Apache Spark and PySpark Installation

Windows

Git Download

Git Download

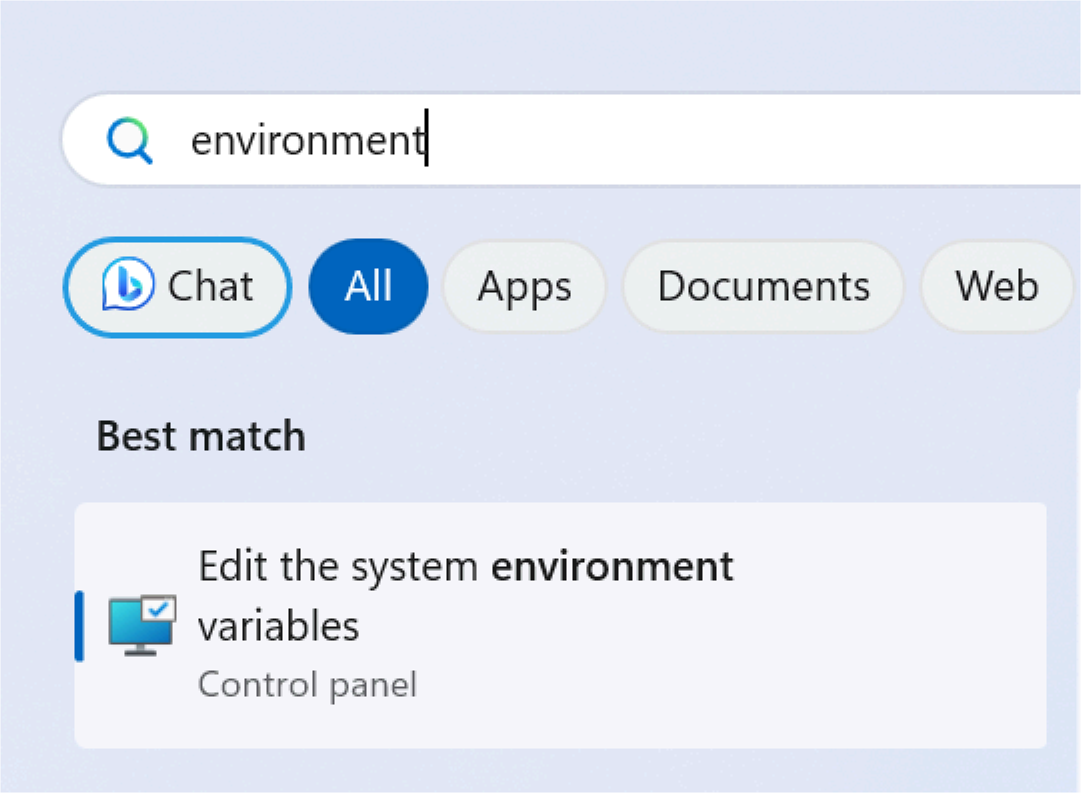
Java Development Environment Setup

Download JDK 11

Download JDK 11

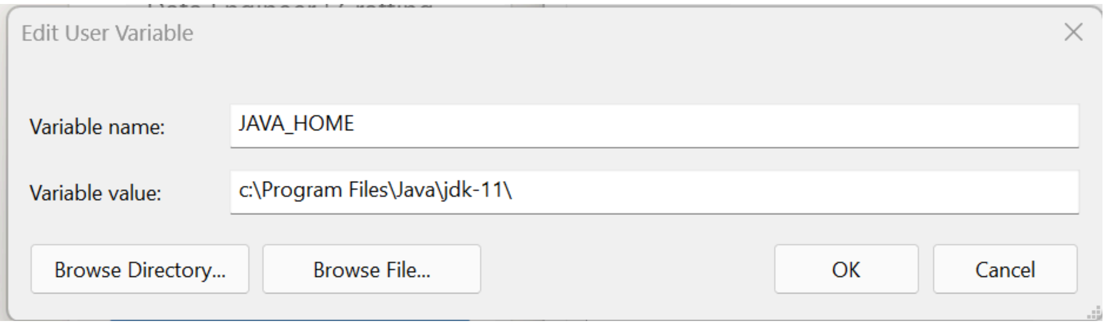
Configure Environment Variables

1. Press the Windows key and search for **Environment** anc click on ***Edit the system environment variables.***


environment

Click the Advanced tab, → Click the Environment Variables button

2. Under User variables click **New** and add the following

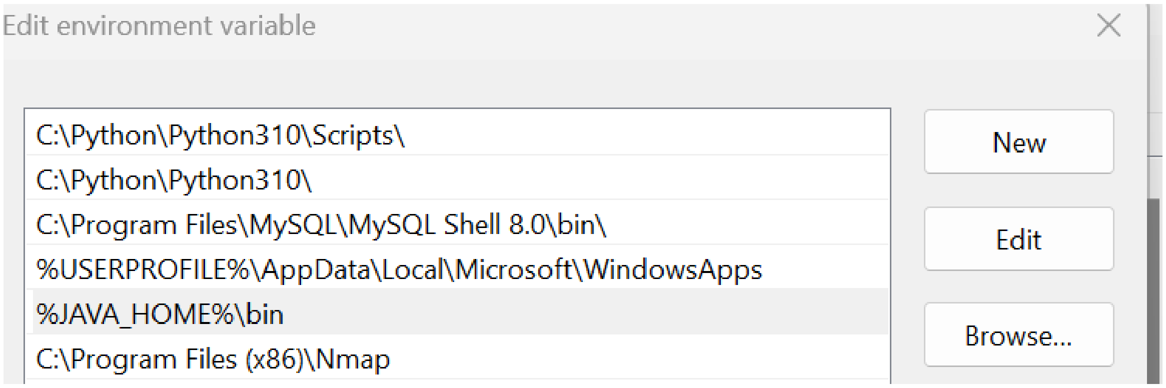

Edit User Variable

Variable name: JAVA_HOME

Variable value: c:\Program Files\Java\jdk-11\

3. Under User variables search for Path variable and click Edit

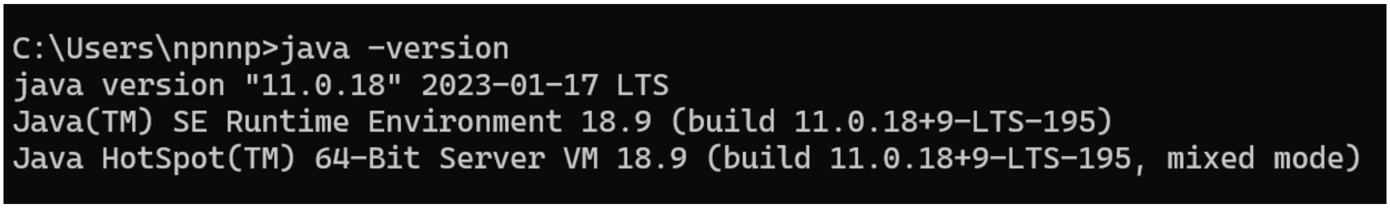
   Add the following



C:\Python\Python310\Scripts\
C:\Python\Python310\
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
%JAVA_HOME%\bin
C:\Program Files (x86)\Nmap

Click OK

4. Check on cmd, see below:



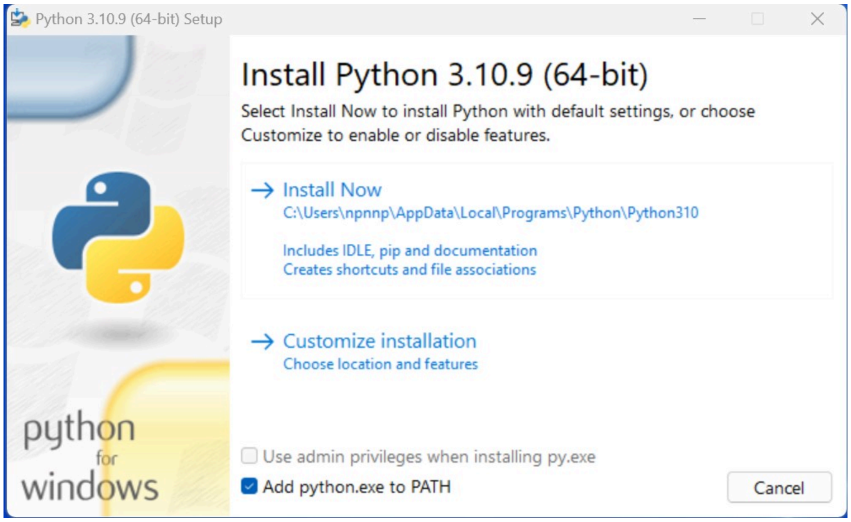
