PySpark Environment Setup

We will install Spark locally on a Windows 10 Laptop and verify Spark operation with Python using PySpark. We do not use a Docker version locally, but instead install a Windows 10 version directly as a Windows application. Our use of Spark requires these languages present before installation:

* Java 8
* Python
* Scala

Each language will have required environment variables that must be defined, and PATH entries. Details for these languages are described below, following the overview.

## **Spark Installation Overview**

Once xxxx.

Java Related Settings:

You can download the Java 8 SDK from this link <https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>. Follow the on-screen installation instructions from the installer. For windows 10, set these Java Environment Variables:

* JAVA\_OPTIONS=-Xmx512M -Xms512M
* JAVA\_HOME=D:\Java\jdk1.8.0\_241

The PATH should have %JAVA\_HOME%\bin. Here is a sample Java installation verification:

D:\Temp>java -version

java version "1.8.0\_241"

Java(TM) SE Runtime Environment (build 1.8.0\_241-b07)

Java HotSpot(TM) 64-Bit Server VM (build 25.241-b07, mixed mode)

Python Related Settings:

Obtain a windows Python installation from this link: <https://www.python.org/downloads/windows/>. We use Python 3.8.1 for this example. A quick overview of the Windows 10 installation is found at <https://realpython.com/installing-python/#step-2-run-the-installer>. Make sure that Python is on the Windows execution path. For example, my system has these PATH entries:

* D:\Program Files\Python38\Scripts\
* D:\Program Files\Python38\

Python installation verification is:

D:\Temp>python --version

Python 3.8.1

Scala Related Settings

Scala **is** a found at <https://www.scala-lang.org/download/>. There are several methods of installing Scala shown at this site. For this example, we just download the Scala Windows MSI installer (scala-2.13.1.msi.) Please set the SCALA\_HOME environment variable and the PATH should include the Scala home. For example:

* SCALA\_HOME=D:\bin\scala
* PATH has %SCALA\_HOME%\bin

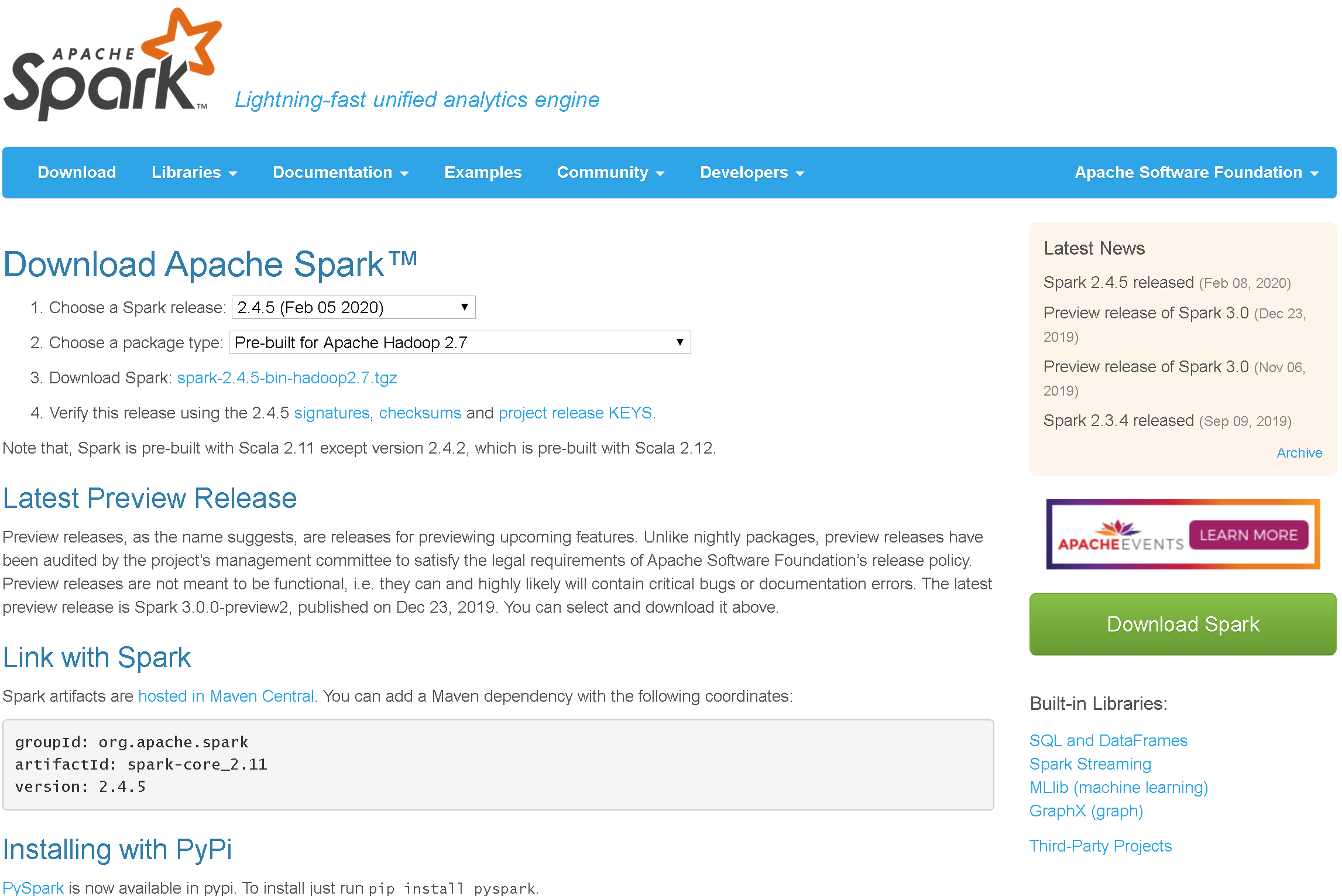
You may wish to review this additional link for Scala installation: <https://www.tutorialspoint.com/scala/scala_environment_setup.htm>. Example Scala verification is:

D:\Temp>scala -version

Scala code runner version 2.13.1 -- Copyright 2002-2019, LAMP/EPFL and Lightbend, Inc.

Spark Related Settings

The Spark download URL used for this article: <https://www.apache.org/dyn/closer.lua/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz>.



Spark Supporting Utilities

A Windows 10 installation of Spark requires the WINUTILS utility.

## Sample Collected Files

D:\Tools\DEV>tree/f

D:.

├───Java

│ jdk-8u241-docs-all.zip

│ jdk-8u241-windows-x64.exe

│

├───Python

│ pycharm-community-2019.3.3.exe

│ python-3.8.1-amd64.exe

│

├───Scala

│ ideaIC-2019.3.2.exe

│ scala-2.13.1.msi

│

├───Spark

│ spark-2.4.5-bin-hadoop2.7.tgz

## Web Resources

1. UCSD instructions (2020): <https://mas-dse.github.io/DSE230/installation/windows/>.
2. Created around 3/2018: <https://medium.com/big-data-engineering/how-to-install-apache-spark-2-x-in-your-pc-e2047246ffc3>.
3. Created around 10/2019: <https://www.knowledgehut.com/blog/big-data/how-to-install-apache-spark-on-windows>.
4. A good overview from 1/2019: <https://medium.com/@naomi.fridman/install-pyspark-to-run-on-jupyter-notebook-on-windows-4ec2009de21f>.
5. Some winutils.exe replacement background: <https://dataintoresults.com/post/apache-spark-on-windows-without-winutils-exe/>.