**Setting up pyspark in windows (Win):**

\*\*\* Can’t say it’s the best way to run pyspark in Win because Linux is more compatible for java apps.

Identify the Pyspark-Hadoop version you want to use in windows. Here I’m using

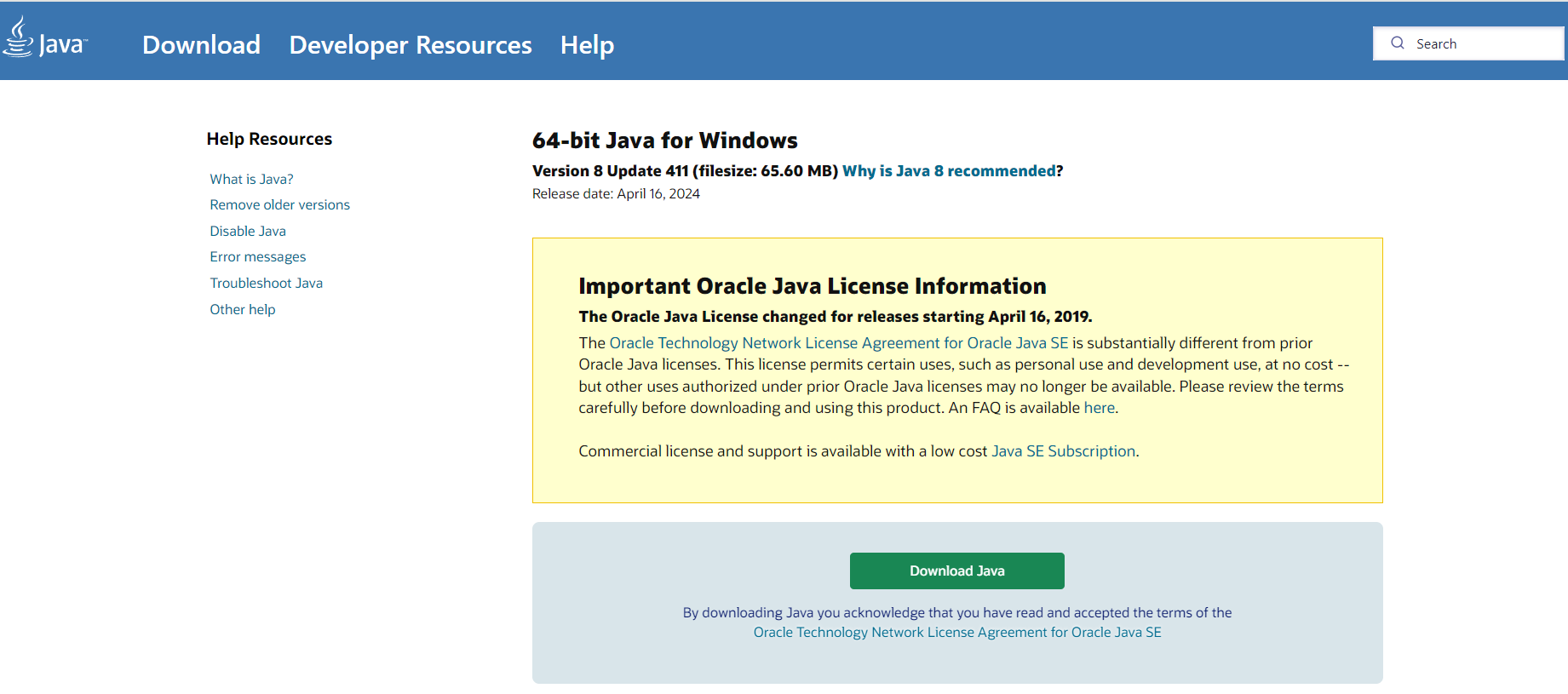
Pyspark version = 3.4.3

Hadoop version =3.0.0

Steps:

1. Download 64-bit Java 8 for Win using <https://www.java.com/en/download/> .

Download and install the jre. If you have any other jdk, don’t forget to uninstall it before installing the new one.



* 1. Set system env variable for Java.

