## **Install Apache Spark on Windows**

**Introduction**

Apache Spark is an open-source framework that processes large volumes of stream data from multiple sources. Spark is used in distributed computing with machine learning applications, data analytics, and graph-parallel processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.

This guide will show you **how to install Apache Spark on Windows**and test the installation**.**

**Prerequisite**: Install JAVA 8 SDK

Make sure you installed JAVA 8 SDK on your system by using this command --> **java -version**

This above command to know the java version installed on the system.

### **STEP 1: Go to Apache spark official download page:**

1. Open a browser and navigate to <https://spark.apache.org/downloads.html>.

2. Under the Download Apache Spark heading, there are two drop-down menus. Use the current latest version.

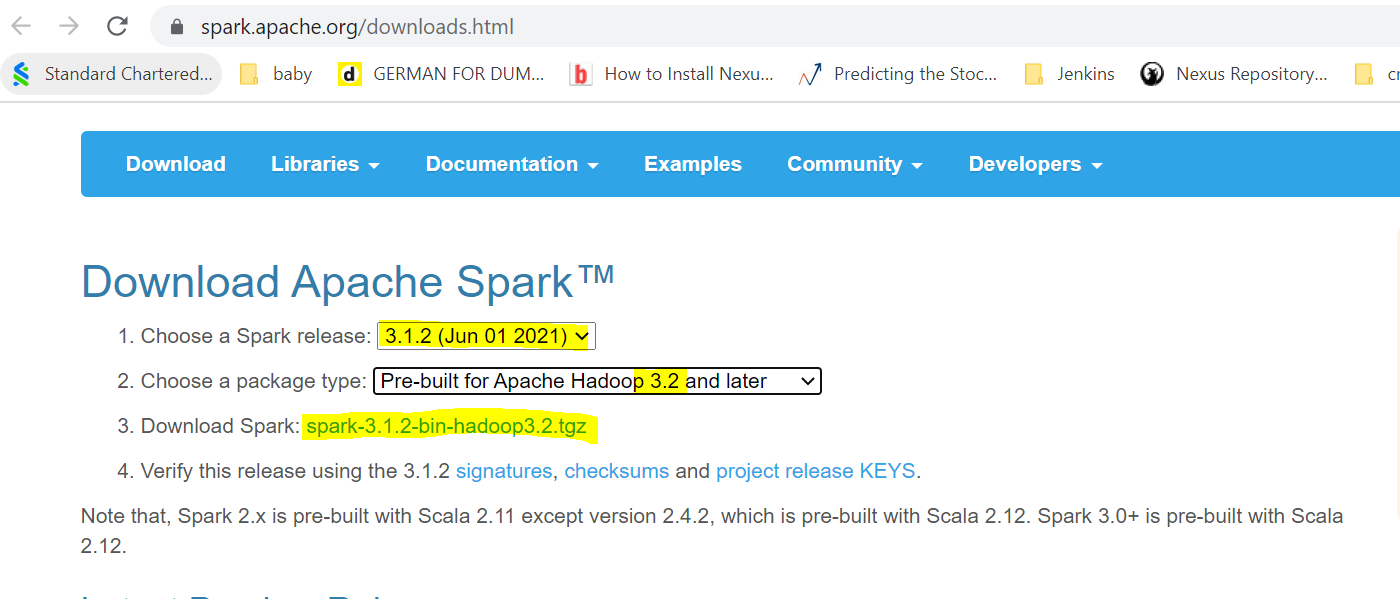
* In our case, in ***Choose a Spark release***drop-down menu select **3.1.2(Jun 01 2021)**.
* In the second drop-down **Choose a package type,** leave the selection **Pre-built for Apache Hadoop 3.2 and later**.

