



# How To Install Apache Spark On Windows 10

Posted May 28, 2020

WINDOWS

BIG DATA

[Home](#) / [Web Servers](#) / [How to Install Apache Spark on Windows 10](#)

☰ Contents



## 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.

This guide will show you **how to install Apache Spark on Windows 10** and test the installation.



## Prerequisites

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

## Install Apache Spark on Windows

This is a vertical advertisement for Intel PhoenixNAP FlexServers. At the top, it features the Intel logo and the PhoenixNAP logo. Below the logos, the text 'FLEXSERVERS' is written in red, followed by 'PUT THE PEDAL TO BARE METAL' in large, bold, white capital letters. Underneath this, a red horizontal line is followed by the text 'Vertical CPU scaling. Advanced performance. Excellent price.' in white. The bottom half of the ad features a glowing blue and red speedometer graphic with digital data overlays. At the very bottom, there is a red rounded rectangular button with the white text 'ORDER NOW'.

Installing Apache Spark on Windows 10 may seem complicated to novice users, but this simple tutorial will have you up and running. If you already have Java 8 and Python 3 installed, you can skip the first two steps.

## Step 1: Install Java 8

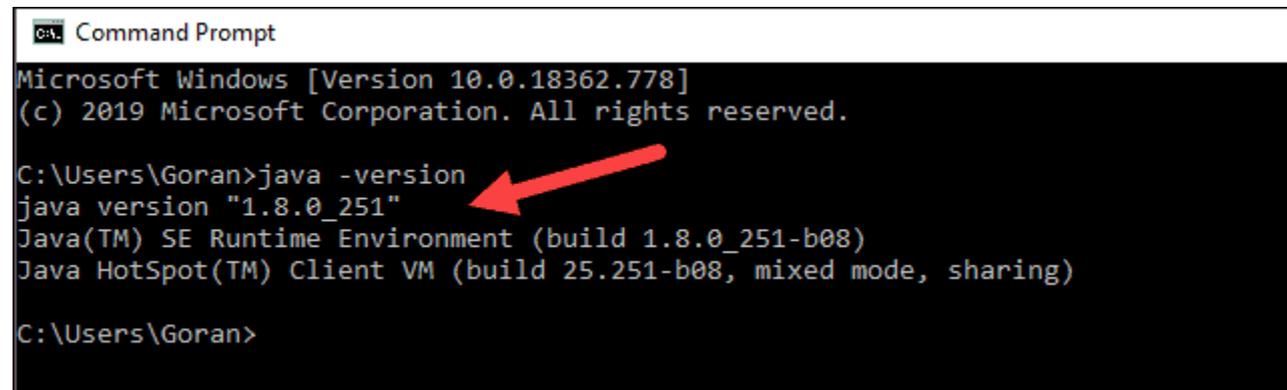
Apache Spark requires Java 8. You can check to see if Java is installed using the command prompt.

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:



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

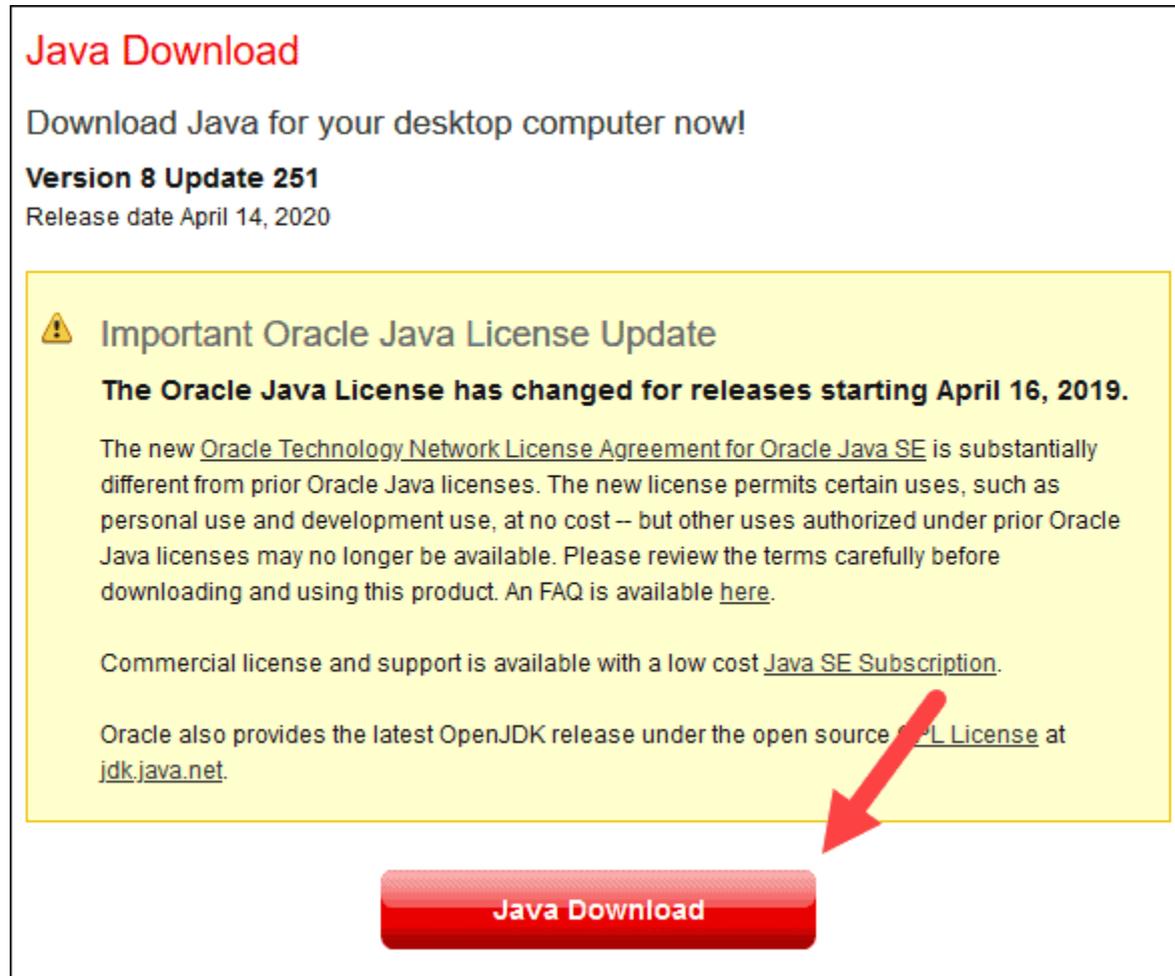
C:\Users\Goran>java -version
java version "1.8.0_251"
Java(TM) SE Runtime Environment (build 1.8.0_251-b08)
Java HotSpot(TM) Client VM (build 25.251-b08, mixed mode, sharing)

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](http://jdk.java.net).

**Java Download**

A red arrow points from the bottom right of the license update box to the 'Java Download' button.

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.

---

**Note:** At the time this article was written, the latest Java version is 1.8.0\_251. Installing a later



version will still work. This process only needs the Java Runtime Environment (JRE) – the full Development Kit (JDK) is not required. The download link to JDK is <https://www.oracle.com/java/technologies/javase-downloads.html>.