```
C:\Users\npnnp>java -version
java version "11.0.18" 2023-01-17 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.18+9-LTS-195)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.18+9-LTS-195, mixed mode)
```

Python Installation

Step 01 : Download Python from the link given below

Download python-3.10.9

Step 02 : Click python-3.10.9.exe which you have downloaded and click on Add Python 3.10.9 to PATH as shown in the screenshot and click on Customize installation then click Next



Python 3.10.9 (64-bit) Setup

Install Python 3.10.9 (64-bit)

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

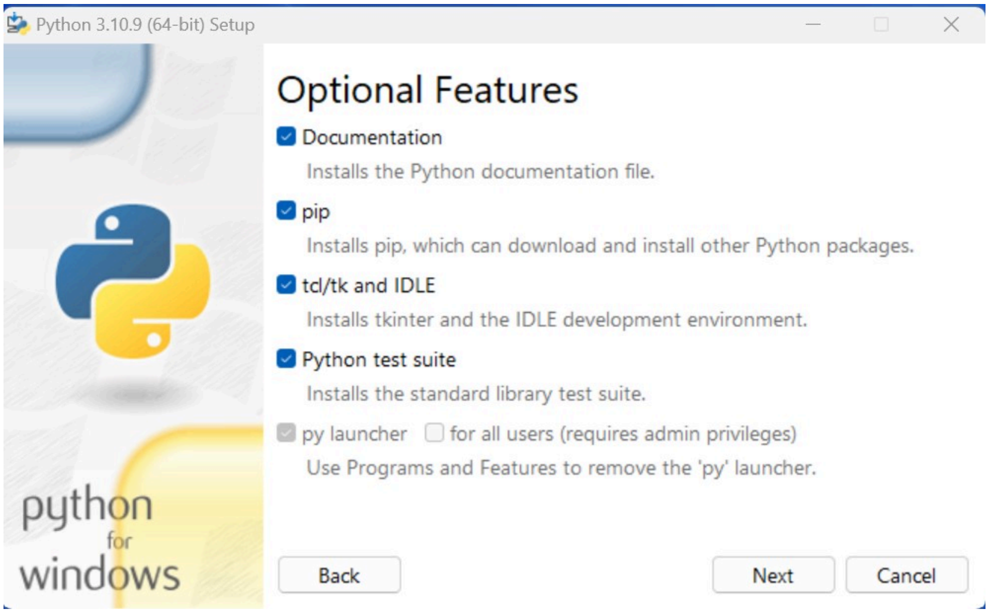
→ Install Now
C:\Users\npnnp\AppData\Local\Programs\Python\Python310

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

□ Use admin privileges when installing py.exe
☑ Add python.exe to PATH

Step 03 : Click Next



Python 3.10.9 (64-bit) Setup

Optional Features

☑ Documentation
Installs the Python documentation file.

