Installing Apache Spark and Python

Windows

Install JDK (Java Development Kit – **8, 11, or 17**) from:

http://www.oracle.com/technetwork/java/javase/downloads/index.html

Download a **pre-built** version of **Apache Spark 3** from:

https://spark.apache.org/downloads.html

1. Extract the Spark archive, and copy its **contents** into **C:\spark**
2. In **c:\spark\conf** folder:
    1. Copy log4j2.properties.template file to log4j2.properties.
    2. Edit log4j2.properties and change the error level from "info" to "error" for log4j.rootCategory
3. Right-click your Windows menu, select System. Click on "Advanced System Settings" and then the "Environment Variables" button.
    1. Add the following new USER variables:

        ```
        SPARK_HOME c:\spark
        PYSPARK_PYTHON python
        ```

    2. Add the following path to your PATH user variable:

        ```
        %SPARK_HOME%\bin
        ```

    3. Close the environment variable screen and the control panels.

Install **Anaconda for Python 3** from:

anaconda.com

1. **Python 3.10** environment. If greater, activate a 3.10 environment.

    Anaconda prompt

    ```
    conda env list
    conda create –n py310 python=3.10
    conda activate py310
    ```

Install **jupyter notebooks**

```
pip install notebook
```

Test:

From Anaconda Prompt:

```
pyspark
rdd = sc.textFile("README.md")
rdd.count()
```

MacOS

Step 1: Install Apache Spark

Using Homebrew

1. Install Homebrew if you don't have it already by entering this from a terminal prompt:
   ```
   /usr/bin/ruby -e "$(curl -fsSL
   https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
   ```
2. Enter:
   ```
   brew install openjdk@17
   brew install scala
   brew install apache-spark
   ```
3. Create a log4j.properties file via
   1. cd /opt/homebrew/Cellar/apache-spark/3.5.2/libexec/conf (substitute 3.5.2 for the version actually installed – the path may be slightly different on your system.)
   2. cp log4j2.properties.template log4j2.properties
4. Edit the log4j2.properties file and change the log level from "info" to "error" on log4j.rootCategory.

Step 2: Install Anaconda

Install the latest Anaconda for Python 3 from anaconda.com, if you don't already have Python installed.

Step 3: Test

From Anaconda Prompt:

```
pyspark
rdd = sc.textFile("README.md")
rdd.count()
```

Linux

1. Install Java, Scala, and Spark according to the particulars of your specific OS.

   https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/

2. Install the latest **Anaconda for Python 3** from anaconda.com
3. Test

   From Anaconda Prompt:

   ```
   pyspark
   rdd = sc.textFile("README.md")
   rdd.count()
   ```