---

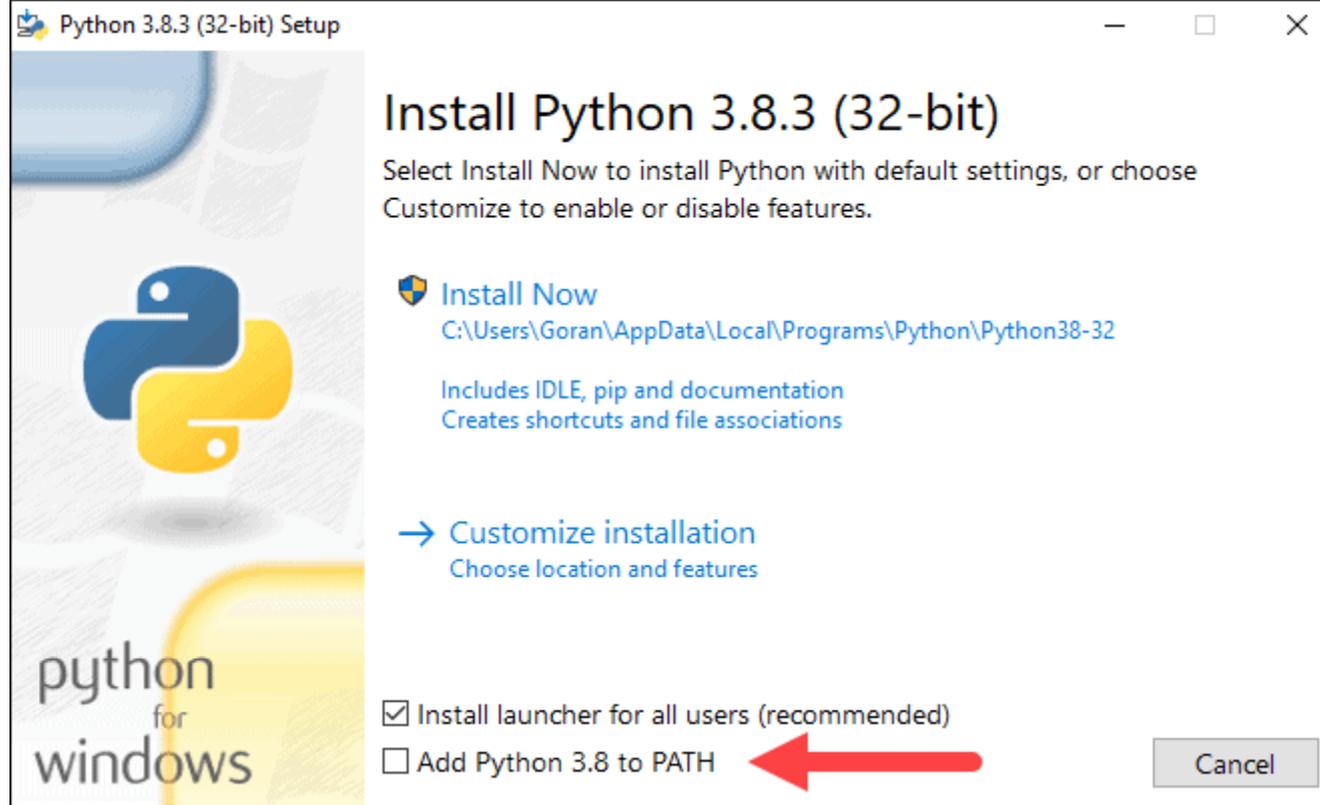
## Step 2: Install Python

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.



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**.



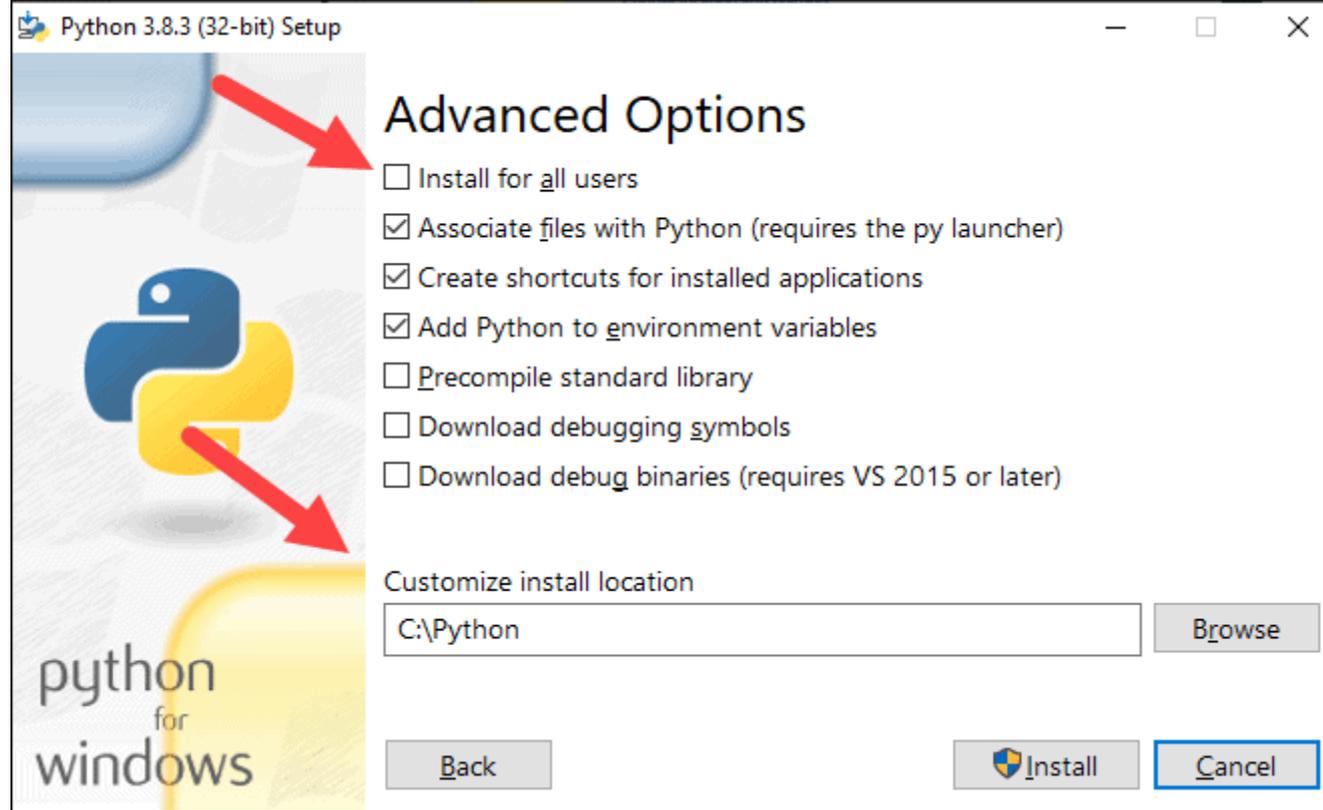
6. You can leave all boxes checked at this step, or you can uncheck the options you do not want.