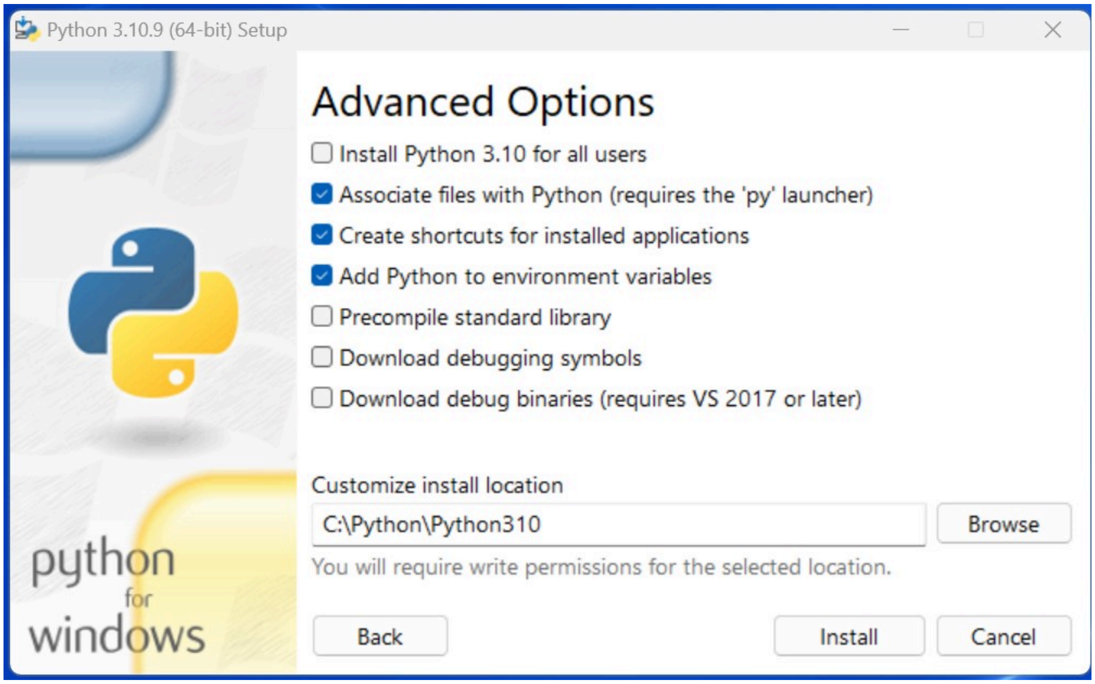
☑ pip
Installs pip, which can download and install other Python packages.

☑ tcl/tk and IDLE
Installs tkinter and the IDLE development environment.

☑ Python test suite
Installs the standard library test suite.

☑ py launcher □ for all users (requires admin privileges)
Use Programs and Features to remove the 'py' launcher.

Step 04 : Configure Customize install location as C:\Python\Python310 shown and the Ensure Add Python to environment variables is checked then click Next



Advanced Options

- ☐ Install Python 3.10 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 2017 or later)

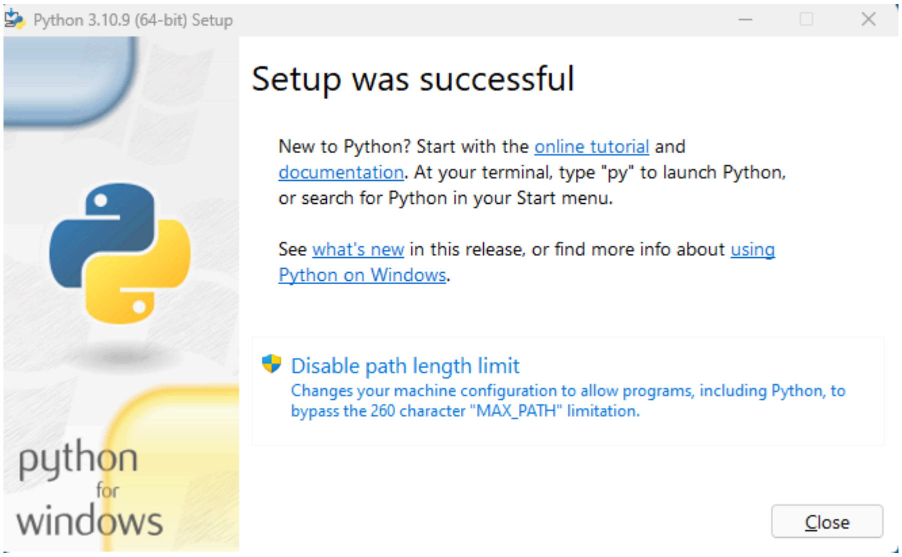
Customize install location

C:\Python\Python310

You will require write permissions for the selected location.

