



## BDI

**Name:** Kartik Jolapara

**Branch:** Computer Engineering

**SAP ID:** 60004200107

**Batch:** B1

### EXPERIMENT NO. 6 Installation and Configuration of Apache Spark

**AIM:** Install and Configure Apache Spark

#### **THEORY:**

Apache Spark is an open-source data processing framework for large volumes of data from multiple sources. Spark is used in distributed computing for processing machine learning applications, data analytics, and graph-parallel processing on single-node machines or clusters.

Owing to its lightning-fast processing speed, scalability, and programmability for Big Data, Spark has become one of the most widely used Big Data distributed processing frameworks for scalable computing.

Thousands of companies, including tech giants like Apple, Facebook, IBM, and Microsoft, use Apache Spark. Spark Installation is simple and can be done in a variety of ways. It provides native bindings for programming languages, including Java, Scala, Python, and R.

#### **Steps in Apache Spark Installation Step 1: Install Java 8**

Apache Spark requires Java 8. You can check to see if Java is installed using the command prompt.

Open the command line by clicking Start > type *cmd* > click Command Prompt. Type the following command in the command prompt:

`java -version`

If Java is installed, it will respond with the following output:

```
Microsoft Windows [Version 10.0.18362.778]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\Goran>java -version
java version "1.8.0_251"
Java(TM) SE Runtime Environment (build 1.8.0_251-b08)
Java HotSpot(TM) Client VM (build 25.251-b08, mixed mode, sharing)

C:\Users\Goran>
```

Your version may be different. The second digit is the Java version – in this case, Java 8. If you don't have Java installed:

1. Open a browser window, and navigate to <https://java.com/en/download/>.



## Java Download

Download Java for your desktop computer now!

**Version 8 Update 251**

Release date April 14, 2020



### Important Oracle Java License Update

**The Oracle Java License has changed for releases starting April 16, 2019.**

The new [Oracle Technology Network License Agreement for Oracle Java SE](#) is substantially different from prior Oracle Java licenses. The new license permits certain uses, such as personal use and development use, at no cost – but other uses authorized under prior Oracle Java licenses may no longer be available. Please review the terms carefully before downloading and using this product. An FAQ is available [here](#).

Commercial license and support is available with a low cost [Java SE Subscription](#).

Oracle also provides the latest OpenJDK release under the open source [GPL License](#) at [jdk.java.net](#).

  
**Java Download**

2. Click the Java Download button and save the file to a location of your choice.
3. Once the download finishes double-click the file to install Java.

### Step 2: Install Python

1. To install the Python package manager, navigate to <https://www.python.org/> in your web browser.
2. Mouse over the Download menu option and click Python 3.8.3. 3.8.3 is the latest version at the time of writing the article.
3. Once the download finishes, run the file.



Shri Vile Parle Kelavani Mandal's

**DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)




Downloads	Documentation	Community	Success Stories	News
All releases	<h3>Download for Windows</h3> <p>Python 3.8.3</p> <p><b>Note that Python 3.5+ cannot be used on Windows XP or earlier.</b></p> <p>Not the OS you are looking for? Python can be used on many operating systems and environments. View the full list of downloads.</p>			
Source code				
Windows				
Mac OS X				
Other Platforms				
License				
Alternative Implementations				

4. Near the bottom of the first setup dialog box, check off *Add Python 3.8 to PATH*. Leave the other box checked.
5. Next, click *Customize installation*.

Python 3.8.3 (32-bit) Setup

## Install Python 3.8.3 (32-bit)

Select **Install Now** to install Python with default settings, or choose **Customize** to enable or disable features.



