Apache Spark Installation Guide – Windows

Pre-requisites

- Java Version: JDK 17.0.17
- Python Version: 3.10.0
- Spark Version: 3.5.7
- Hadoop Version: 3.3.x

Installation Steps

Step 1

Download and install Java from the following URL:
https://www.oracle.com/java/technologies/downloads/#java17-windows

Step 2

Download and install Python from:
https://www.python.org/ftp/python/3.10.0/python-3.10.0-amd64.exe

Step 3

Download Apache Spark (version 3.5.7 built for Hadoop 3) from:
https://www.apache.org/dyn/closer.lua/spark/spark-3.5.7/spark-3.5.7-bin-hadoop3.tgz

Step 4

Extract the downloaded Spark archive and rename the folder to 'spark'.

Step 5

Move the 'spark' folder to the root of drive C: (C:\spark).

Step 6

Download 'winutils.exe' and 'hadoop.dll' from the following GitHub repository:
https://github.com/robguilarr/spark-winutils-3.3.1/blob/master/hadoop-3.3.1/bin

Step 7

Create the following directories:
C:\hadoop\bin
C:\hadoop\tmp

Step 8

Copy the downloaded 'winutils.exe' and 'hadoop.dll' files into:
C:\hadoop\bin

Step 9

Set the following Environment Variables (under System Variables):
SPARK_HOME = C:\spark
JAVA_HOME = C:\Program Files\Java\jdk-17
HADOOP_HOME = C:\hadoop
PYSPARK_PYTHON =
C:\Users\<userName>\AppData\Local\Programs\Python\Python310\python.exe

Step 10

Edit the 'Path' variable and add the following entries:
%HADOOP_HOME%\bin
%SPARK_HOME%\bin
%JAVA_HOME%\bin

Step 11

Configuring pyspark in pycharm
Open pycharm and create project with “Project venv” and choose version python “3.10”
Open file -> Settings -> Project:<projectname> -> Project Structure -> Add Content Root
Add the below 2 files
C:\spark\python\lib\py4j-0.10.9.7-src.zip
C:\spark\python\lib\pyspark.zip