Step 05

Disable path length limit



Setup was successful

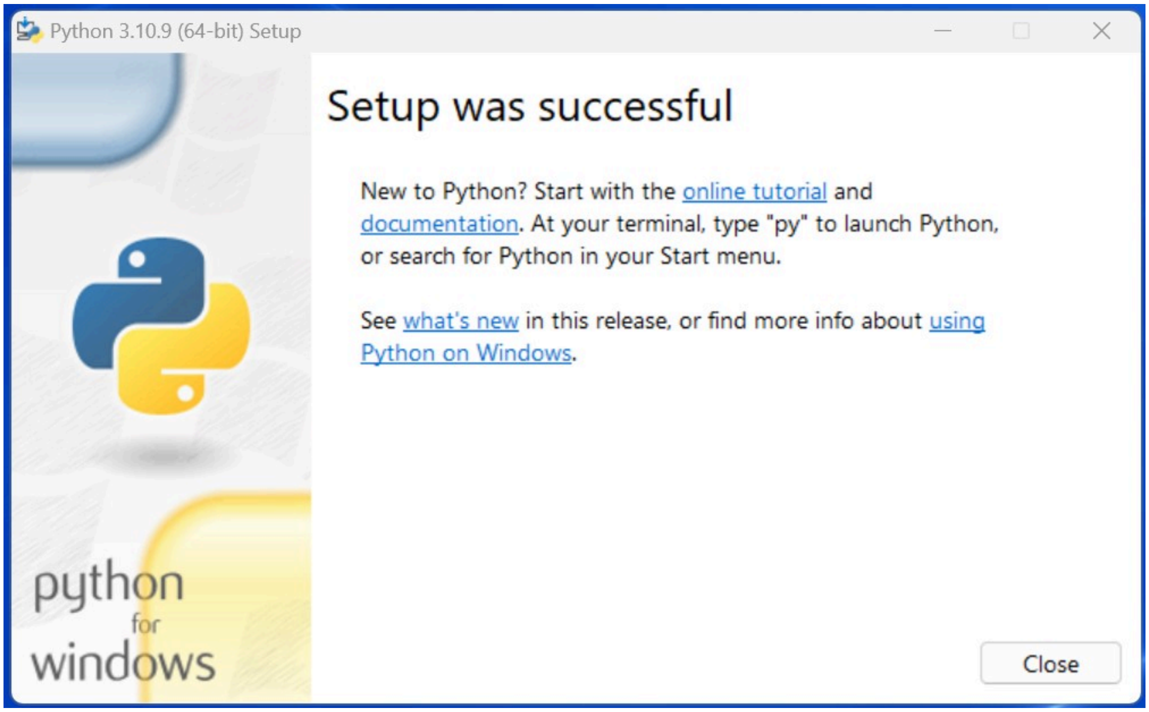
New to Python? Start with the online tutorial and documentation. At your terminal, type "py" to launch Python, or search for Python in your Start menu.

See what's new in this release, or find more info about using Python on Windows.

Disable path length limit

Changes your machine configuration to allow programs, including Python, to bypass the 260 character "MAX_PATH" limitation.

Step 06 : Click Close



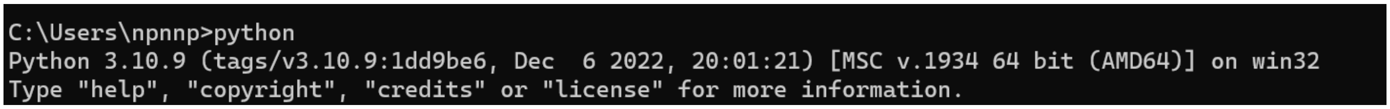
Setup was successful

New to Python? Start with the online tutorial and documentation. At your terminal, type "py" to launch Python, or search for Python in your Start menu.

See what's new in this release, or find more info about using Python on Windows.

Step 04 : Check for correct installation

Open Command Prompt



```
C:\Users\npnnp>python
Python 3.10.9 (tags/v3.10.9:1dd9be6, Dec  6 2022, 20:01:21) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
```

PyCharm Installation

Step 01 : Download **PyCharm Community Edition**

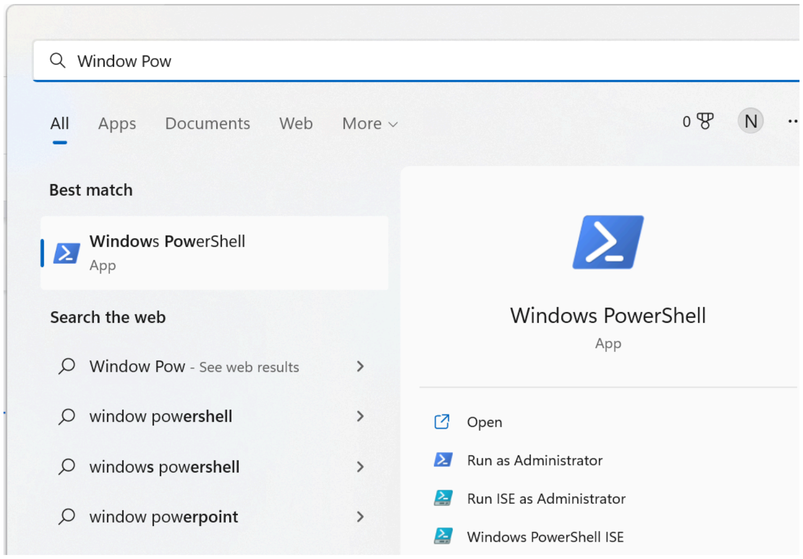
Navigate to the given link and scroll down to find PyCharm Community Edition

⤓ Download PyCharm Community Edition

PyCharm Community Edition

The IDE for Pure Python Development

Edit Windows Powershell (SKIP)



Open Powershell in **Administrator Mode** and execute the below command and give 'Y' wherever it prompts

```
$> Set-ExecutionPolicy -ExecutionPolicy RemoteSigned -Scope LocalMachine
```

Scala Installation

[Scala 2.12.8]