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**.

---

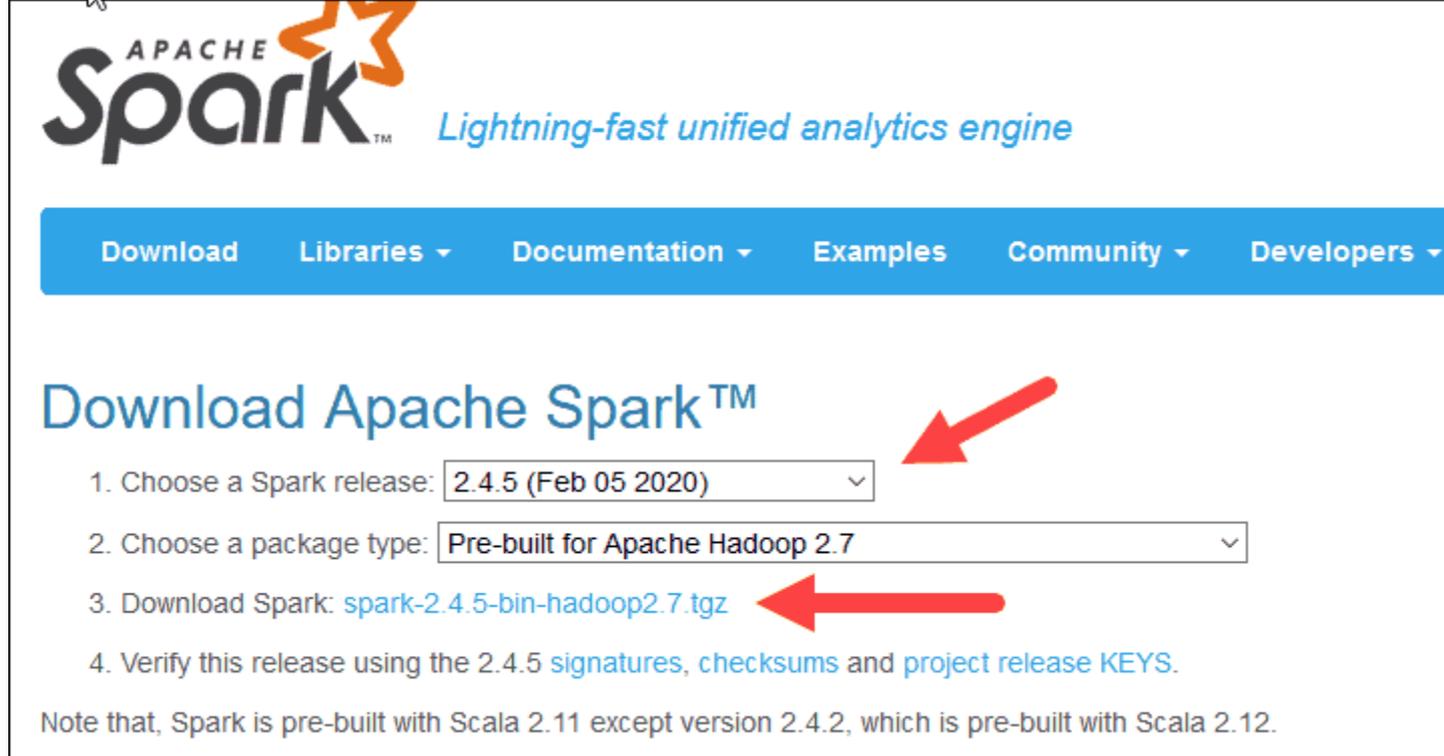


**Note:** For detailed instructions on how to install Python 3 on Windows or how to troubleshoot potential issues, refer to our [Install Python 3 on Windows](#) guide.