JAVA\_HOME = <jre\_path>

1. Download the Pyspark version from Apache site and select

Pyspark version = 3.4.3

Hadoop version =3.0.0

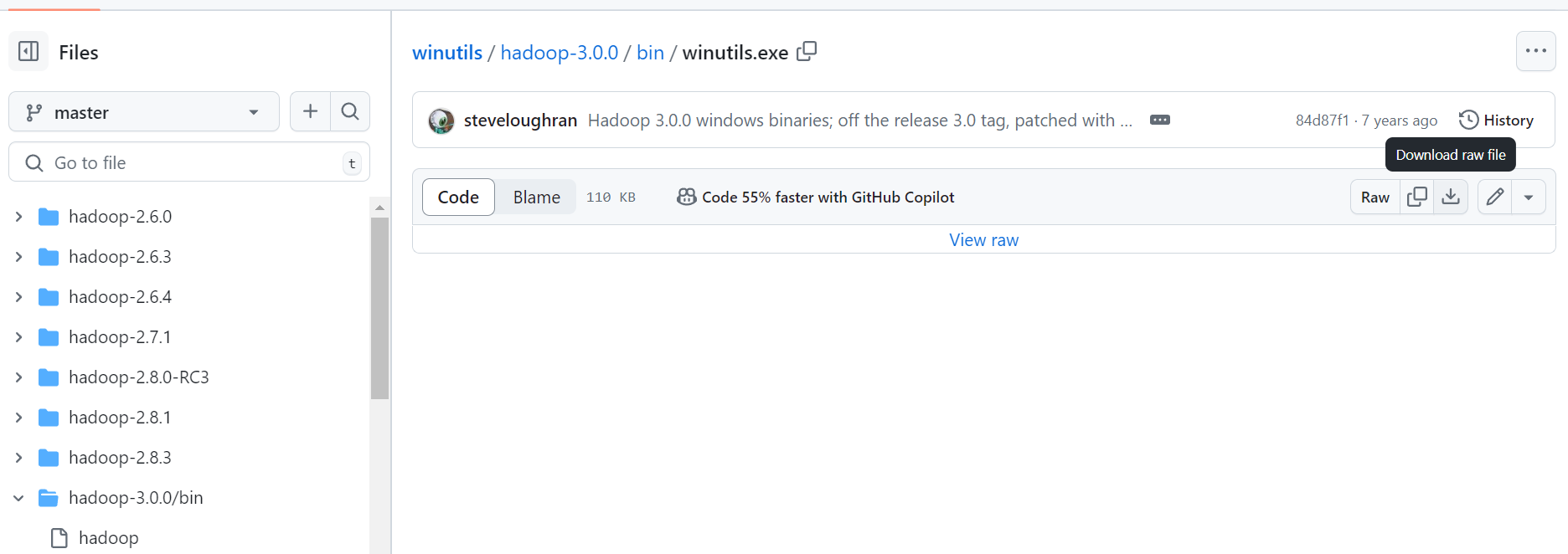
Refer <https://sparkbyexamples.com/pyspark/how-to-install-and-run-pyspark-on-windows/> .

* 1. Extract the downloaded zip file.
  2. Move the extracted folder to C:\apps\ folder (Create ‘apps’ folder if C drive doesn’t contain apps folder).
  3. Set system env variable for spark.

SPARK\_HOME = C:\apps\spark-3.4.3-bin-hadoop3-scala2.13

1. Download Winutil.exe for Hadoop v3 from <https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe> .

Winutil.exe for other Hadoop versions are also available.

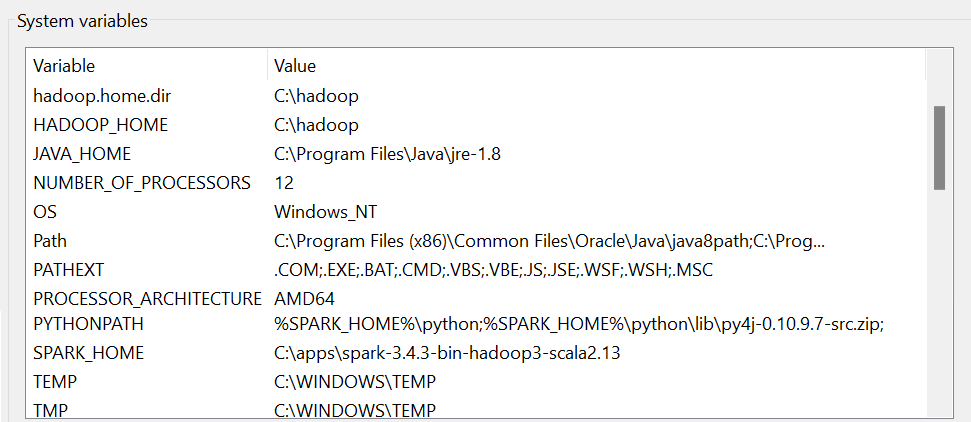


1. Copy the downloaded Winutil.exe to a location.
   1. Create a folder named ‘hadoop’ in C drive.
   2. Create a ‘bin’ folder inside ‘hadoop’ folder.
   3. Copy the downloaded Winutil.exe into bin folder.
   4. Set env variable for Hadoop.