Clone Git Workspace

Open the command prompt and clone the repo into c:\

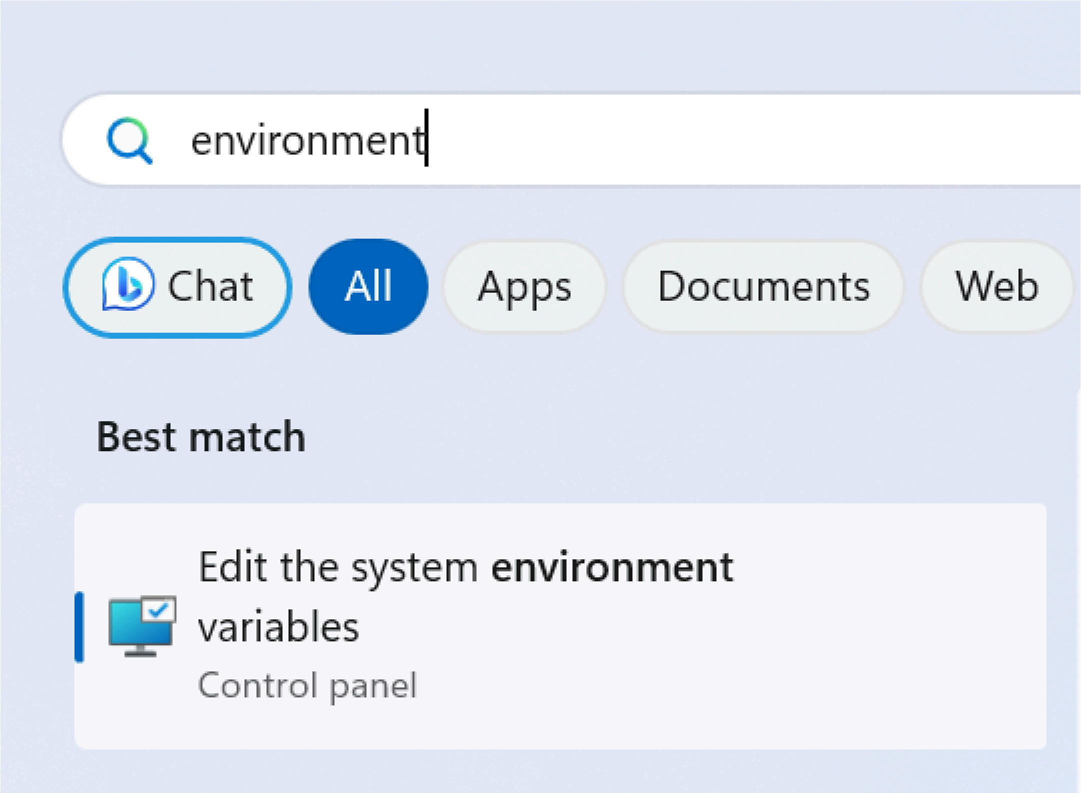
```
cd c:\
git clone https://github.com/naveenpn-trainer/data-engineering-env-setup
```

Apache Spark Installation

1. Download apache-spark.zip to C:\data-engineering-env-setup\
2. Right Click on **apache-spark.zip** , Select 7-Zip and select Extract Here

Configure Environment Variables

1. Press the Windows key and search for **Environment** anc click on ***Edit the system environment variables.***

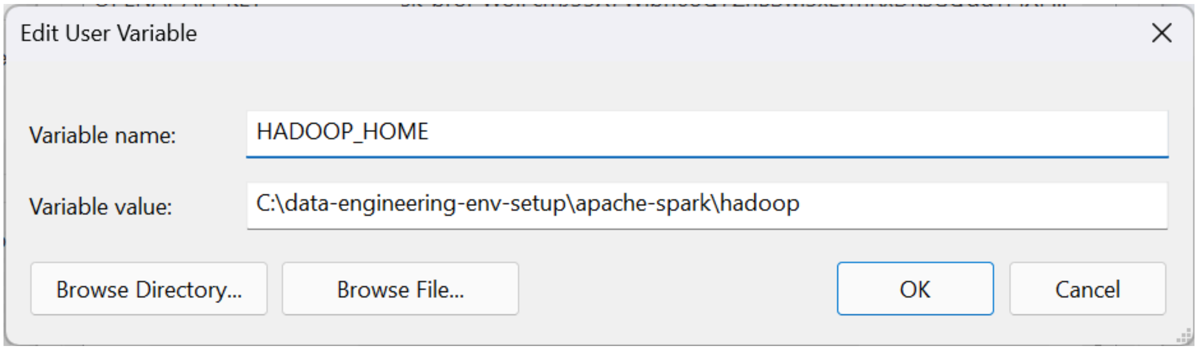


Click the Advanced tab, → Click the Environment Variables button

2. Add the following new User variables Click New and add below mentioned details

Configuring Hadoop

- Variable Name : HADOOP_HOME
- Variable value: C:\data-engineering-env-setup\apache-spark\hadoop