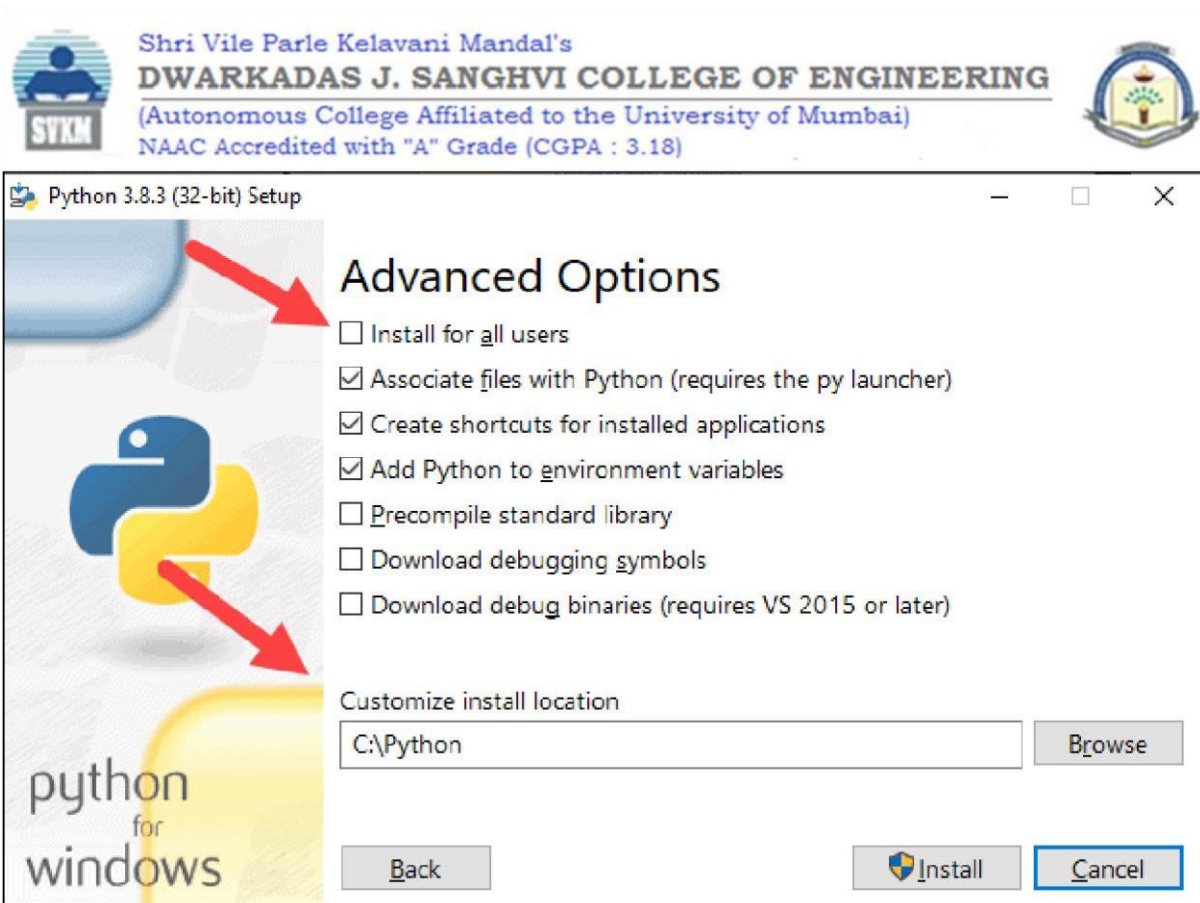
**Install Now**  
C:\Users\Goran\AppData\Local\Programs\Python\Python38-32  
Includes IDLE, pip and documentation  
Creates shortcuts and file associations

→ **Customize installation**  
Choose location and features

☒ Install launcher for all users (recommended)  
☐ Add Python 3.8 to PATH

Cancel

6. You can leave all boxes checked at this step, or you can uncheck the options you do not want.
7. Click **Next**.
8. Select the box **Install for all users** and leave other boxes as they are.
9. Under *Customize install location*, click **Browse** and navigate to the C drive. Add a new folder and name it *Python*.
10. Select that folder and click **OK**.



11. Click Install, and let the installation complete.
12. When the installation completes, click the *Disable path length limit* option at the bottom and then click Close.
13. If you have a command prompt open, restart it. Verify the installation by checking the version of Python: `python --version` The output should print Python 3.8.3.

