

# BEAD Workshop Series

# Installation and Configuration Instructions



ISS-BA-BEAD/CPM/Workshop/ Installation 01 - PySpark on Windows based Hadoop Local Cluster  
PySpark on Windows based Hadoop Local Cluster  
© 2015-2023 NUS. All rights reserved.

## Table of Contents

Introduction .....	3
Simple Installation Video .....	3
Pre-Requisite.....	3
Download Sequence .....	3
Install 7 zip, JDK, Winutils (Hadoop), Scala, Spark, and Python .....	5
7Zip Installation.....	5
JDK Installation.....	5
Python Installation.....	6
Scala Installation .....	6
Hadoop (winutils.exe) Setup.....	7
Spark Installation .....	7
Path Setup.....	8
Environment Testing.....	10
Testing the pre-requisites .....	10
Testing Spark Installation.....	10
Install PyCharm .....	11
Create a PySpark Project using PyCharm.....	11
Connect to PySpark libraries.....	13
Test the PySpark codes .....	16
Concluding Remarks.....	17

## Introduction

This tutorial is to help participants install PySpark on a Windows 10 machine. The sequence of installation would involve multiple tools such as PySpark, PyCharm and Hadoop. For this firstly we need the pre-requisite installation including 7Zip, JDK (Java Development Kit), Python (we use version 3.8.X) and Hadoop (via winutils version ). Then we proceed to install Spark, PyCharm and test the PySpark coded. This set of instructions would guide you through the entire setup and configuration processes on Windows Platform.

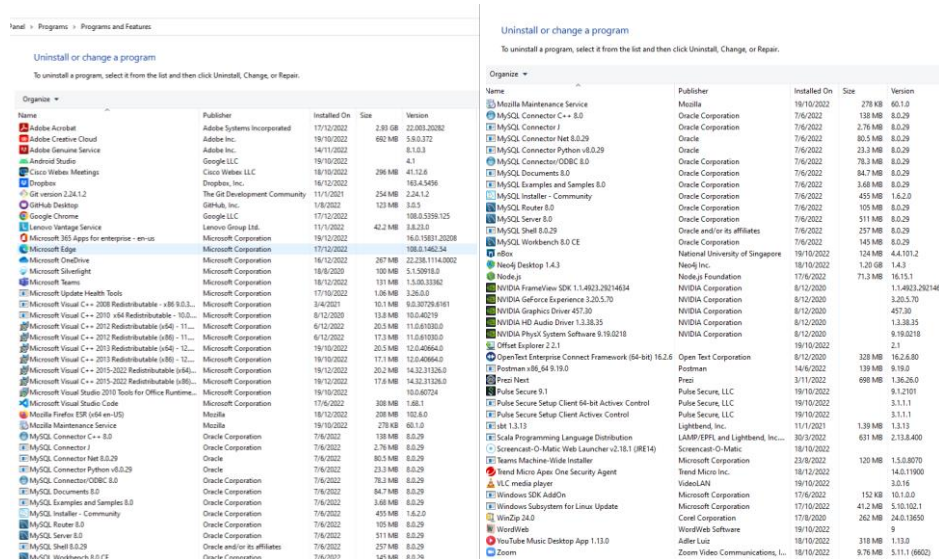
*Note: If you are a mac user you may find this link helpful to install similar tools in your environment.*

## Simple Installation Video

In case the participant wants to recreate the setup in their personal devices, they can use the below video log to guide them to install the needed technical environment as shown in the following links:

## Pre-Requisite

Since a clean installation can save a lot of efforts and version mismatch issues, we highly recommend that you do not have multiple python/java installation before starting this installation workshop. For instance, screen below ensures the same.



## Download Sequence

The following open source / community software stack were installed in a Windows 10 to help participants fast track their learning using PySpark on Hadoop cluster:

1. 7 zip tool downloads [link](#). Download a copy of this tool (.exe 64-bit Windows x64 version) into a special designated download folder (in this case D:\BEAD\Downloads).
2. Signup for Java download via Oracle. Download the windows installer of Java Development Kit 11 from [link](#) (Version jdk-11.0.16.1\_windows-x64\_bin.exe)

Community Edition via the created account to same download folder. Please note that you will be asked to accept Oracle's networking license agreement in the process.

- For Hadoop 3.3.1 winutils, please download from [link](#) to same download folder (winutils.exe you may have to use 'Save link as' from the right click context menu by mousing over the executable file).
- Spark Framework can be downloaded from the [link](#) to same download folder. Make sure you choose the 3.3.1 version labelled oct 22 targeting Hadoop 3.3 in the dropdown list and then click download on the step 3 spark-3.3.1-bin-hadoop3.tgz file link as shown below:

## Download Apache Spark™

1. Choose a Spark release: **3.3.1 (Oct 25 2022)** ▼

2. Choose a package type: **Pre-built for Apache Hadoop 3.3 and later** ▼

3. Download Spark: [spark-3.3.1-bin-hadoop3.tgz](#)

- As mentioned in the download page we need Scala distribution 2.13 for using this framework. It can be downloaded from [link](#) into the same download folder (file windows installer file from scala-2.13.1.msi). Scroll down to screen below

### Other resources

You can find the installer download links for other operating systems, as well as documentation and source code archives for Scala 2.13.1 below.

Archive	System	Size
<a href="#">scala-2.13.1.tgz</a>	Mac OS X, Unix, Cygwin	18.77M
<a href="#">scala-2.13.1.msi</a>	Windows (msi installer)	115.13M
<a href="#">scala-2.13.1.zip</a>	Windows	18.81M
<a href="#">scala-2.13.1.deb</a>	Debian	582.81M
<a href="#">scala-2.13.1.rpm</a>	RPM package	115.52M
<a href="#">scala-docs-2.13.1.bxz</a>	API docs	48.58M
<a href="#">scala-docs-2.13.1.zip</a>	API docs	99.67M
<a href="#">scala-sources-2.13.1.tar.gz</a>	Sources	

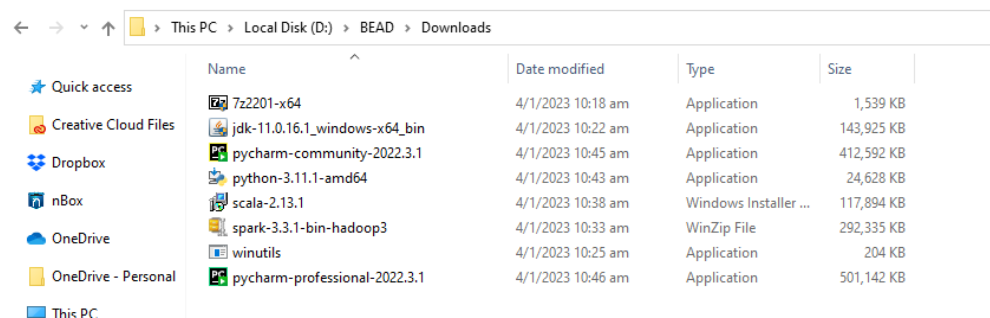
- Python Language Compiler and Interpreter (3.7+ is mandated by Hadoop, we will use the latest 3.11.1 windows installer 64 bit version from [link](#) downloaded to same folder). Screen shown below:

### Files

Version	Operating System	Description	MD5 Sum	File Size	GPG	Sigstore
<a href="#">Gzipped source tarball</a>	Source release		5c986b2865979b393aa50a31c65b64e8	26394378	SIG	CRT SIG
<a href="#">XZ compressed source tarball</a>	Source release		4efe92ad728875c77d3b9b2e8d3bc44a	19856648	SIG	CRT SIG
<a href="#">macOS 64-bit universal2 installer</a>	macOS	for macOS 10.9 and later	7c4d83ac21cf1e0470aa133ef6a1fff6	42665618	SIG	CRT SIG
<a href="#">Windows embeddable package (32-bit)</a>	Windows		cc960a3a6d5d1529117c463ac00aae43	9557137	SIG	CRT SIG
<a href="#">Windows embeddable package (64-bit)</a>	Windows		f16900451e15abe1ba3ea657f3c7fe9e	10538985	SIG	CRT SIG
<a href="#">Windows embeddable package (ARM64)</a>	Windows		405185d5ef1f436f8dbc370a868a2a85	9763968	SIG	CRT SIG
<a href="#">Windows installer (32-bit)</a>	Windows		a592f5db4f45ddc3a46c0ae465d3bee0	24054000	SIG	CRT SIG
<a href="#">Windows installer (64-bit)</a>	Windows	Recommended	3a02deed117ff4dbc1188d201ad164a	25218984	SIG	CRT SIG
<a href="#">Windows installer (ARM64)</a>	Windows	Experimental	3a98e0f9754199d99a7a97a6dadb0d91	24355528	SIG	CRT SIG

- Download PyCharm IDE Community Edition ([Latest](#)) to same download folder or you can sign up using nus matriculated email to obtain a one-year copy of Professional Edition from the same link.

When all downloads are complete your download folder will look like screen shot shown below:

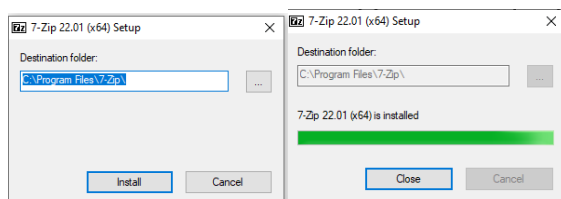


## Install 7 zip, JDK, Winutils (Hadoop), Scala, Spark, and Python

In this section, participants will install the pre-requisite tools for setting up spark and pyspark via PyCharm later. We will also set up the Spark framework in this section.

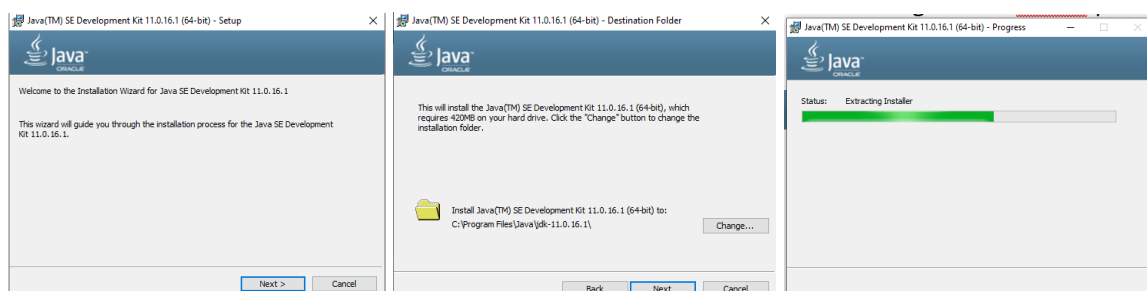
### 7Zip Installation

7-Zip is free software with open source. 7-Zip is popular for dealing with spark related zip and tar files. 7-Zip supports a wide range of formats: 7z, XZ, BZIP2, GZIP, TAR, ZIP and WIM. Installations is simple as shown in screenshots below:



### JDK Installation

The JDK is a development environment for building applications using the Java programming language. The JDK includes tools useful for developing and testing programs written in the Java programming language and running on the Java™ platform. Installation (accepting defaults) is simple as shown in screenshots below:



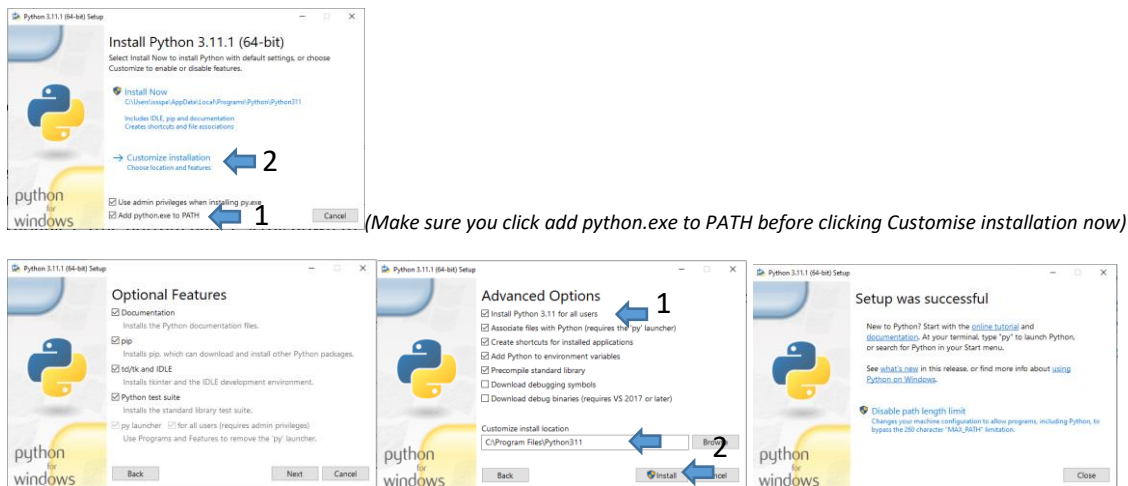


Once installation is complete click Close.

## Python Installation

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically typed, and garbage collected. It supports multiple programming paradigms, including structured, object-oriented, and functional programming.

Installation (accepting defaults) is simple as shown in screenshots below:



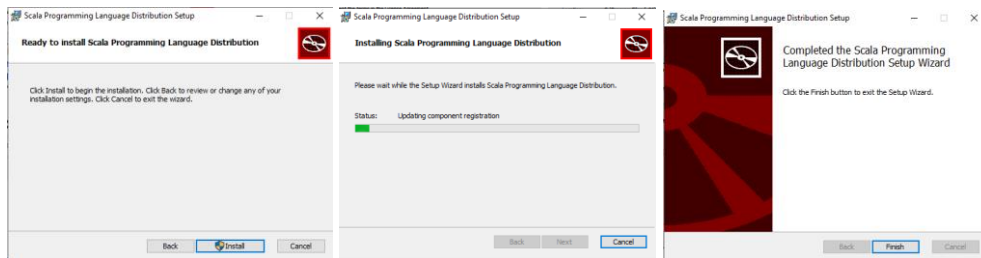
Once installation is complete click Close.

## Scala Installation

Scala is a strong statically typed general-purpose programming language that supports both object-oriented programming and functional programming. Designed to be concise, many of Scala's design decisions are aimed to address criticisms of Java.

Installation (accepting defaults) is simple as shown in screenshots below:

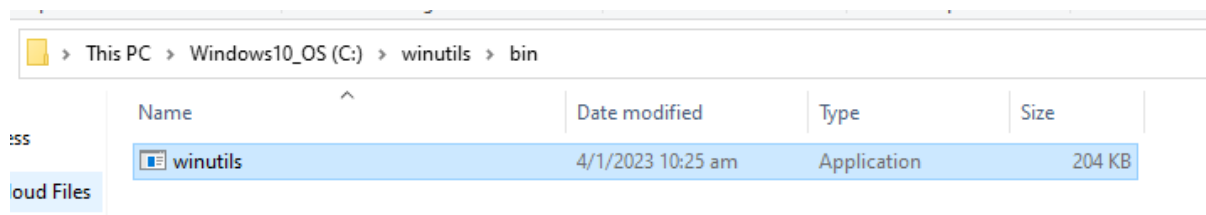




Once installation is complete click Finish.

