**Setting up a Big-data Platform for Machine Learning** **using Spark, Python, and SQL Server**

Apache Spark is an open-source framework that processes large volumes of stream data from multiple sources. Spark is used in distributed computing with parallel machine learning models and parallel data analytics. Python is an open-source programming language can be used to handle big data analytics using PySpark libraries. SQL Server is a database management system capable to store and manage big data. The current document describes the process of setting a big-data platform on Windows 10 operating system using Spark, Python, and SQL Server.

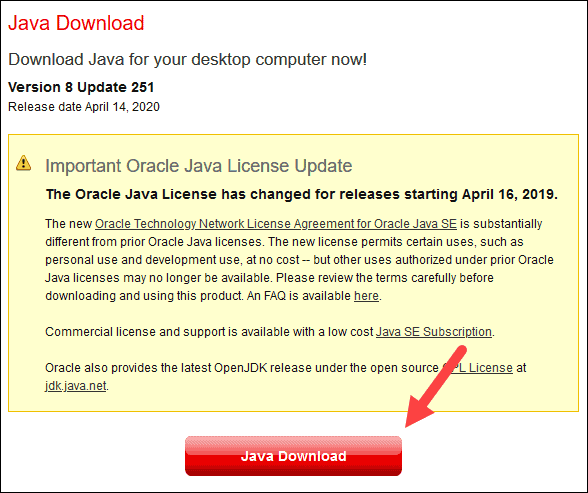
**Prerequisites**

* A standalone system or virtual machine running Windows 10
* A user account with administrator privileges (required to install software, modify file permissions, and modify system PATH)
* A tool to extract .tar files, such as 7-Zip

**Step 1: Install Java**

Apache Spark requires Java Runtime Environment (JRE). Install JRE from the following link:

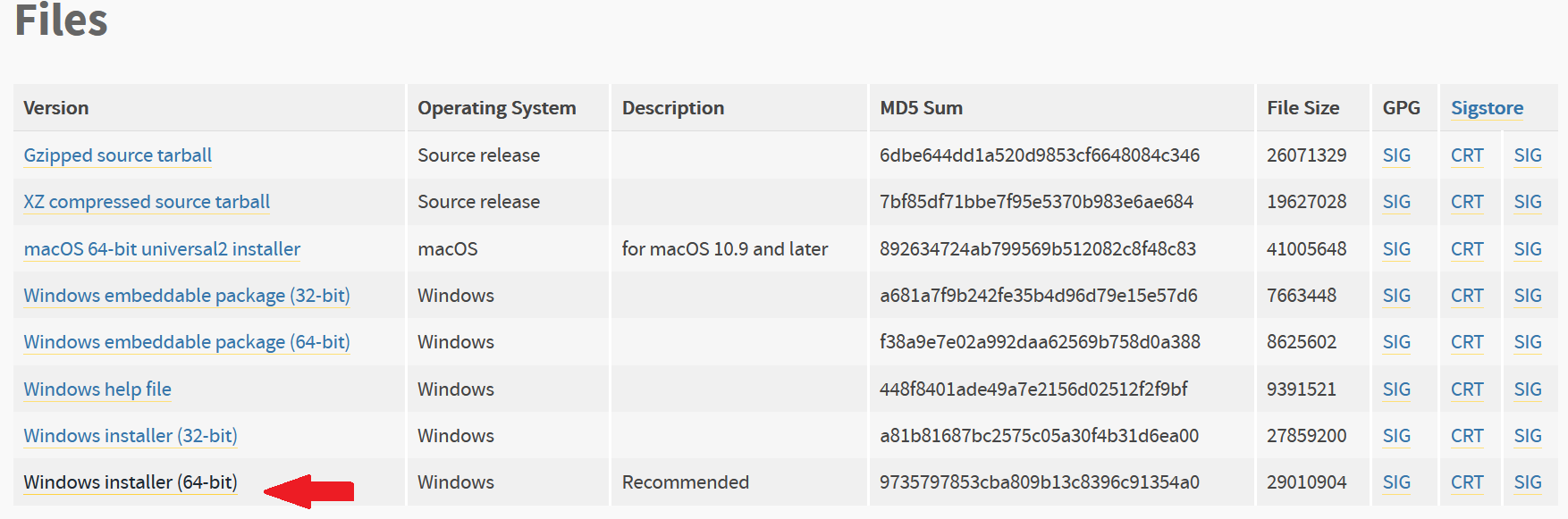
<https://java.com/en/download/>



**Step 2: Install Python**

Download the Python 3.10 from the following link:

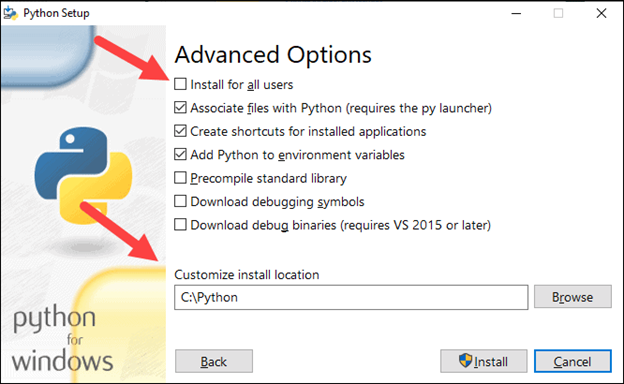
<https://www.python.org/downloads/release/python-31010/>



Next, run the downloaded file and click **Customize installation**.



Under *Customize install location,* click **Browse** and navigate to the C drive. Add a new folder and name it *Python*. Select the box **Install for all users** and leave other boxes as they are.



**Step 3: Install Apache Spark**

Download Spark 3.3.2 from the following link:

<https://www.apache.org/dyn/closer.lua/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz>

Create a new folder named *Spark* in the root of your C: drive then extract the file into *C:\Spark* (e.g., using 7-Zip). Now, your *C:\Spark* folder has a new folder *spark-3.3.2-bin-hadoop3* with the necessary files inside.

**Step 4: Add winutils.exe File**

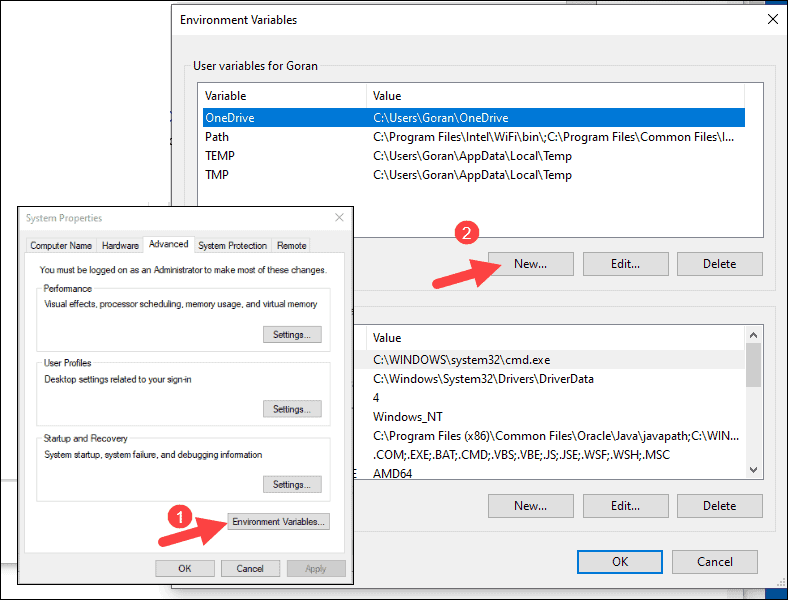
Download the **winutils.exe** file for the Spark installation you downloaded.

<https://github.com/kontext-tech/winutils/raw/master/hadoop-3.3.0/bin/winutils.exe>

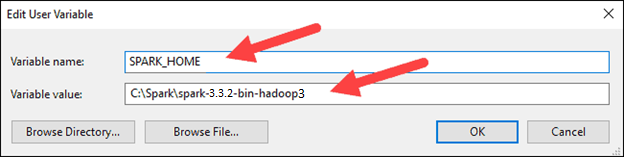
Now, create **Hadoop** folder on C: drive and **bin** folder inside **Hadoop** and copy the winutils.exe file to **C:\hadoop\bin**.

**Step 5: Configure Environment Variables**

Select the result labeled **Edit the system environment variables**from start menu then click **New** in the next window.