3. Click the **spark-3.1.2-bin-hadoop3.2.tgz**link.

Or

Below is the direct link for download Spark latest 3.1.2

<https://www.apache.org/dyn/closer.lua/spark/spark-3.1.2/spark-3.1.2-bin-hadoop3.2.tgz>



4. A page with a list of mirrors loads where you can see different servers to download from. Pick any from the list and save the file to your Downloads folder.

### Step 4: Verify Spark Software File

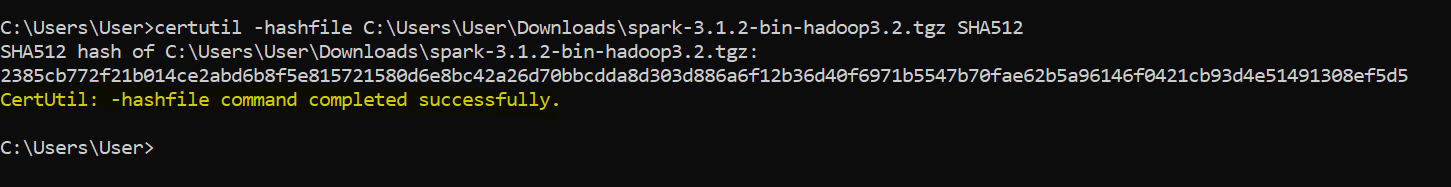
1. Verify the integrity of your download by checking the **checksum** of the file. This ensures you are working with unaltered, uncorrupted software.

2. Navigate back to the Spark Download page and open the **Checksum** link, preferably in a new tab.

3. Next, open a command line and enter the following command:

**certutil -hashfile C:\Users\User\Downloads\spark-3.1.2-bin-hadoop3.2.tgz SHA512**

4. Change the User to your username. The system displays a long alphanumeric code, along with the message **Certutil: -hashfile completed successfully**.



### Step 5: Install Apache Spark

Installing Apache Spark involves **extracting the downloaded file** to the desired location.

1. Create a new folder named Spark in the root of your C: drive. From a command line, enter the following:

cd \

mkdir Spark

2. In Explorer, locate the Spark file you downloaded.

3. Right-click the file and extract it to C:\Spark using the tool you have on your system (e.g., 7-Zip).

4. Now, your C:\Spark folder has a new folder spark-3.1.2-bin-hadoop3.2 with the necessary files inside.

### Step 6: Add winutils.exe File

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

1. Navigate to this URL <https://github.com/cdarlint/winutils> and inside the **bin** folder, locate **winutils.exe**, and click it.

* **Note**: Choose the same version as the package type you choose for the Spark .tgz file you chose in Step 1 *“Spark Download” (in my case: hadoop-3.2.0)*
* You need to navigate inside the **hadoop-X.X.X folder**, and inside the **bin** folder you will find **winutils.exe**



1. Find the **Download**button on the right side to download the file.

Or

If you chose the same version as me (hadoop-3.2.0) here is the direct link: <https://github.com/cdarlint/winutils/blob/master/hadoop-3.2.1/bin/winutils.exe>