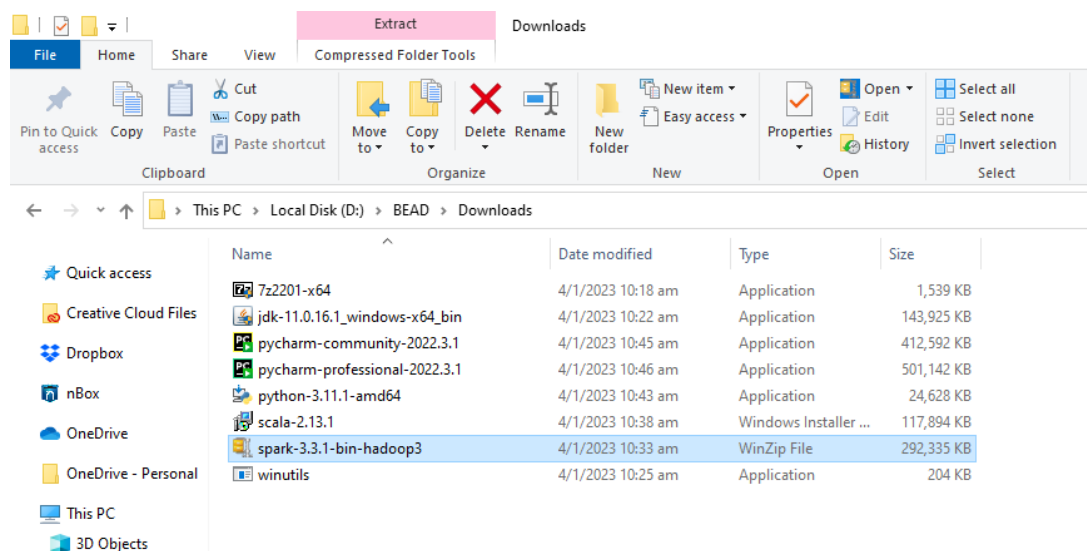
## Hadoop (winutils.exe) Setup

Create a folder called C:\winutils\bin. Copy the winutils.exe over to this folder. Screen below:



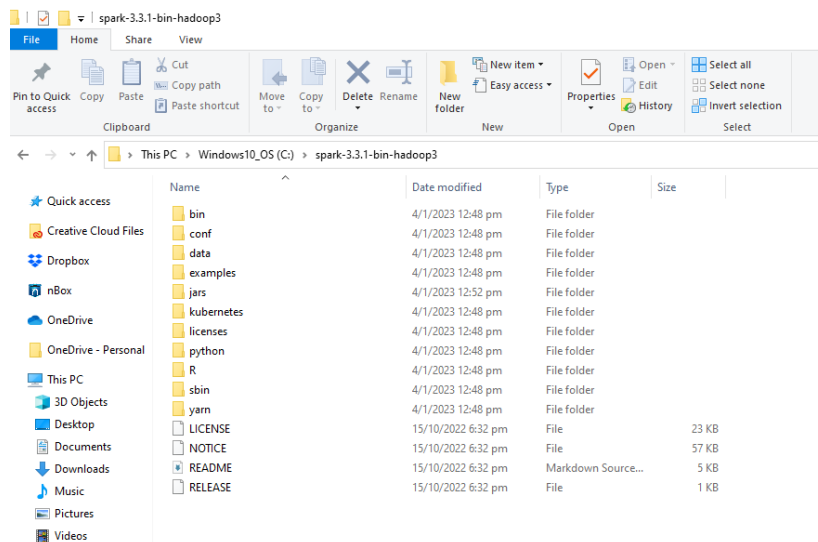
## Spark Installation

Spark files are downloaded and extracted from the previously downloaded tgz compressed file (**PLEASE USE ONLY 7ZIP**)



Use 7Zip>Extract here. Note you will have to deal with double compression. Use 7zip to unzip twice and move the unzipped final folder to the C drive – it looks as below:





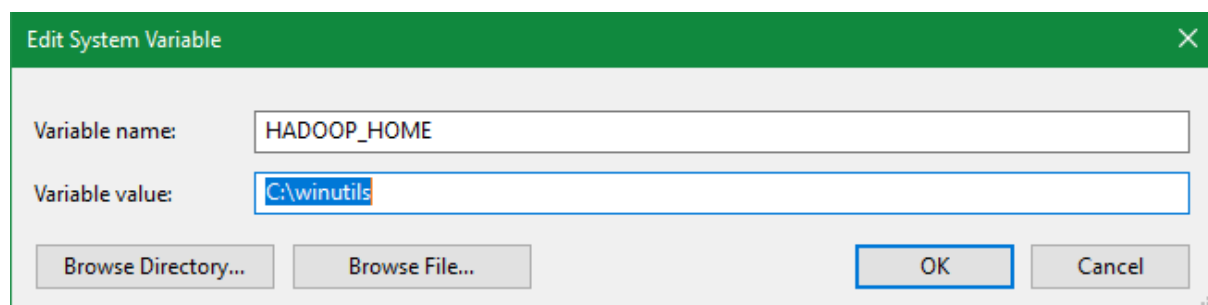
*Note: If you have issues with one or two jar file owing to file size or name, you can move them to respective folders manually.*

## Path Setup

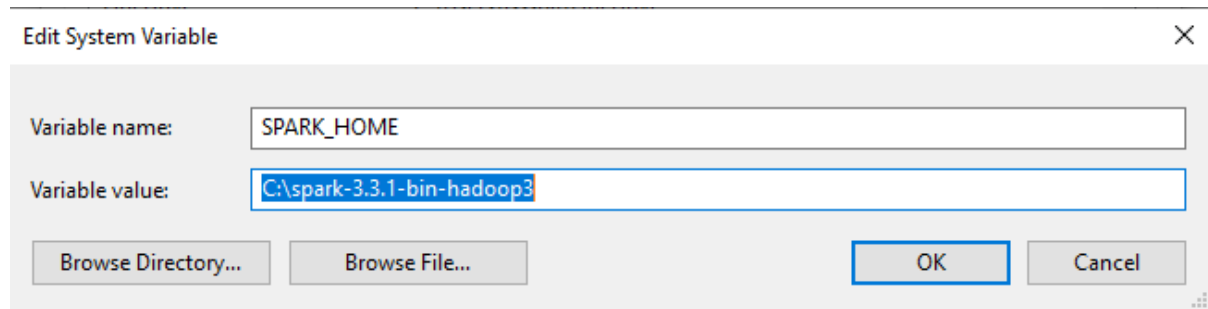
To edit the PATH environment variable in Windows 10:

- Launch "Control Panel" ⇒ (Optional) "System and Security" ⇒ "System" ⇒ Click "Advanced system settings" on the left pane.
- Switch to "Advanced" tab ⇒ Click "Environment Variables" button.
- Under "System Variables" (the bottom pane), scroll down to select variable "Path" ⇒ Click "Edit..."
- You shall see a TABLE listing all the existing PATH entries. Click "New" ⇒ Click "Browse" and navigate to your JDK's "bin" directory, i.e., "C:\Program Files\Java\jdk-11.0.16.1\bin", where {x} is your installation update number ⇒ Select "Move Up" to move this entry all the way to the TOP.
- Create a new path variable called **JAVA\_HOME** and this variable points to 'C:\Program Files\Java\jdk-11.0.16.1'
- Create new path variable **HADOOP\_HOME** is created and this variable points to C:\winutils
- Create new path variable **SPARK\_HOME** is created and this variable points to C:\winutils