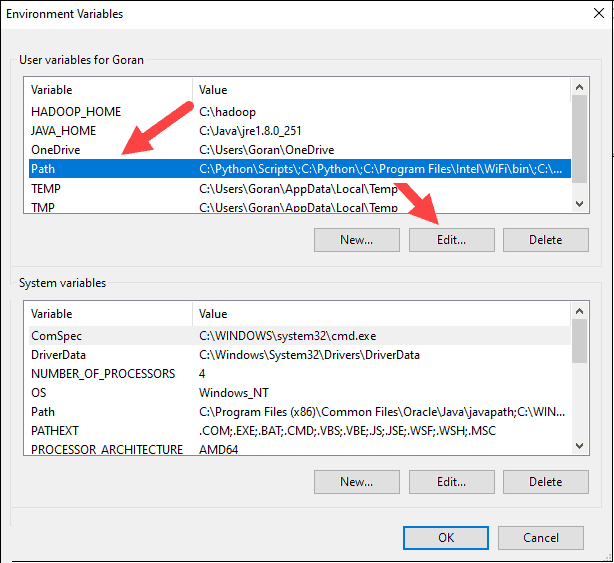


For *Variable Name* type ***SPARK\_HOME*** and for *Variable Value* type **C:\Spark\spark-3.3.2-bin-hadoop3** and click OK

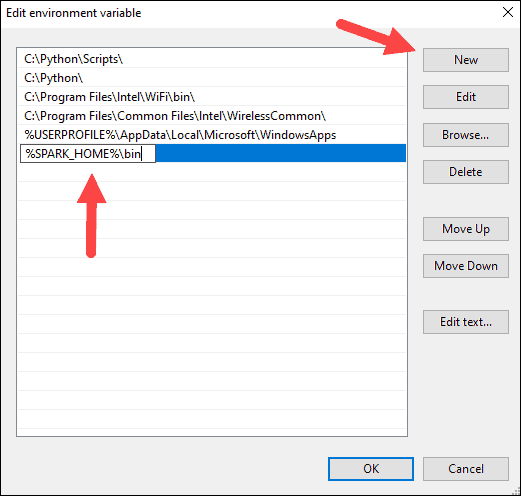


Repeat this step for three other environment variables. For the first one, use **PYSPARK\_DRIVER\_PYTHON** as *Variable Name* and path to *python.exe* (e.g., C:\Users\MyUserName\AppData\Local\Programs\Python\Python310\python.exe) as *Variable Value.* For the second variable, use **PYSPARK\_PYTHON** as *Variable Name* and path to *python.exe* as *Variable Value*. Lastly, define **PYTHONPATH** as a variable with value of **C:\Spark\spark-3.3.2-bin-hadoop3\python.**

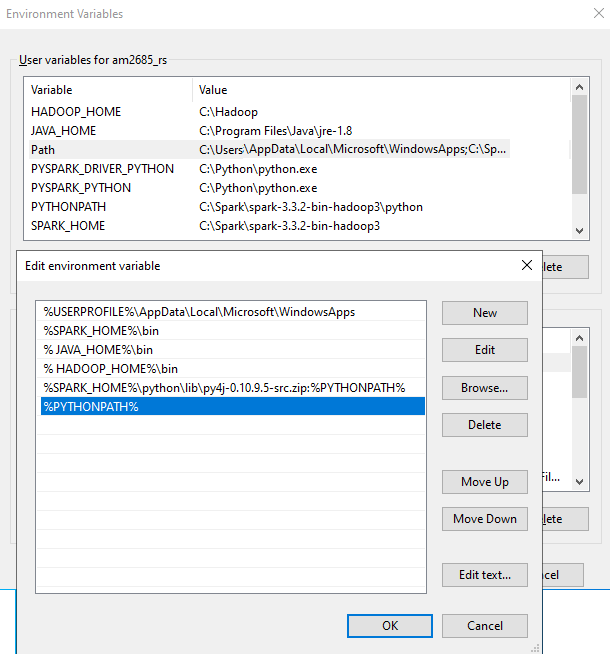
In the top box, click the **Path** entry, then click **Edit**.