3. Now, create new folders **Hadoop**and **bin** on C: using Windows Explorer or the Command Prompt.

4. Copy the winutils.exe file from the Downloads folder to **C:\hadoop\bin**.

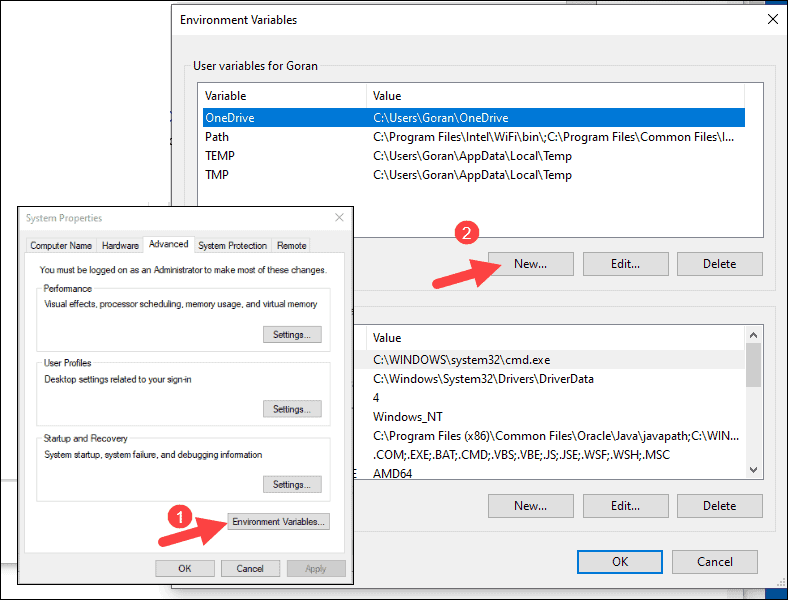
### Step 7: Configure Environment Variables

Configuring environment variables in Windows adds the Spark and Hadoop locations to your system PATH. It allows you to run the Spark shell directly from a command prompt window.

1. Click **Start** and type environment.

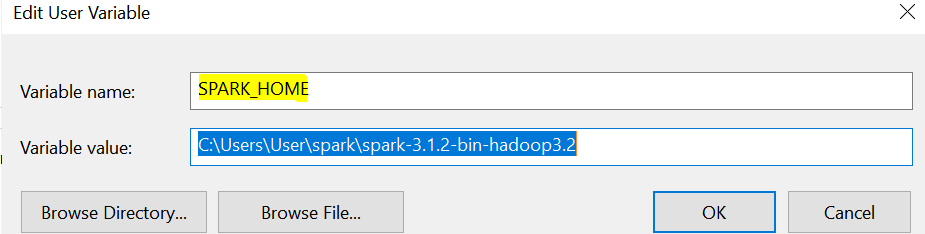
2. Select the result labeled **Edit the system environment variables**.

3. A System Properties dialog box appears. In the lower-right corner, click **Environment Variables** and then click **New** in the next window.

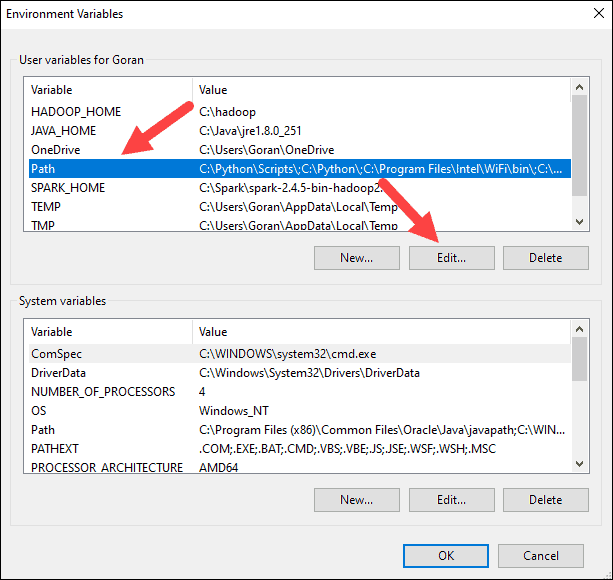


4. For Variable Name type ***SPARK\_HOME***.

5. For Variable Value type **C:\Spark\spark-3.1.2-bin-hadoop3.2** and click OK. If you changed the folder path, use that one instead.

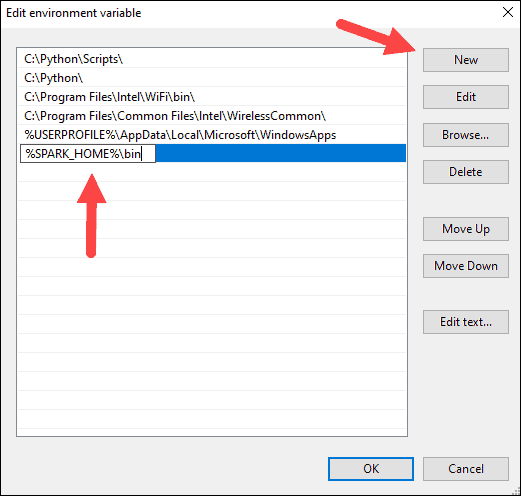


6. In the top box, click the **Path** entry, then click **Edit**. Be careful with editing the system path. Avoid deleting any entries already on the list.