---

### 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.



APACHE  
**Spark**™ Lightning-fast unified analytics engine

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.

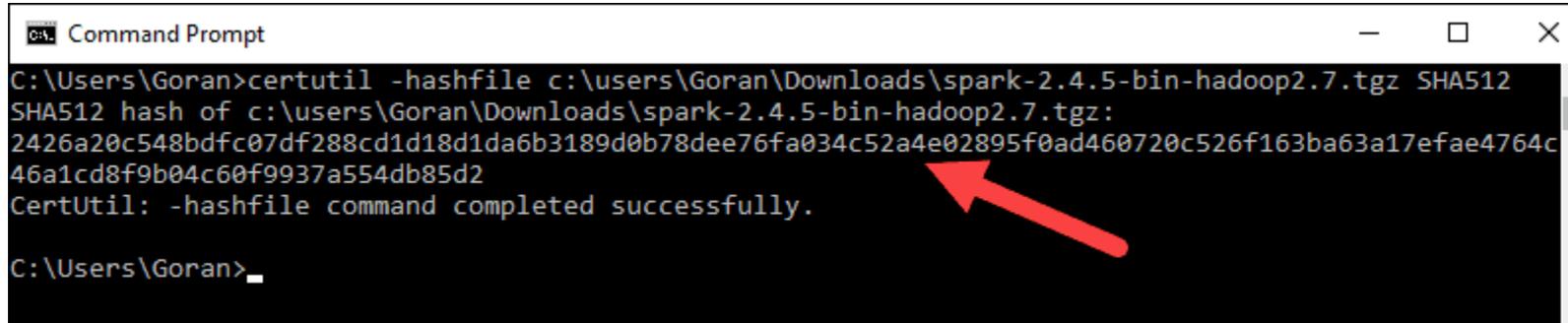
4. A page with a list of mirrors loads where you can see different servers to download from. Pick any from the list and save the file to your Downloads folder.

### 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.

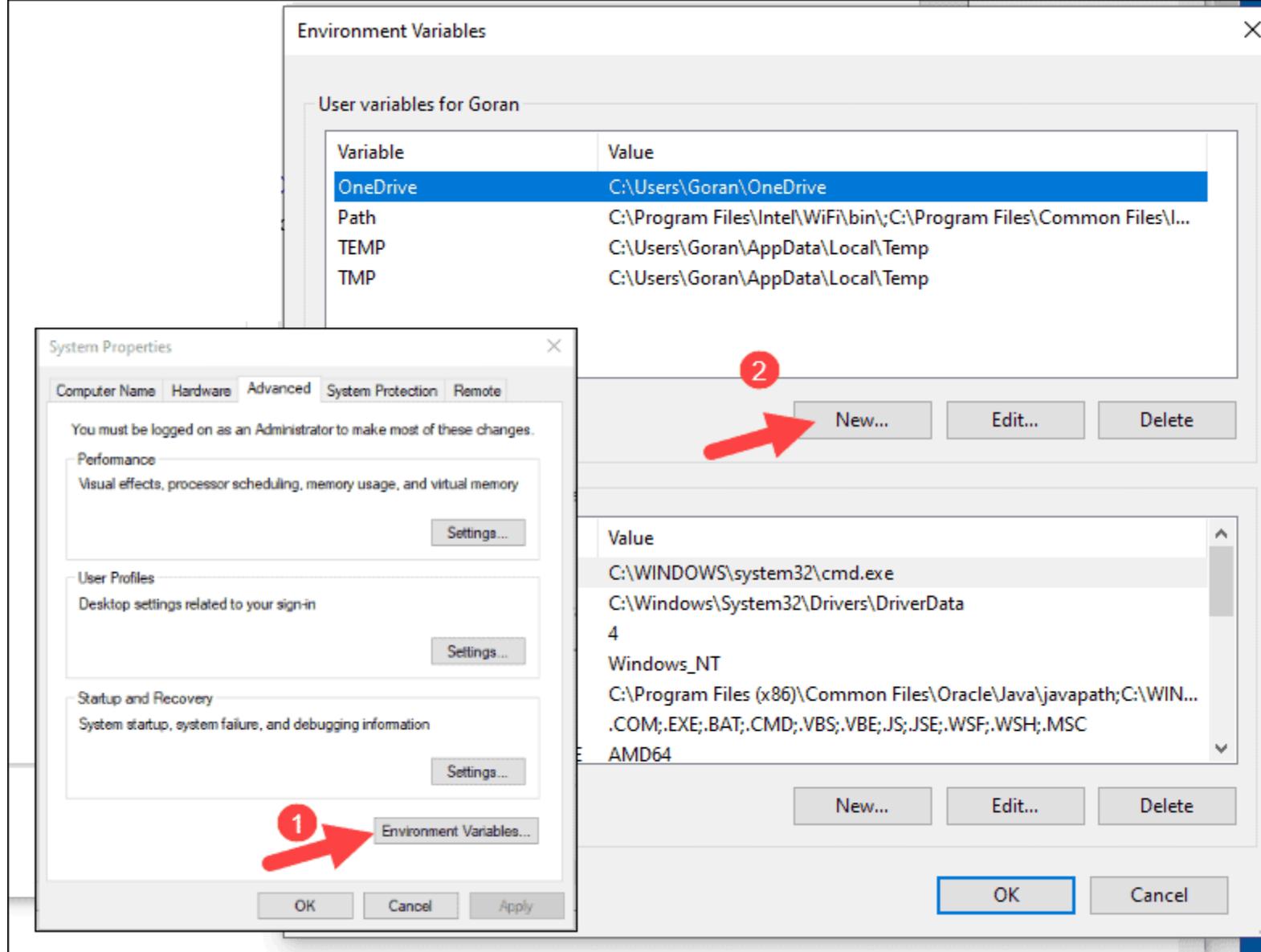
3. Now, create new folders **Hadoop** and **bin** on C: using Windows Explorer or the Command Prompt.