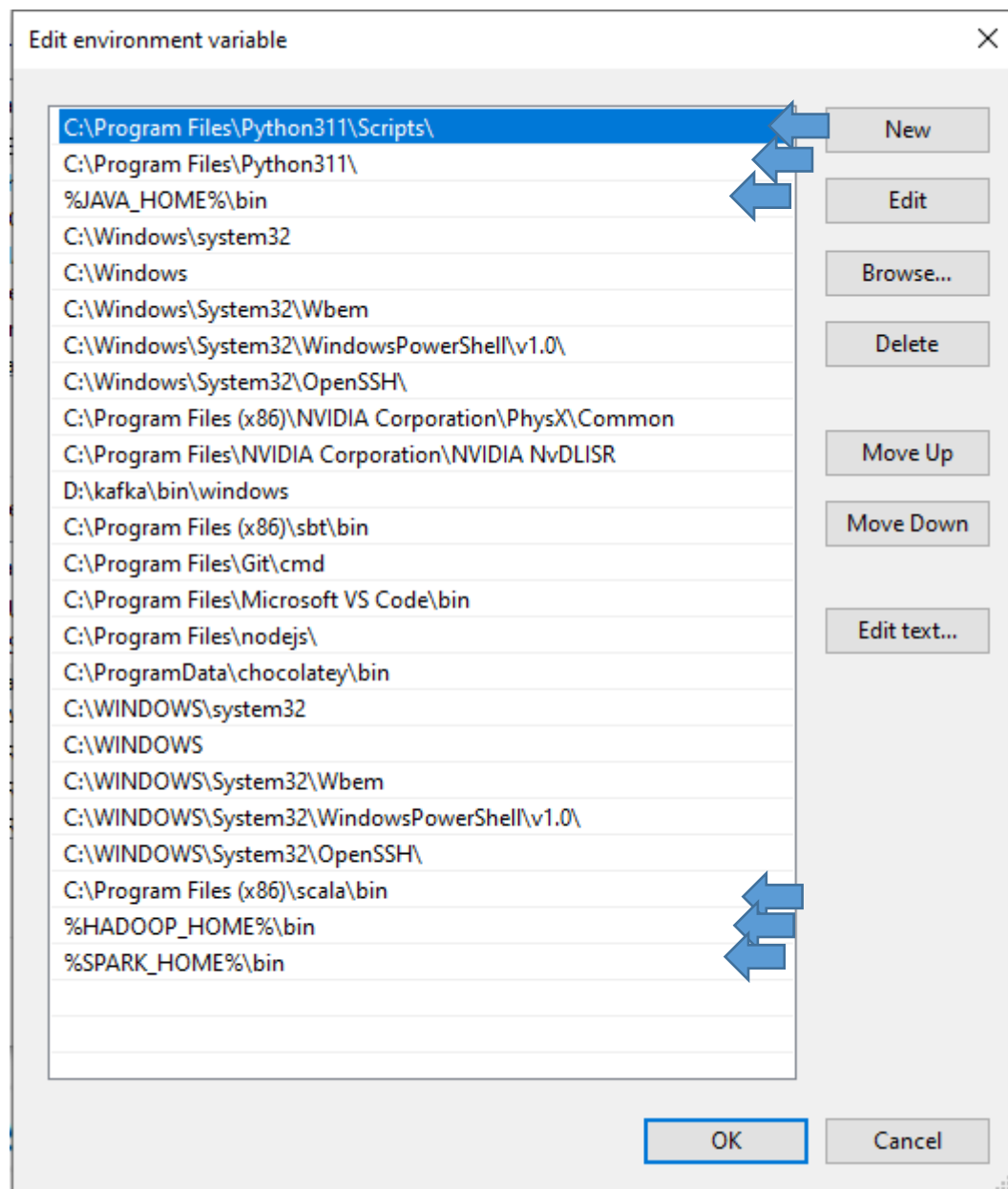
Set New Path Variables (HADOOP\_HOME and SPARK\_HOME)







Check if the system variable **PATH** includes Java bin folder, Scala bin folder and Python bin/scripts folder. Also add Hadoop and spark bin folders as shown below:



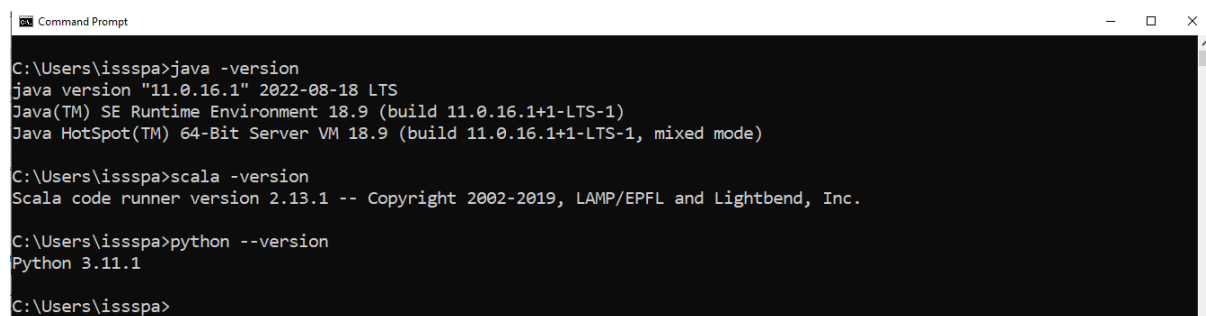
## Environment Testing

In this section, we will verify if the installation process we have followed thus far is working. For this we can use the old cmd prompt or powershell utility that is new in Windows 10 in administrative mode.

### Testing the pre-requisites

Launch a CMD via one of the following means:

1. Click "Search" button ⇒ Type "cmd" ⇒ Choose "Command Prompt", or
2. Right-click "Start" button ⇒ run... ⇒ enter "cmd", or you can also use power shell.
3. Type "java -version" and verify the proper installation of JDK from previous steps.
4. Type scala -version and verify the proper installation of Scala from previous steps.
5. Type python --version and verify the proper installation of Python from previous steps.



```

C:\Users\issspa>java -version
java version "11.0.16.1" 2022-08-18 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.16.1+1-LTS-1)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.16.1+1-LTS-1, mixed mode)

C:\Users\issspa>scala -version
Scala code runner version 2.13.1 -- Copyright 2002-2019, LAMP/EPFL and Lightbend, Inc.

C:\Users\issspa>python --version
Python 3.11.1

C:\Users\issspa>
  
```

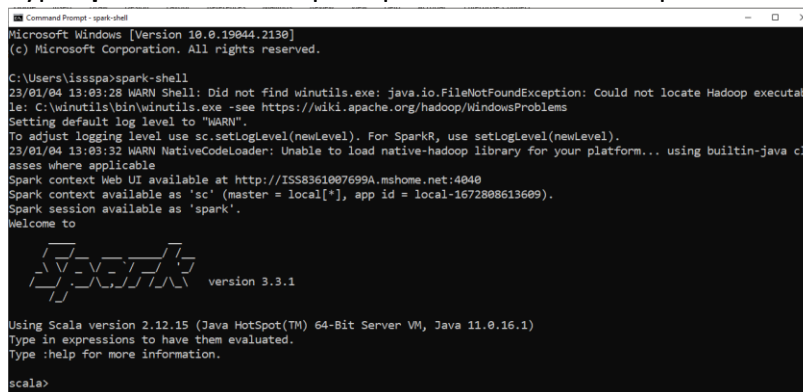
```

C:\Users\issspa> java -version
java version "11.0.16.1" 2022-08-18 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.16.1+1-LTS-1)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.16.1+1-LTS-1, mixed mode)
PS C:\Users\issspa> scala -version
Scala code runner version 2.13.1 -- Copyright 2002-2019, LAMP/EPFL and Lightbend, Inc.
PS C:\Users\issspa> python --version
Python 3.11.1
PS C:\Users\issspa>
  