### Step 3: Download Apache Spark



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 non-preview version.
  - In our case, in *Choose a Spark release* drop-down menu select 2.4.5 (Feb 05 2020).
  - In the second drop-down *Choose a package type*, leave the selection Pre-built for Apache Hadoop 2.7.
3. Click the *spark-2.4.5-bin-hadoop2.7.tgz* link.





Download Libraries ▾ Documentation ▾ Examples Community ▾ Developers ▾

## Download Apache Spark™

1. Choose a Spark release:  
2. Choose a package type:
3. Download Spark: [spark-2.4.5-bin-hadoop2.7.tgz](#) 
4. Verify this release using the 2.4.5 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.


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\username\Downloads\spark-2.4.5-bin-hadoop2.7.tgz SHA512
4. Change the username to your username. The system displays a long alphanumeric code, along with the message Certutil: -hashfile completed successfully.

```
Command Prompt
C:\Users\Goran>certutil -hashfile c:\users\Goran\Downloads\spark-2.4.5-bin-hadoop2.7.tgz SHA512
SHA512 hash of c:\users\Goran\Downloads\spark-2.4.5-bin-hadoop2.7.tgz:
2426a20c548bdfc07df288cd1d18d1da6b3189d0b78dee76fa034c52a4e02895f0ad460720c526f163ba63a17efae4764c
46a1cd8f9b04c60f9937a554db85d2
CertUtil: -hashfile command completed successfully.

C:\Users\Goran>
```



5. Compare the code to the one you opened in a new browser tab. If they match, your download file is uncorrupted.

### Step 5: Install Apache Spark

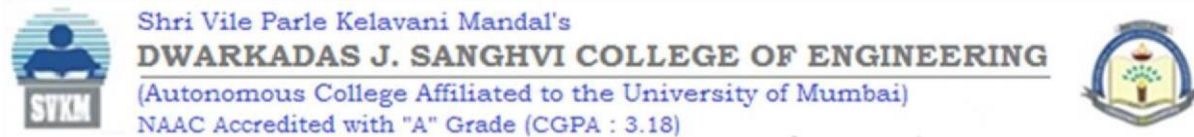
Installing Apache Spark involves extracting the downloaded file to the desired location.

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-2.4.5-bin-hadoop2.7` 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.