Create a **New** entry and enter the path of the bin from Spark folder using **%SPARK\_HOME%\bin**



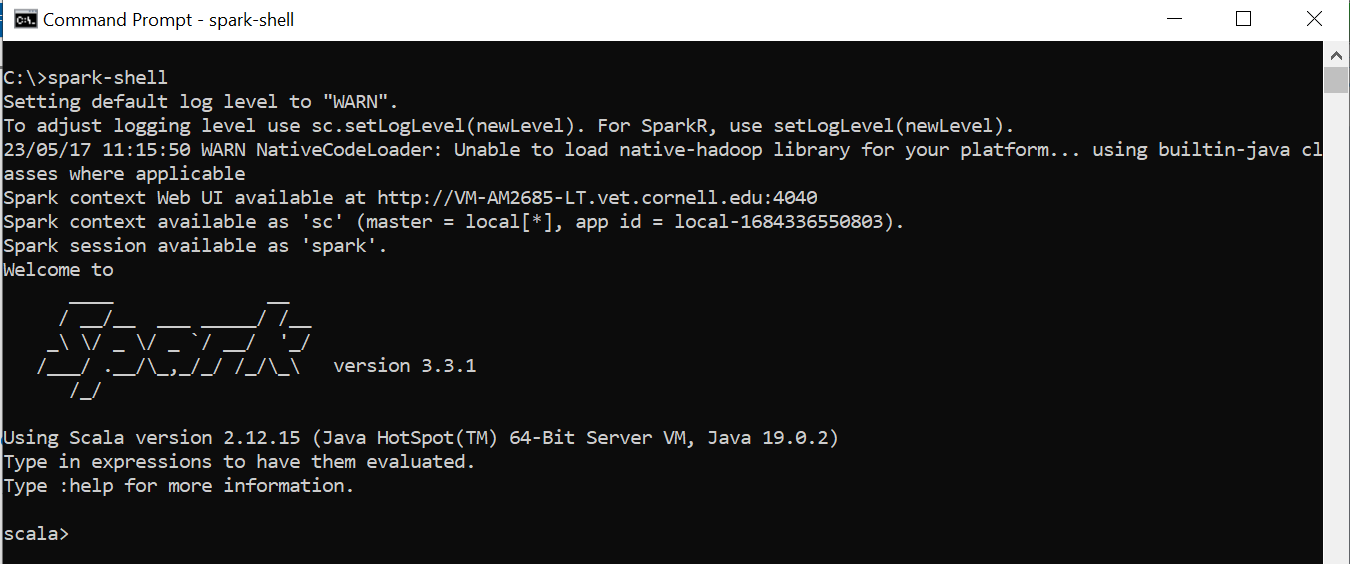
Repeat this step for Hadoop and Java. For Hadoop, the variable name is **HADOOP\_HOME** and for the value use the path of the folder you created earlier: **C:\hadoop.** Add the **Path** variableusing **%HADOOP\_HOME%\bin.** For Java, the variable name is **JAVA\_HOME** and for the value use the path to your Java runtime environment directory (e.g., **C:\Program Files\Java\** **jre-1.8**). Add the **Path** variableusing **% JAVA\_HOME%\bin.** Also add **%SPARK\_HOME%\python\lib\[Name of py4j source zip file]:%PYTHONPATH%** and **%PYTHONPATH%** to the **Path** variable.

****

Click **OK** to close all open windows then reboot the system.

To check the process, open a new command-prompt window using the right-click and **Run as administrator** and type **spark-shell** to launch Spark.

The system should display several lines indicating the status of the application. You may get a Java pop-up. Select **Allow access** to continue. Finally, the Spark logo appears, and the prompt displays the **Scala shell**.



**Step 6: Install SQL Server**

Install the Developer Edition of SQL Server (Basic option) using the following link:

<https://go.microsoft.com/fwlink/p/?linkid=2215158&clcid=0x409&culture=en-us&country=us>

A screenshot of a computer

Description automatically generated with medium confidence

Lastly, install SQL Management Studio by clicking on the **Install SSM**S button. In the Download SQL Server Management Studio page click on *Free Download for SQL Server Management Studio* link (e.g., <https://aka.ms/ssmsfullsetup>) then download and execute the file.

Then run SQL Management Studio and enter a point (.) as server name then press **Connect**.

A screenshot of a computer

Description automatically generated with medium confidence

Connecting to the local instance of SQL Server confirms that the installation was successful.