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

## Step 7: Configure Environment Variables

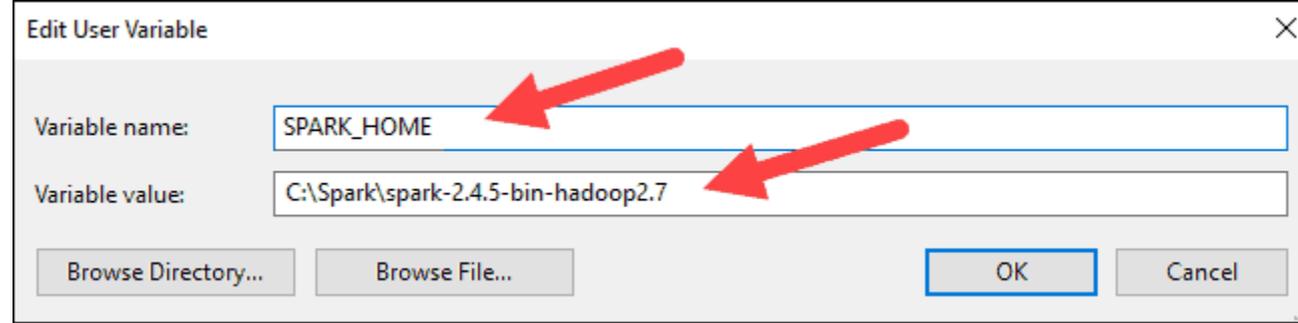
This step adds the Spark and Hadoop locations to your system PATH. It allows you to run the Spark shell directly from a command prompt window.

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.