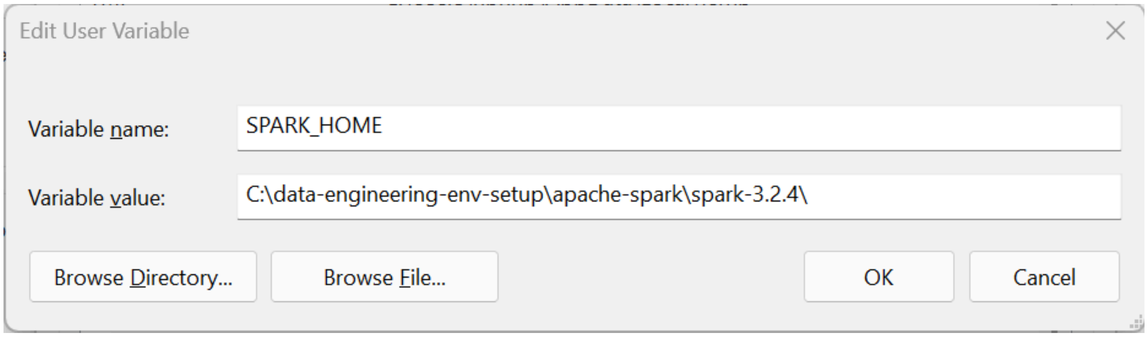


Variable name: HADOOP_HOME

Variable value: C:\data-engineering-env-setup\apache-spark\hadoop

Configuring Spark

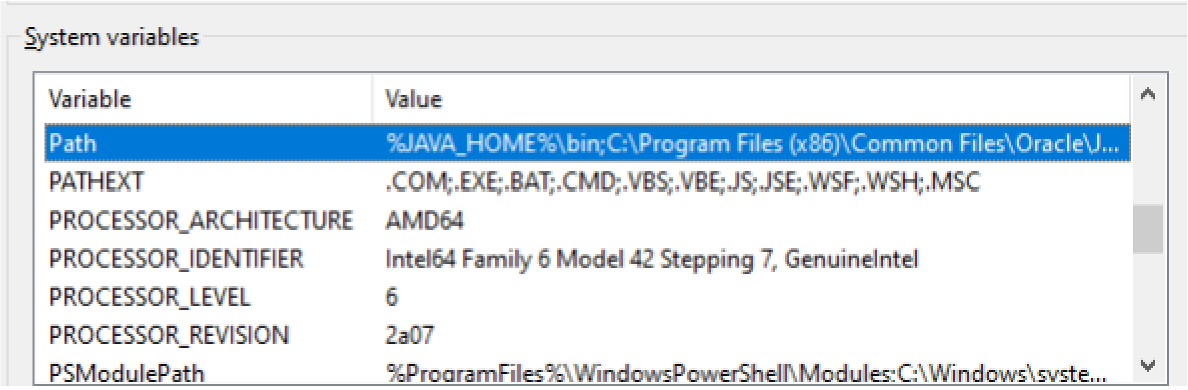
- Variable Name : SPARK_HOME
- Variable Value : C:\big-data-engineering-masters-program\frameworks\apache-spark\spark-3.2.4-bin-hadoop2.7



Variable name: SPARK_HOME

Variable value: C:\data-engineering-env-setup\apache-spark\spark-3.2.4\

3. Now click on Path Variable under User variables

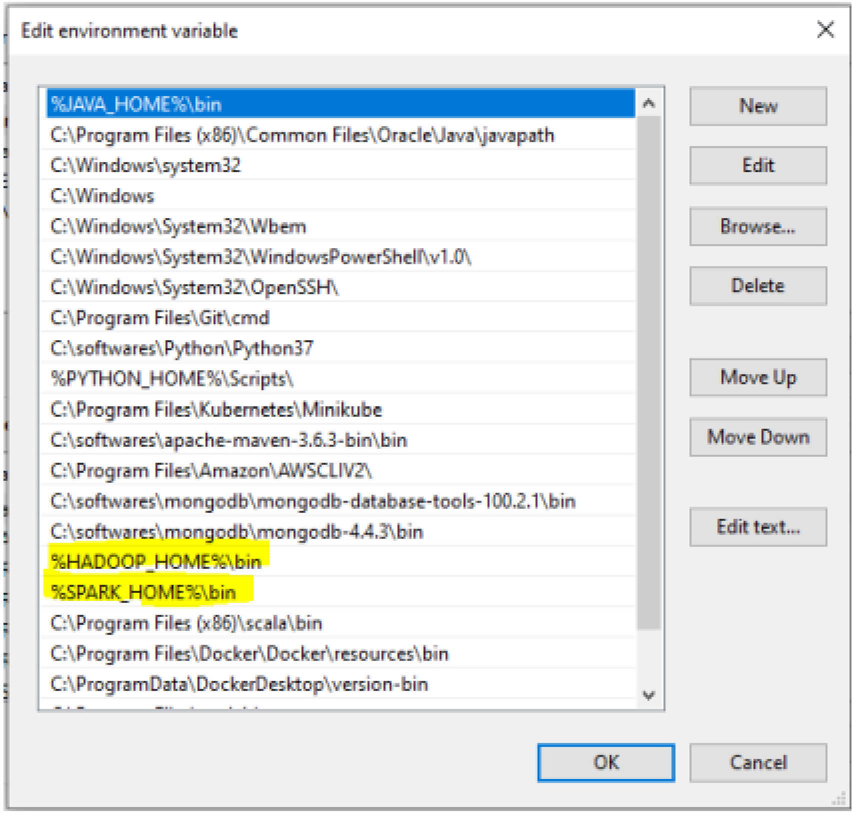


Variable	Value
Path	%JAVA_HOME%\bin;C:\Program Files (x86)\Common Files\Oracle\J...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 42 Stepping 7, GenuineIntel
PROCESSOR_LEVEL	6
PROCESSOR_REVISION	2a07
PSModulePath	%ProgramFiles%\WindowsPowerShell\Modules;C:\Windows\syste...