7. You should see a box with entries on the left. On the right, click **New**.

8. The system highlights a new line. Enter the path to the Spark folder **C:\Spark\spark-3.1.2-bin-hadoop3.2\bin**. We recommend using **%SPARK\_HOME%\bin**to avoid possible issues with the path.



9. Repeat this process 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 **C:\hadoop\bin**to the **Path variable**field, but we recommend using **%HADOOP\_HOME%\bin**.
* For Java, the variable name is **JAVA\_HOME** and for the value use the path to your Java JDK directory (in our case it’s **C:\Program Files\Java\jdk1.8.0\_251**).

10. Click **OK** to close all open windows.

**Note:** Star by restarting the Command Prompt to apply changes. If that doesn't work, you will need to reboot the system.

### Step 8: Launch Spark

1. Open a new command-prompt window using the right-click and **Run as administrator**:

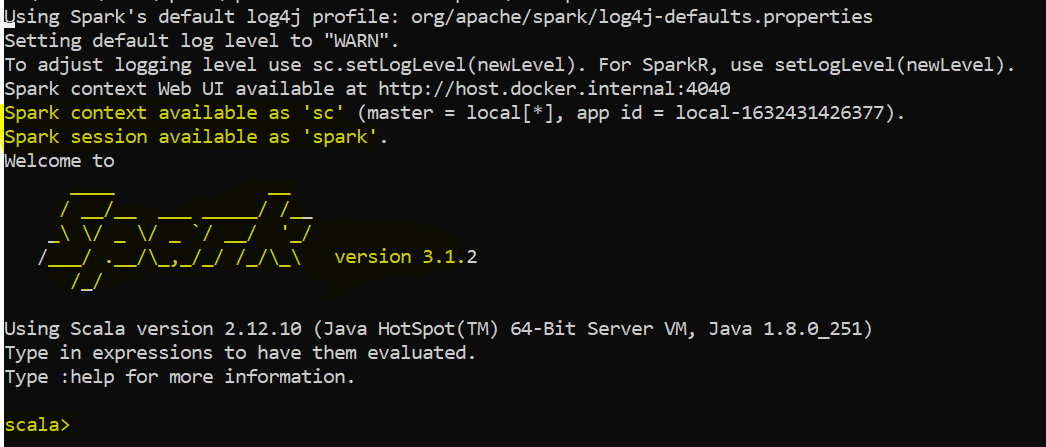
2. To start Spark, enter:

C:\Spark\spark-3.1.2-bin-hadoop3.2\bin\spark-shell

If you set the **environment path** correctly, you can type **spark-shell** to launch Spark.

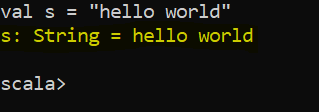
3. The system should display several lines indicating the status of the application. You may get a Java pop-up. Select **Allow access** to continue.

Finally, the Spark logo appears, and the prompt displays the **Scala shell**.