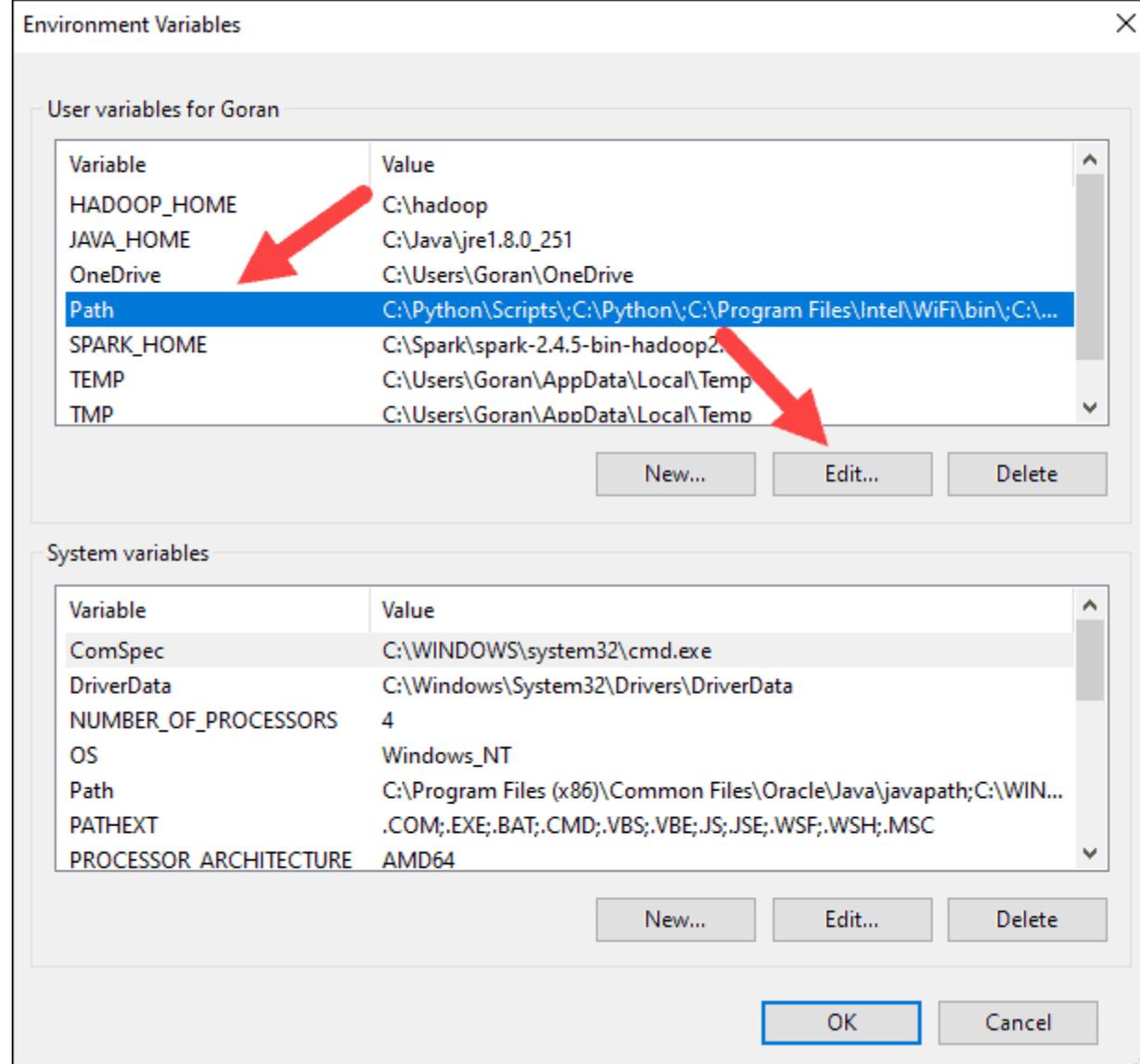


4. For *Variable Name* type **SPARK\_HOME**.

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.