```

### Testing Spark Installation

Launch a CMD via one of the following means:

1. Click "Search" button ⇒ Type "cmd" ⇒ Choose "Command Prompt", or
2. Right-click "Start" button ⇒ run... ⇒ enter "cmd", or you can also use power shell.
3. Type **spark-shell** in the prompt to see similar interpreter as shown below:



```

Microsoft Windows [Version 10.0.19044.2130]
(c) Microsoft Corporation. All rights reserved.

C:\Users\issspa>spark-shell
23/01/04 13:03:28 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Could not locate Hadoop executable: C:\winutils\bin\winutils.exe -see https://wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/01/04 13:03:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://ISS8361087699A.mshome.net:4040
Spark context available as 'sc' (master = local[*], app id = local-1672808613609).
Spark session available as 'spark'.
Welcome to

  ____  __  _ __
 ___/  /_  /  /_  _ __
/ _ \ / __/  / __/  / __/
/ ___//_/  /_/  /_/  /_/
version 3.3.1

Using Scala version 2.12.15 (Java HotSpot(TM) 64-Bit Server VM, Java 11.0.16.1)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
  
```

4. Open another cmd prompt. Type pyspark in the prompt to see as shown below:

```

Microsoft Windows [Version 10.0.19044.2130]
(c) Microsoft Corporation. All rights reserved.

C:\Users\issga>pyspark
Python 3.11.1 (tags/v3.11.1:a7a450f, Dec 6 2022, 19:58:39) [MSC v.1924 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
23/01/04 13:05:06 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Could not locate Hadoop executab
le: C:\winutils\bin\winutils.exe -see https://wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/01/04 13:05:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
23/01/04 13:05:08 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to
Spark version 3.3.1

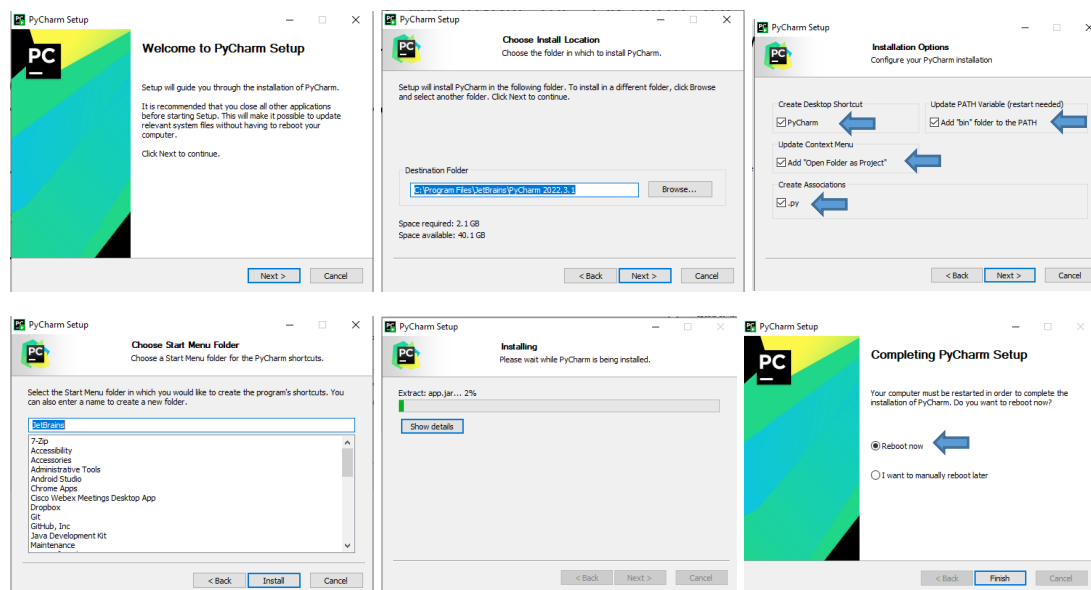
Using Python version 3.11.1 (tags/v3.11.1:a7a450f, Dec 6 2022 19:58:39)
Spark context Web UI available at http://ISS8361007699A.ms.home.net:4041
Spark context available as 'sc' (master = local[*], app id = local-1672888708510).
SparkSession available as 'spark'.
>>> sc

```

Close both the open command prompt windows. If your setup works thus far, please continue to the next section.

## Install PyCharm

In this section, participants will install the PyCharm IDE. PyCharm is an integrated development environment used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django.



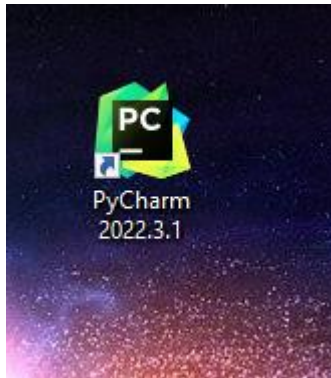
*Note that here we have chosen to set the shortcut, path variable and also associate .py files to this IDE. Also we chose Reboot Now. You can choose per your needs.*

Once installation is complete click Finish.

## Create a PySpark Project using PyCharm

This section teaches you to create a python project and connect to the PySpark libraries. Follow the below instructions as guided by the screenshots for setting up a Python project.

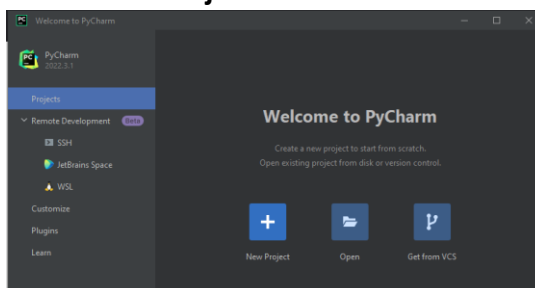
1. Open the PyCharm IDE that you have installed previously.



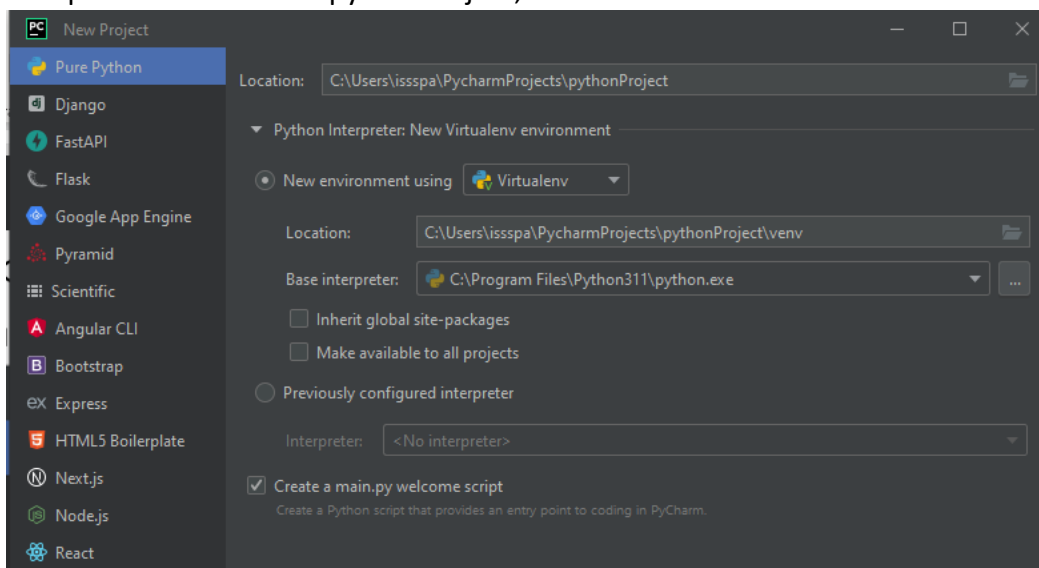
If you are opening for the first time, you may be asked to browse through some IDE tips, until you reach the below screen.

