PySpark Documentation:

Step1: Download the PyCharm community version from following link

https://www.jetbrains.com/pycharm/download/#section=windows

Step2: Download spark Hadoop

https://dlcdn.apache.org/spark/spark-3.4.0/spark-3.4.0-bin-hadoop3.tgz

Check the actual site, the version may be different.

Step3: Download winutils

https://github.com/steveloughran/winutils

Step4: Download java jdk

https://www.oracle.com/in/java/technologies/downloads/#jdk20-windows

Check the actual site, the version may be different.

Step5: Go to download folder and install PyCharm first.

Step6: Install Java

Step7:

Go to C – drive, create new folder by name "**Spark**"

Copy the downloaded zip file of "**Spark 3.1.2-bin-hadoop-2.7**" (**the version may be different**)

Extract the file here

We get folder by name "**Spark 3.1.2-bin-hadoop-2.7"** (**the version may be different**)

Step8:

Again, go back to C- drive

And create new folder by name "**Hadoop**".

Inside the **Hadoop** folder ---create new folder named "**bin**"

Inside the bin folder copy that downloaded **winutils** file.

Step9:

Go to system environmental variables

Inside the system variable select **new**

Variable name: HADOOP_HOME

Variable values: C:\Hadoop

Note: Give the path of Hadoop folder which we have created in C drive, C drive → Hadoop…. select the path till **Hadoop** folder only.

Click Ok

Step10:

Inside the system variable select **new**

Variable name: SPARK_HOME

Variable values: C:\Spark\spark-3.1.2-bin-hadoop2.7 (**the version may be different**)

Note: Give the path of Spark folder which we have created in C drive, C drive → Spark →Spark-3.1.2-bin-handoop2.7…. select the path till **spark-3.1.2-bin-hadoop2.7** folder name only.

Click Ok

Step11:

Inside the system variable select **new**

Variable name: JAVA_HOME

Variable values: C:\Program Files\Java\jdk1.8.0_321 (**the version may be different**)

Note: Give the path of Java folder, go to C drive → Program files → Java → jdk 1.8.0_321….select the path till **jdk 1.8.0_321 folder** name only.

Step12:

Now go to environmental variable just above the system variable

Select path → edit → new → %SPARK_HOME%\bin

Select path → edit → new → %HADOOP_HOME%\bin

Select path → edit → new → %SPARK_HOME%\python

Select path → edit → new → %PYTHONPATH%

Go to C-drive → Spark → spark-3.1.2-bin-hadoop2.7 →python → lib → py4j-0.10.9-src.zip

Select the path till py4j-0.10.9-src.zip upto this zip file

Again go back to environmental variable and select new path and past it over here.

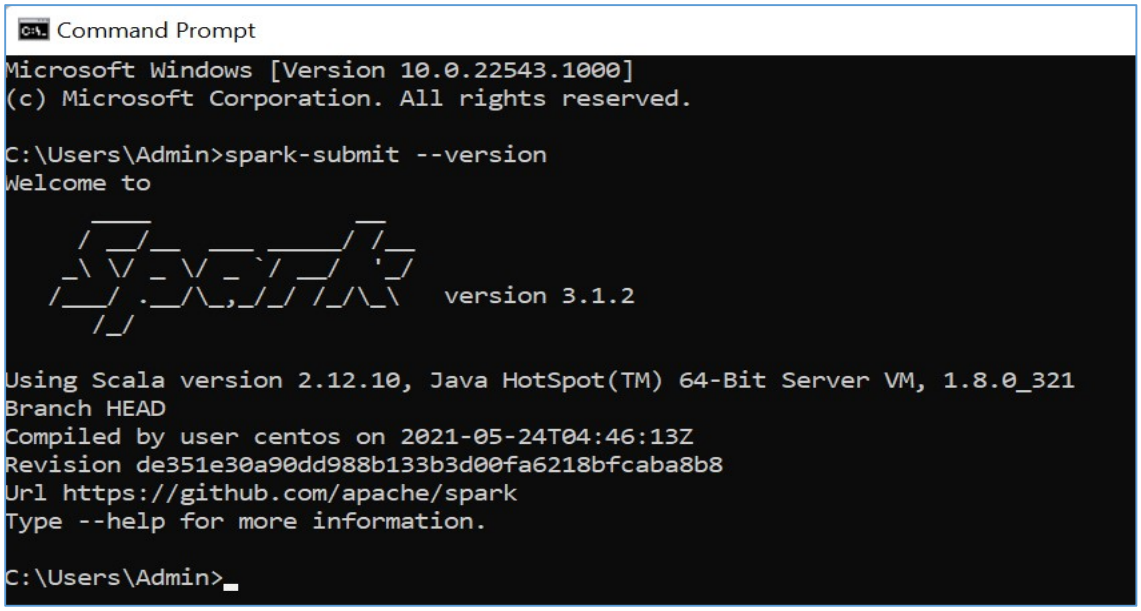
Select path → edit → new → C:\Spark\spark-3.1.2-bin-hadoop2.7\python\lib\py4j-0.10.9-src.zip

Click Ok

Step13: Now go to CMD:

And type the command → **spark-submit --version**

If you get the output as shown below it means Spark has been successfully configured.



Now restart the system.

Step14: Open PyCharm

Click on new Project (+)

Go to file → settings → Project: python project → Python Interpreter → Check the interpreter ..that should be **Python 3.9 latest version.**

Go to file → settings → Project: python project → Project Structure → Add Content Root (right side top corner) → Select Spark → Python → lib → Select both file

py4j-0.10.9-src.zip

pyspark.zip

Apply → Ok.

This is how we have configured our PyCharm with PySpark Interpreter.

Now You can check you PySpark is working properly or not in your PyCharm.

Create the PySpark Session first.

And then try to read any file from your local into Spark dataframe….

To Create session…. Copy this code into pycharm

```
from pyspark.sql import SparkSession
# creating spark session
spark = SparkSession.builder.master("local[1]") \
                    .appName('SparkByExamples.com') \
                    .getOrCreate()

# reading file into spark dataframe
df = spark.read.options(header='True', inferSchema='False', delimiter=',') \
     .csv("C:/Users/Desktop/Data/zipcodes1111.csv")
df.printSchema()
```