4. Add the following paths to your PATH variable:

%HADOOP_HOME%\bin

%SPARK_HOME%\bin



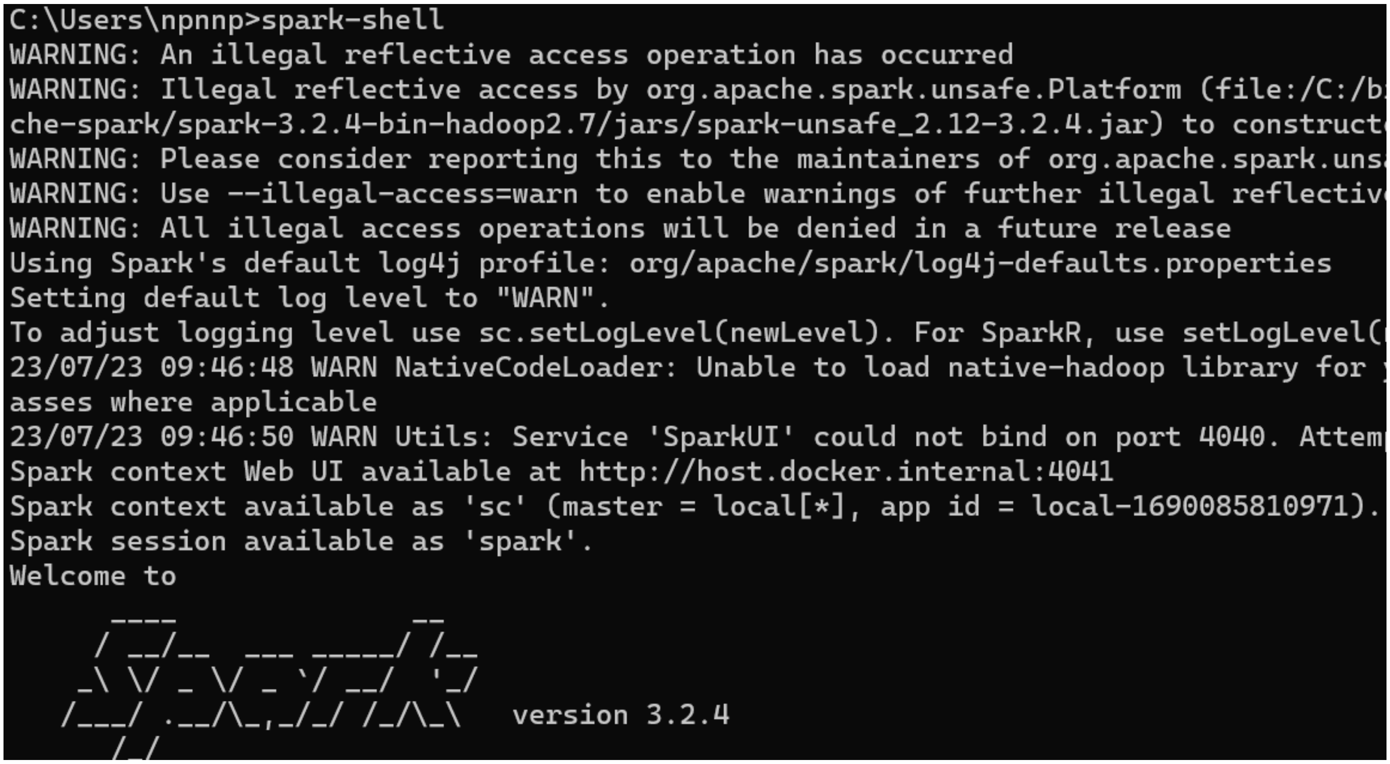
Edit environment variable

```
%JAVA_HOME%\bin
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Windows\System32\OpenSSH\
C:\Program Files\Git\cmd
C:\softwares\Python\Python37
%PYTHON_HOME%\Scripts\
C:\Program Files\Kubernetes\Minikube
C:\softwares\apache-maven-3.6.3-bin\bin
C:\Program Files\Amazon\AWSCLIV2\
C:\softwares\mongodb\mongodb-database-tools-100.2.1\bin
C:\softwares\mongodb\mongodb-4.4.3\bin
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
C:\Program Files (x86)\scala\bin
C:\Program Files\Docker\Docker\resources\bin
C:\ProgramData\DockerDesktop\version-bin
```

Close the environment variable screen and the control panels.

Test for correct Installation

1. Close all the command prompt if opened
2. Open up windows command prompt
3. And type spark-shell



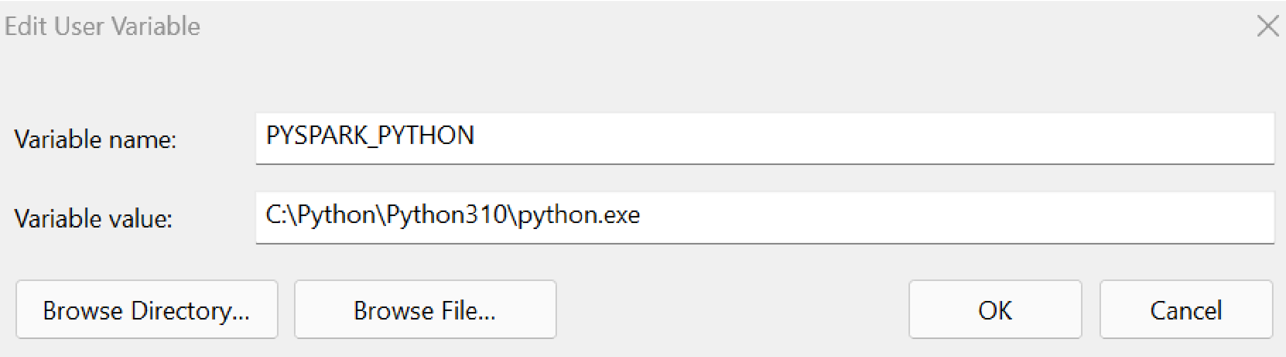
```
C:\Users\npnnp>spark-shell
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/b
che-spark/spark-3.2.4-bin-hadoop2.7/jars/spark-unsafe_2.12-3.2.4.jar) to construct
WARNING: Please consider reporting this to the maintainers of org.apache.spark.uns
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflectiv
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(
23/07/23 09:46:48 WARN NativeCodeLoader: Unable to load native-hadoop library for
asses where applicable
23/07/23 09:46:50 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attem
Spark context Web UI available at http://host.docker.internal:4041
Spark context available as 'sc' (master = local[*], app id = local-1690085810971).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.4
      /_/
```

PySpark Installation

1. Add the following new User variables Click New and add below mentioned details

   Configuring PySpark

   - Variable Name : PYSPARK_PYTHON
   - Variable Value : C:\Python\Python310\python.exe