*Note: Your screens could be a little different, as the screen shown here are from professional edition.*

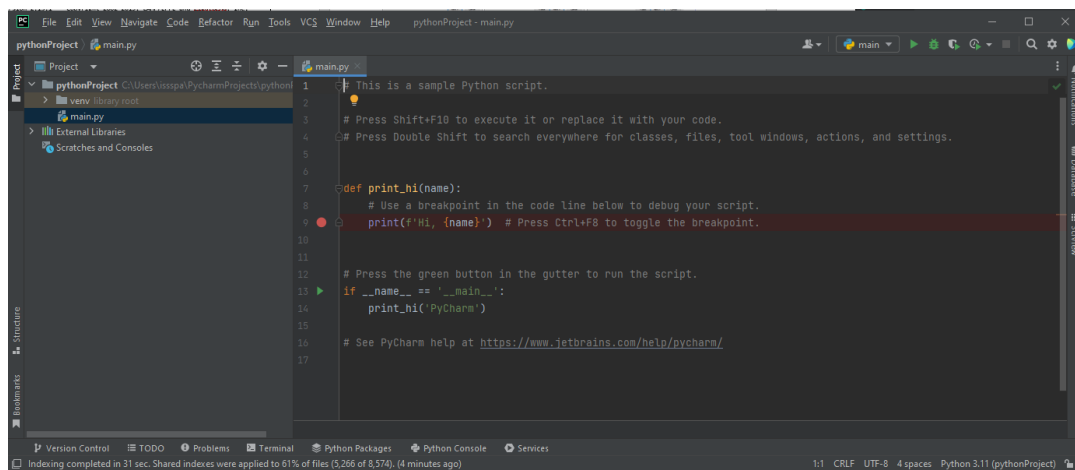
2. Select **New Project**.



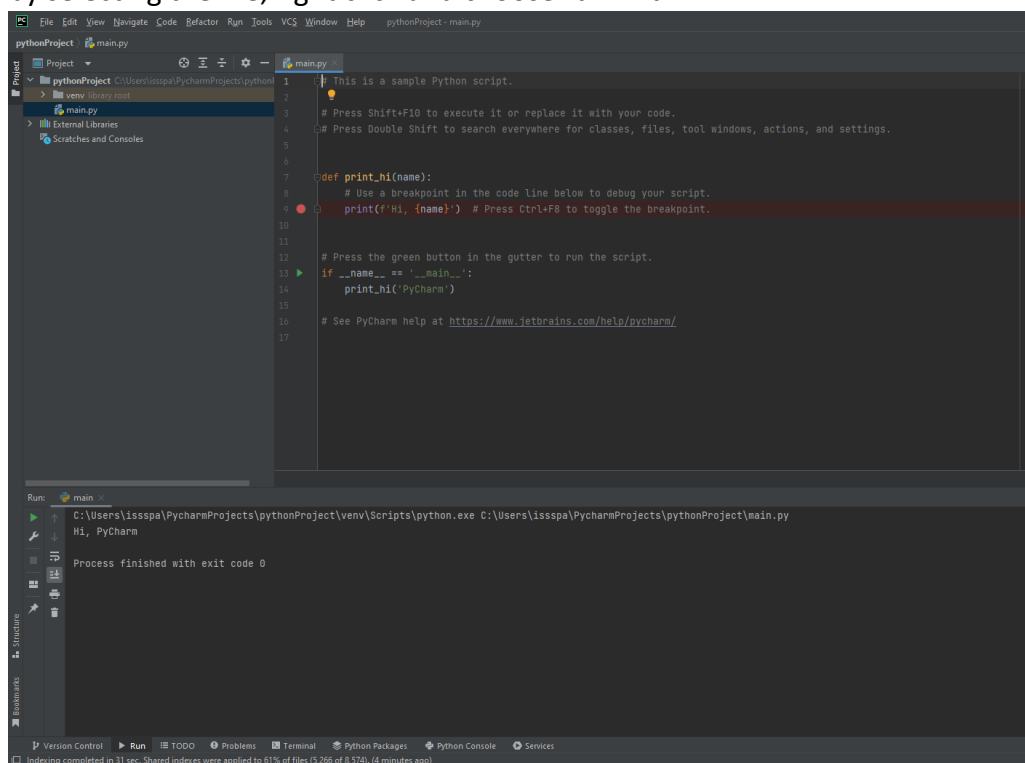
3. In the new project wizard will pop up. Make sure you **pick Python3.11**. Accept default Location: pythonProject; click **Create**.



4. After the tips popup is closes, you will see the below workspace open up.



5. There will be a default main.py file that has a simple print statement. Execute the file by selecting the file, right click and choose **run 'main'**