4. Run below command to check the execution which will return the string “Hello World”

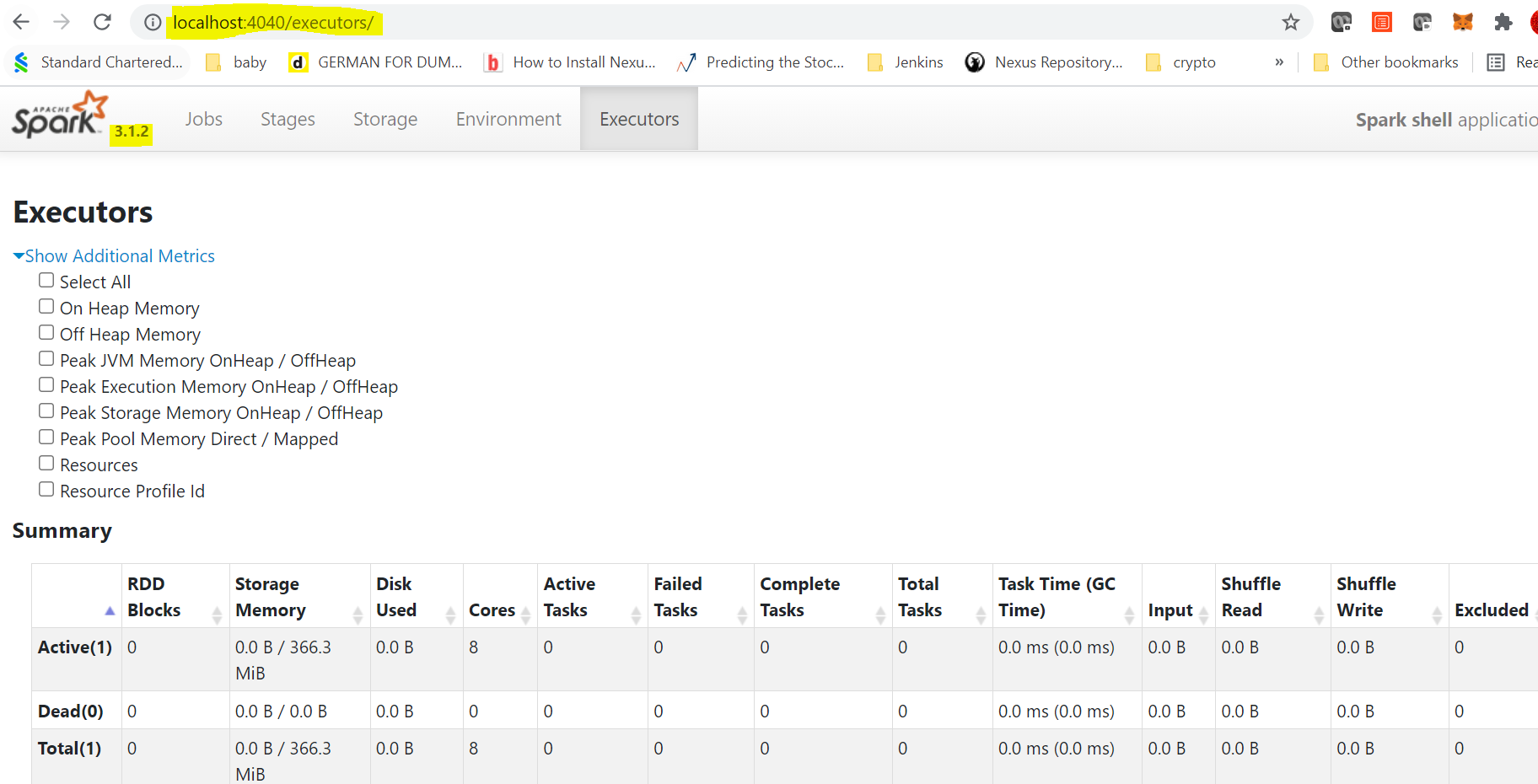
val s = "hello world"



5. Open a web browser and navigate to **http://localhost:4040/**.

6. You can replace **localhost**with the name of your system.

7. You should see an Apache Spark shell Web UI. The example below shows the Executors page.



8. To exit Spark and close the Scala shell, press **ctrl-d**in the command-prompt window.

Conclusion

That is all for installing Apache Spark on windows. We will see the real use cases of the Apache Spark in our demo session on 4th Oct, 2021.