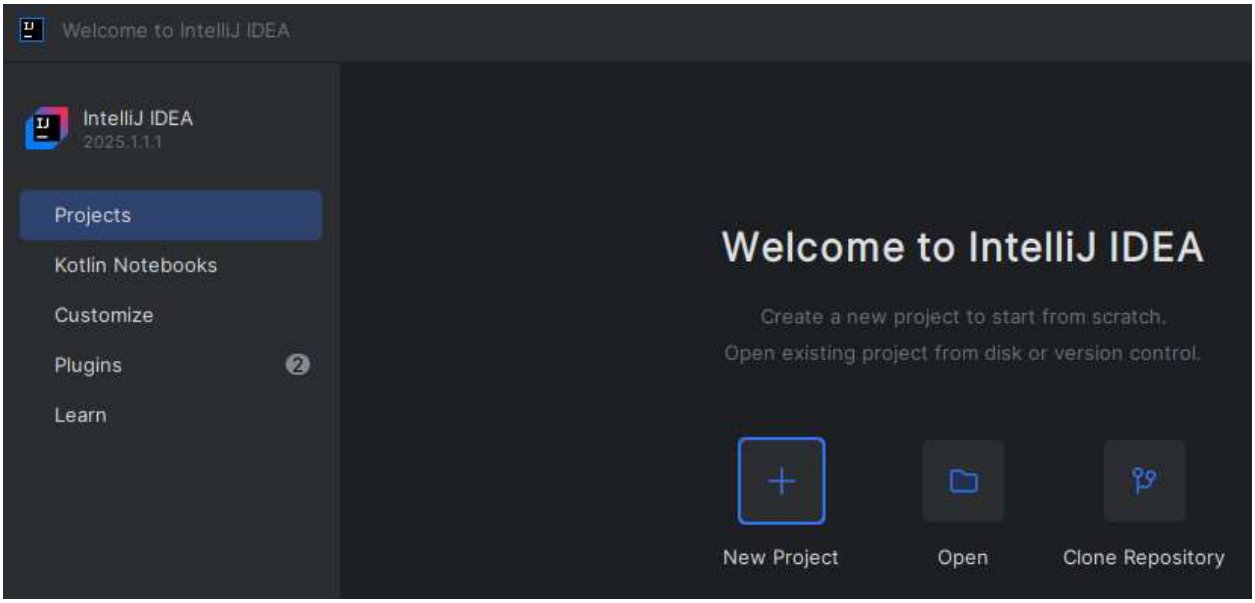
Use the link below to start the download directly.

`https://www.jetbrains.com/idea/download/download-thanks.html?platform=windows&code=IIC`

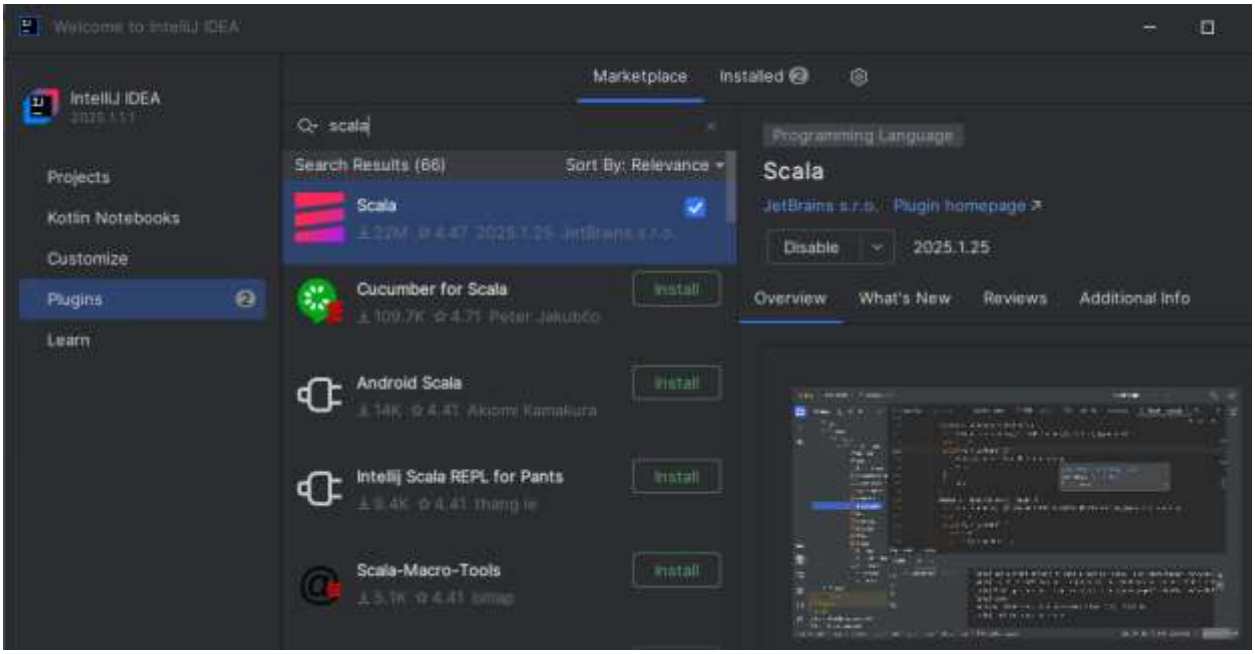
While installing, choose the options *create desktop shortcut* and *add to path/ environment variables*.

After installation is completed, run IntelliJ and install Scala Plug-in.

As shown in the screenshots below, you can select plugins from the main screen, then search for Scala under Market Place and select Install.



After installing Scala plugin a check box will be displayed in place of the *Install* button as in the screenshot below.



Installing Scala language is not mandatory. We can use Spark-shell which is an interactive REPL shell to learn Scala features or IntelliJ for writing code.