6. The following message displays

```

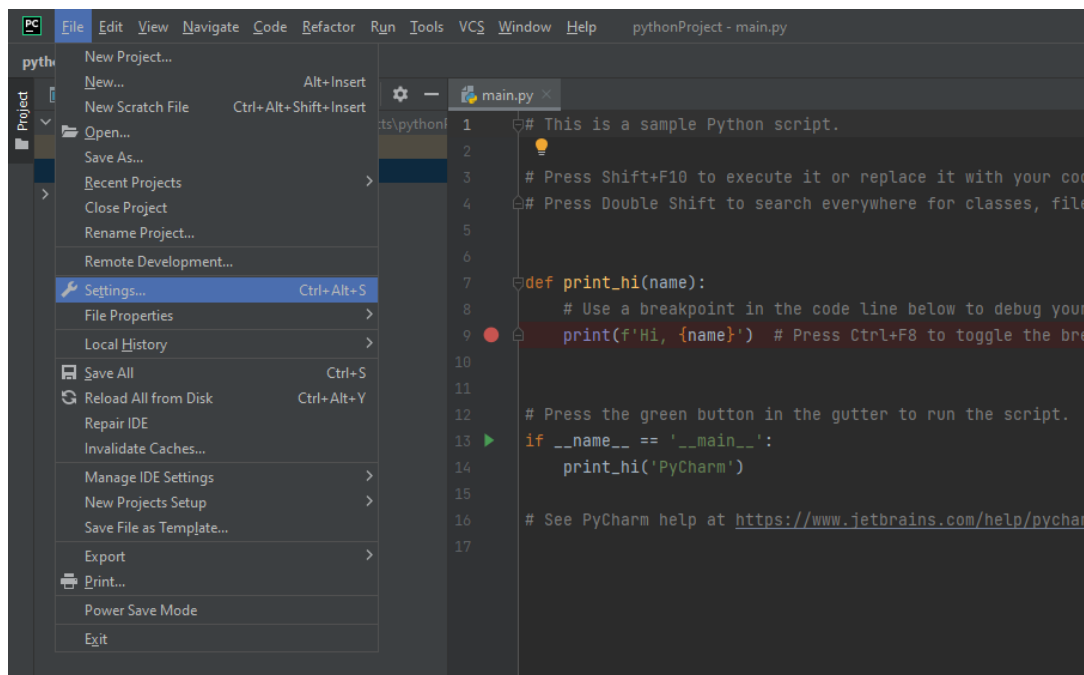
C:\Users\issspa\PycharmProjects\pythonProject\venv\Scripts\python.exe
C:/Users/issspa/PycharmProjects/pythonProject/main.py
Hi, PyCharm

Process finished with exit code 0

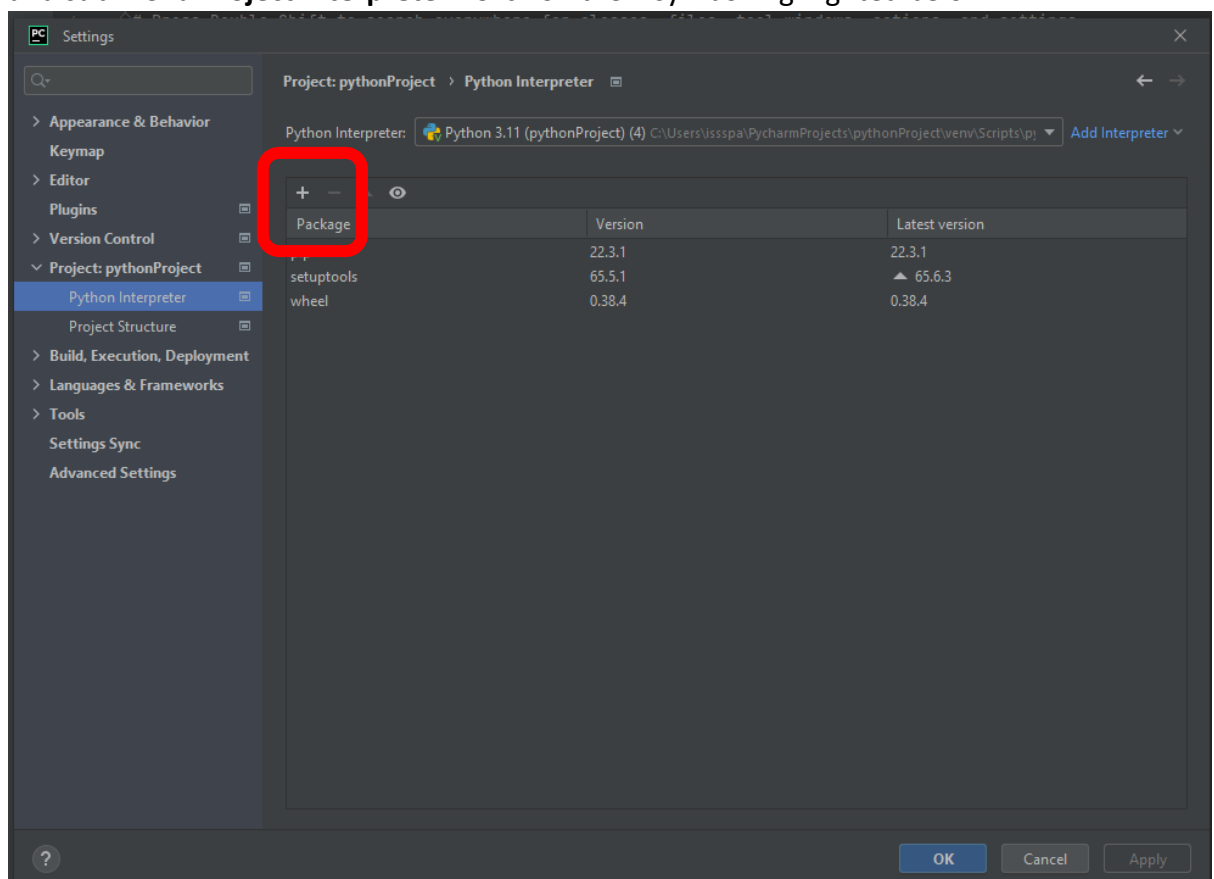
```

## Connect to PySpark libraries

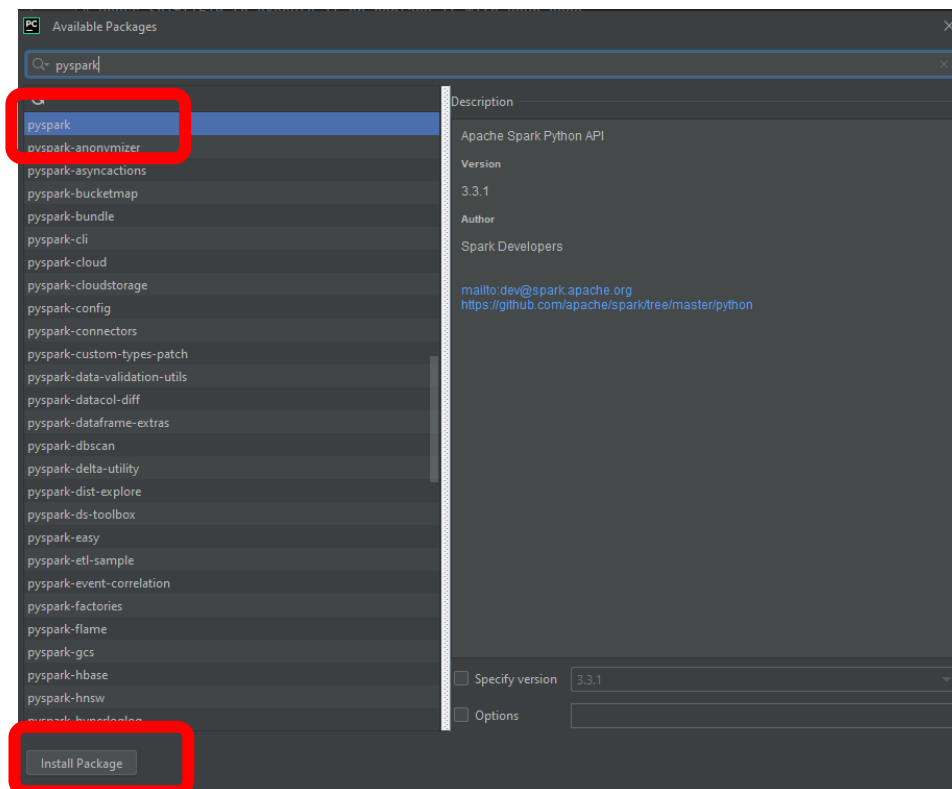
7. Next step is to connect to the spark library. Select the pythonProject, then select File Menu and choose **settings** as shown below