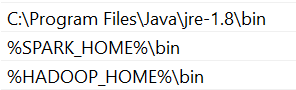
HADOOP\_HOME = C:\hadoop

Hadoop.home.dir = C:\hadoop

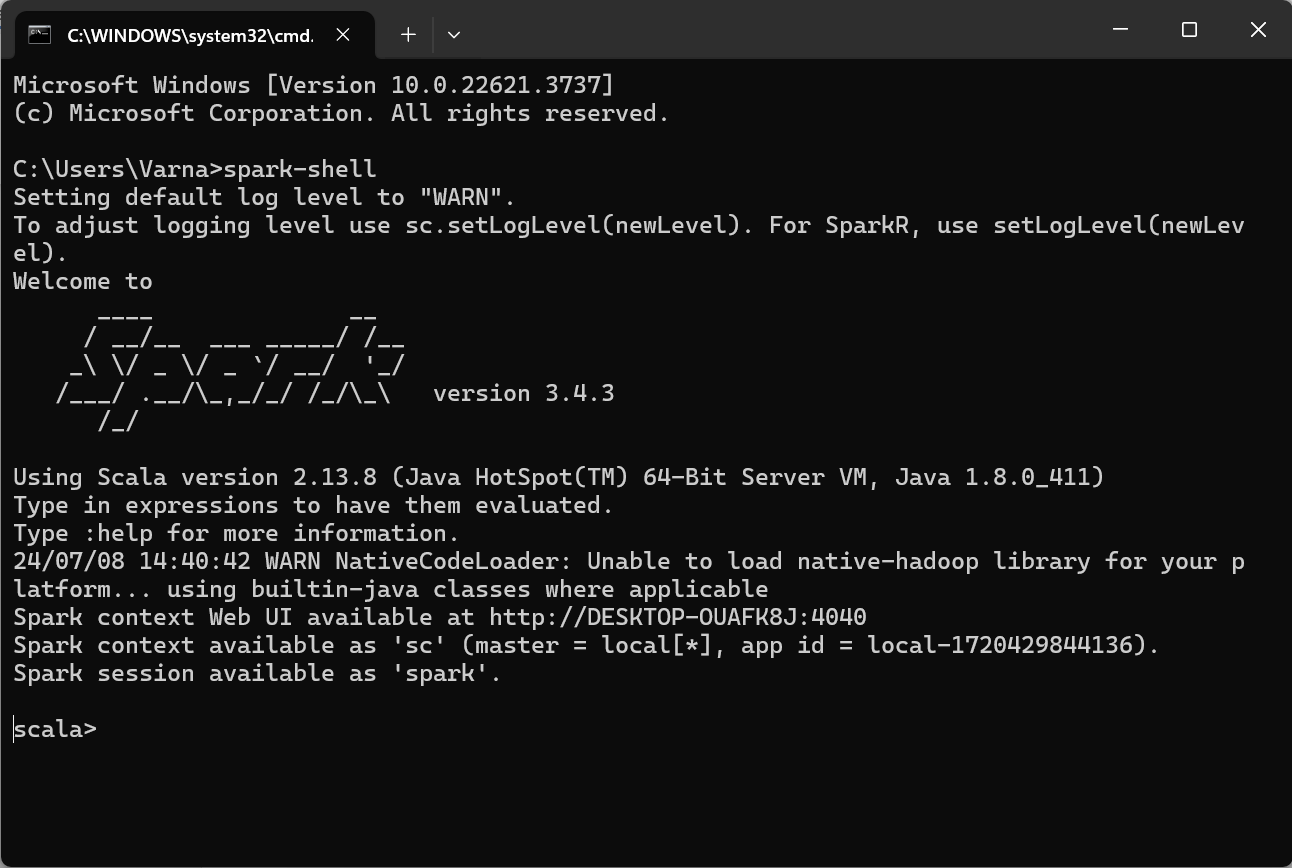
1. System Env variables will look like this:



1. Path variable in System Env variables will look like:



1. Launch spark shell to check the set-up
   1. Open cmd in Admin mode.
   2. $ spark-shell



* 1. Open spark web UI in <http://localhost:4040/> to verify installation
  2. To close the shell,

$ System.exit(0)