Edit User Variable

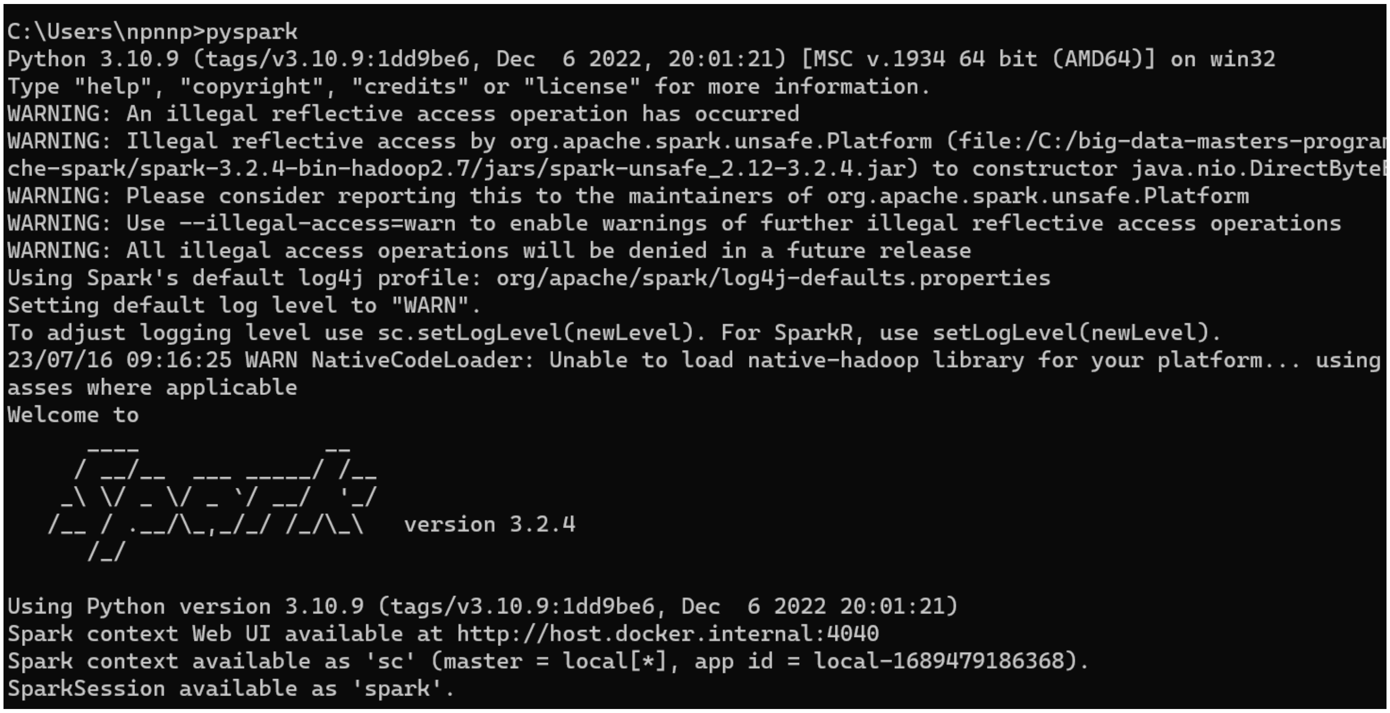
Variable name: PYSPARK_PYTHON

Variable value: C:\Python\Python310\python.exe

2. Open Commands prompt

```
$> python.exe -m pip install --upgrade pip
$> pip install pyspark==3.2.4
```

Test for correct Installation

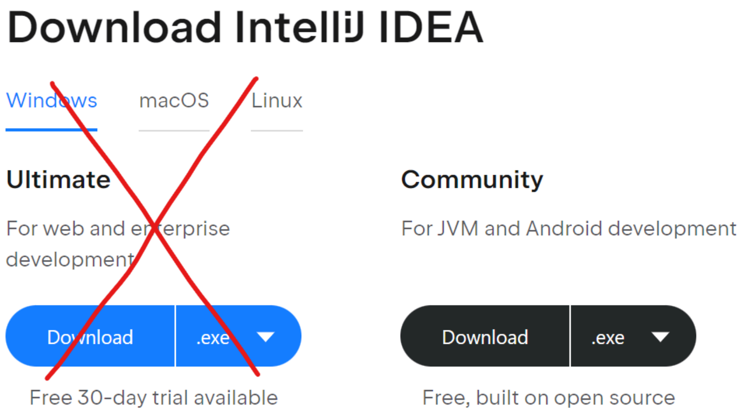


```
C:\Users\npnnp>pyspark
Python 3.10.9 (tags/v3.10.9:1dd9be6, Dec  6 2022, 20:01:21) [MSC v.1934 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/big-data-masters-progra
che-spark/spark-3.2.4-bin-hadoop2.7/jars/spark-unsafe_2.12-3.2.4.jar) to constructor java.nio.DirectByte
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/07/16 09:16:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
asses where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.4
      /_/

Using Python version 3.10.9 (tags/v3.10.9:1dd9be6, Dec  6 2022 20:01:21)
Spark context Web UI available at http://host.docker.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1689479186368).
SparkSession available as 'spark'.
```

Intellij IDEA IDE Installation (Skip for Python Developers)

Step 01 : Download Intellij IDEA Community Edition **(Skip if already downloaded)**


Download IntelliJ IDEA
Windows
macOS
Linux
Ultimate
For web and enterprise development
Download
.exe
Free 30-day trial available
Community
For JVM and Android development
Download
.exe
Free, built on open source

Step 02 : Install Plugins
Launch IntelliJ

1. Click on **Customize** → Change **Color Theme** to Intellij Light

2. Click on **Plugins** → Click on Marketplace →
    a. Search for **Scala** and install
    b. Search for **Batch Scripts Support** and install

Docker Big Data Cluster

1. Download Docker Desktop : https://www.docker.com/products/docker-desktop/

2. Launch Docker Desktop

3. Execute C:\data-engineering-env-setup\multi-node-big-data-cluster\start-multi-node-cluster-windows.bat in command line