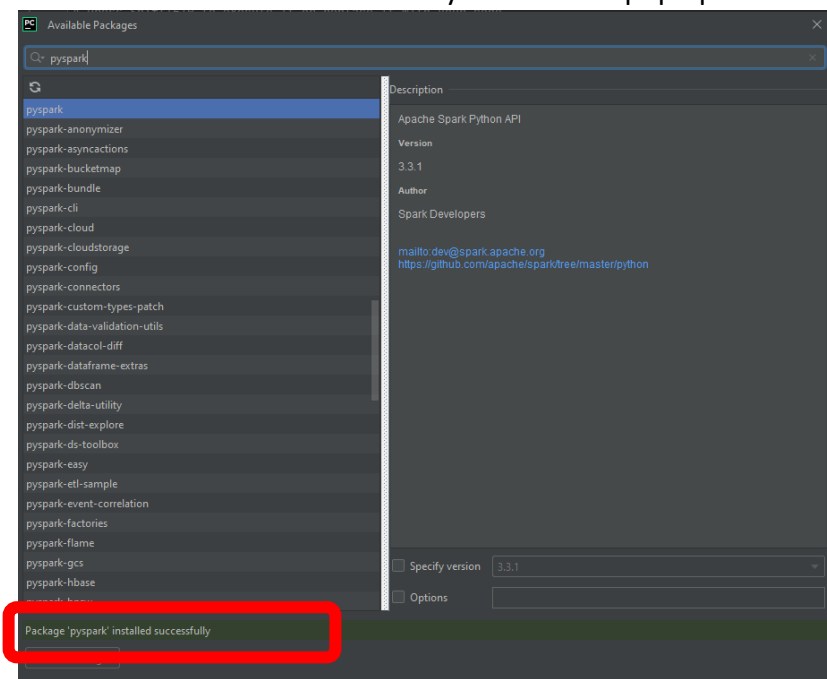
8. The settings wizard opens up as below. Choose the side menu **Project** **pythonProject** and sub menu **Project Interpreter**. Click on the + symbol highlighted below:



9. Once the library window pops up, type pyspark. Select pyspark from the list and click on Install button as highlighted below.



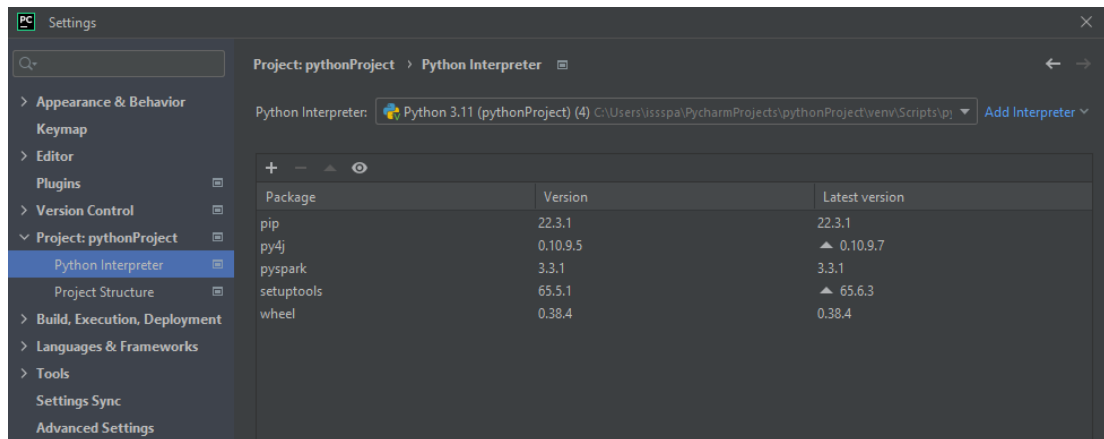
10. Wait for the installation to be over and the successful installation message appears on the bottom of the screen before you close the pop up window.



*Note that here you are installing the latest version 3.3.1. You can also change it to older version by clicking on the specify version radio button and typing the release version.*

11. You will now notice that there are additional necessary libraries added into the project settings.

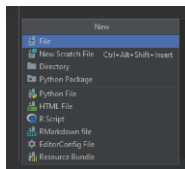




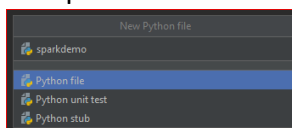
12. Select ok and conform the settings

## Test the PySpark codes

13. Select the **file** menu > **new**. Choose **file** and the list has **python file**, select **python file** as shown below



14. Complete the file name as **pysparkdemo** and select enter key.

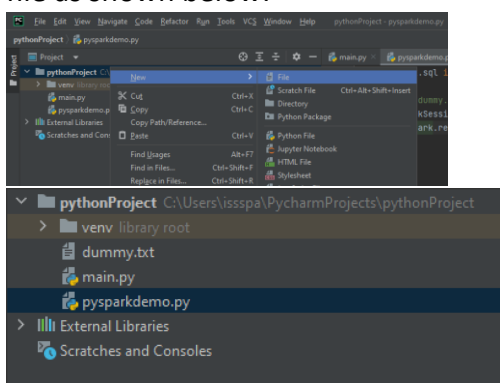


15. Type in the following code:

```
from pyspark.sql import SparkSession

someFile = "dummy.txt"
# the above file is under your pythonProject folder
spark = SparkSession.builder.appName("SimpleApp").getOrCreate()
print(spark.read.text(someFile).count())
```

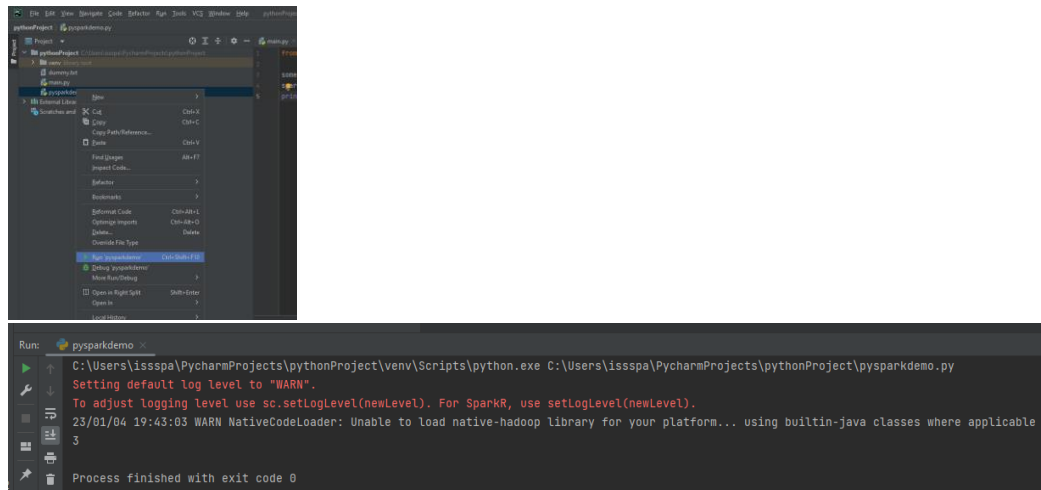
16. Create a file **dummy.txt** in the **pythonProject** folder. You can do this by choosing new file as shown below:



17. Copy the below text and save it in dummy.txt file.

Some  
File  
With random line of text

18. Execute the code by selecting **Run pysparkdemo**. You should see some output like below



The screenshot shows the PyCharm IDE interface. The top part displays the 'Run' menu with options like 'Run', 'Debug', 'Test', etc. The bottom part shows the output console for the 'pysparkdemo' run configuration. The output text is as follows:

```
C:\Users\isspa\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\isspa\PycharmProjects\pythonProject\pysparkdemo.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/01/04 19:43:03 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
3
Process finished with exit code 0
```

*Have fun with coding with python. Feel free to add other spark commands you can think of, use the documentation. Cheers to the craft of creating clean code.*

## Concluding Remarks

Now your Scala and Python kernels for Spark Framework on Windows Platform completely installed. In addition, participants have installed PyCharm IDE and you have learnt how to set up a PySpark project. This would enable participant to continue all the other workshops lined up after this session. We hope this will allow participants to write spark codes that will help process big data for analytics outcomes.

*Have fun coding. Cheers to the craft of creating clean code.*



-----END OF DOCUMENT-----