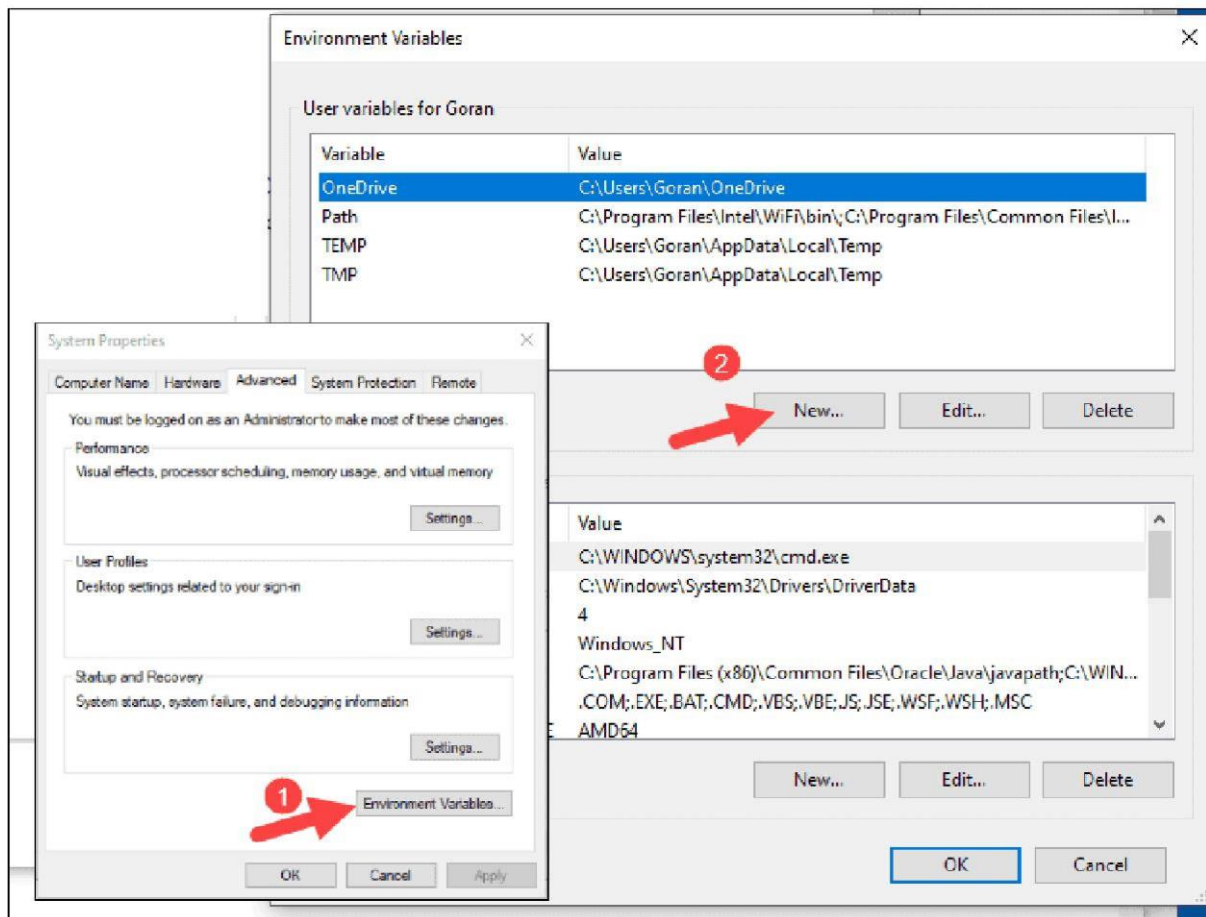
 <a href="#">mapred</a>	some binaries from 273 to 311
 <a href="#">mapred.cmd</a>	some binaries from 273 to 311
 <a href="#">rcc</a>	some binaries from 273 to 311
 <a href="#">winutils.exe</a>	fixed exe and lib 265-312
 <a href="#">winutils.pdb</a>	fixed exe and lib 265-312
 <a href="#">yarn</a>	some binaries from 273 to 311
 <a href="#">yarn.cmd</a>	some binaries from 273 to 311

2. Find the Download button on the right side to download the file.
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`.

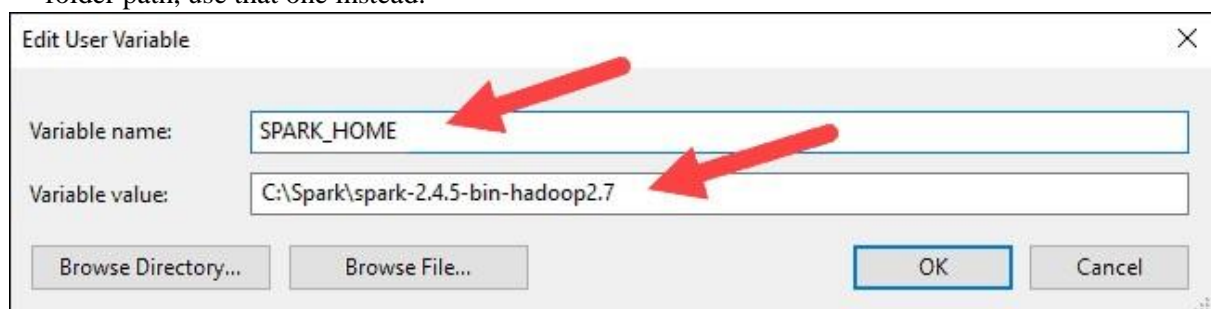
### Step 7: Configure Environment Variables

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-2.4.5-bin-hadoop2.7** 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.



Environment Variables

User variables for Goran

Variable	Value
HADOOP_HOME	C:\hadoop
JAVA_HOME	C:\Java\jre1.8.0_251
OneDrive	C:\Users\Goran\OneDrive
Path	C:\Python\Scripts\;C:\Python\;C:\Program Files\Intel\WiFi\bin\;C:\...
SPARK_HOME	C:\Spark\spark-2.4.5-bin-hadoop2
TEMP	C:\Users\Goran\AppData\Local\Temp
TMP	C:\Users\Goran\AppData\Local\Temp

New... Edit... Delete

System variables

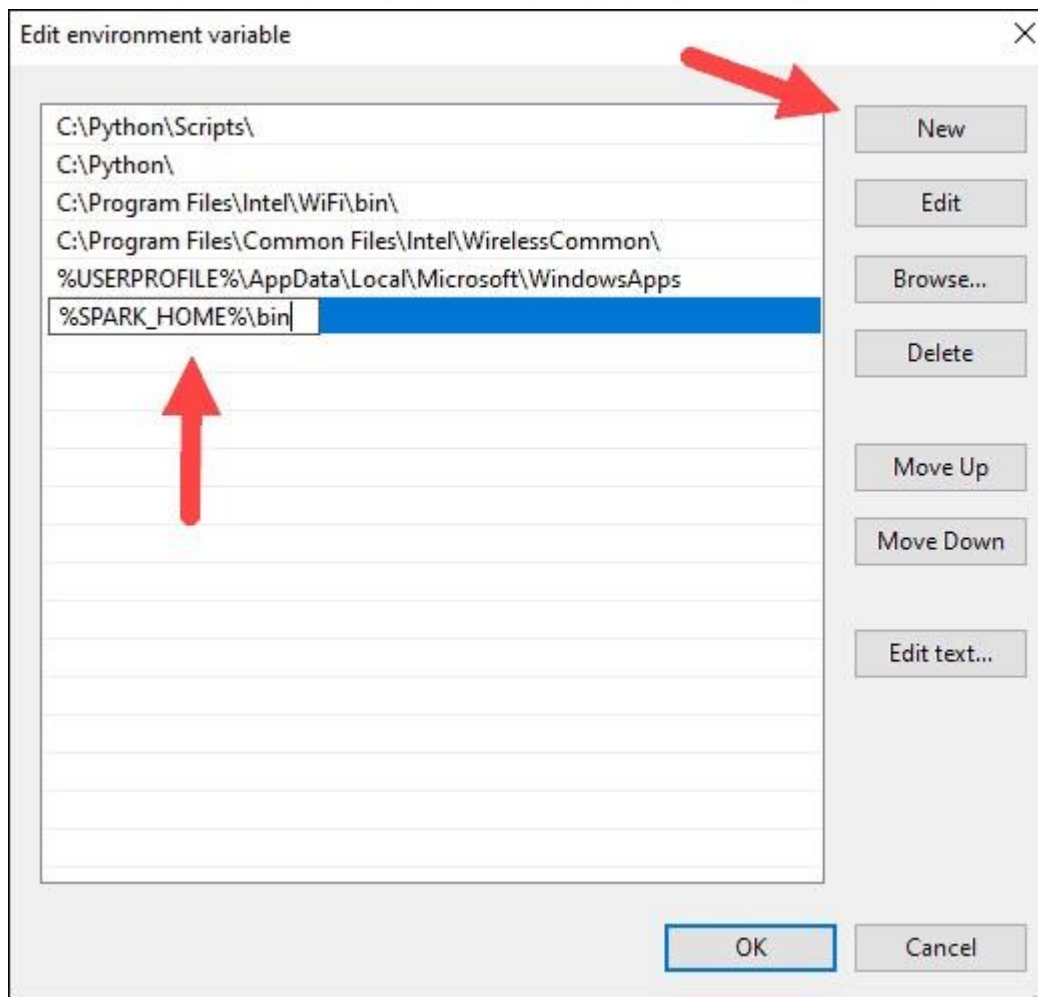
Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	4
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\WIN...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

New... Edit... Delete

OK Cancel

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-2.4.5-bin-hadoop2.7\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: Start 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-2.4.5-bin-hadoop2.7\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., Open a web browser and navigate to `http://localhost:4040/`.
5. You can replace `localhost` with the name of your system.
6. You should see an Apache Spark shell Web UI. The example below shows the *Executors* page.

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

**CONCLUSION:** We have successfully installed and configured Apache Spark in Windows.