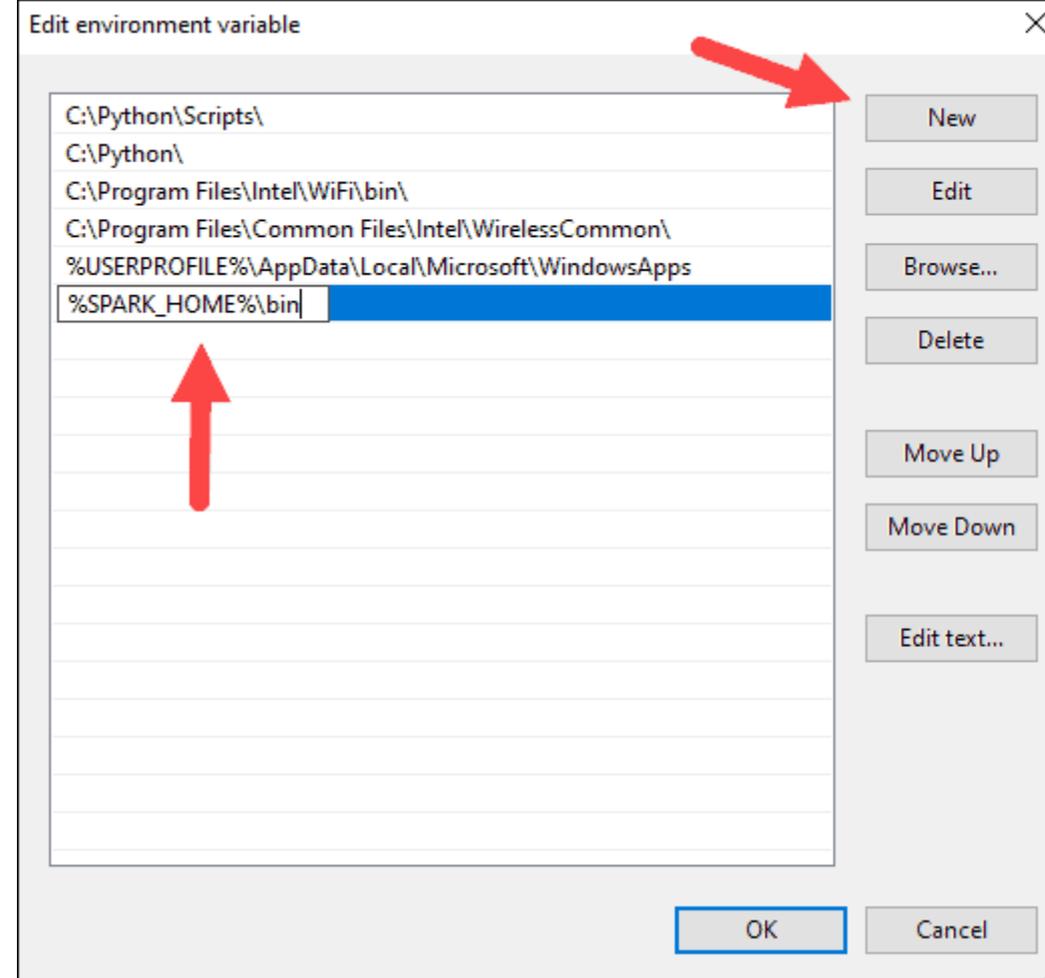


6. In the top box, click the **Path** entry, then click **Edit**. Be careful with editing the system path. Avoid deleting any entries already on the list.



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**.



[Jobs](#)
[Stages](#)
[Storage](#)
[Environment](#)
[Executors](#)
Spark shell application U

## Executors

[Show Additional Metrics](#)

### Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	To
Active(1)	0	0.0 B / 434 MB	0.0 B	4	0	0	0	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0
Total(1)	0	0.0 B / 434 MB	0.0 B	4	0	0	0	0

### Executors

Show  entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Com
driver	DESKTOP-SFBGHOU:61547	Active	0	0.0 B / 434 MB	0.0 B	4	0	0	0

Showing 1 to 1 of 1 entries [Previous](#) [1](#) [Next](#)

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

---

**Note:** If you installed Python, you can run Spark using Python with this command:



pyspark

Exit using `quit()` .

---

## Test Spark

In this example, we will launch the Spark shell and use Scala to read the contents of a file. You can use an existing file, such as the *README* file in the Spark directory, or you can create your own. We created *pnaptest* with some text.

1. Open a command-prompt window and navigate to the folder with the file you want to use and launch the Spark shell.
2. First, state a variable to use in the Spark context with the name of the file. Remember to add the file extension if there is any.

```
val x =sc.textFile("pnaptest")
```

3. The output shows an RDD is created. Then, we can view the file contents by using this command to call an action:

```
x.take(11).foreach(println)
```

```
Administrator: Command Prompt - spark-shell
Welcome to
  ____
 /  _ \
/_  /  \
 \_/    \
  version 2.4.5

Using Scala version 2.11.12 (Java HotSpot(TM) Client VM, Java 1.8.0_251)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val x = sc.textFile("pnaptest")
20/05/17 23:04:46 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming
yes
x: org.apache.spark.rdd.RDD[String] = pnaptest MapPartitionsRDD[1] at textFile at <console>:24

scala> x.take(11).foreach(println)
Apache Spark Environment Variables for Windows 10

Variable Name: SPARK_HOME
Variable Value: C:\Spark\spark-2.4.5-bin-hadoop2.7
Path: %SPARK_HOME%\bin

Variable Name: HADOOP_HOME
Variable Value: C:\hadoop
Path: %HADOOP_HOME%\bin

phoenixnap.com

scala>
```

This command instructs Spark to print 11 lines from the file you specified. To perform an action on this file (**value x**), add another value **y**, and do a map transformation.

4. For example, you can print the characters in reverse with this command:

```
val y = x.map(_.reverse)
```

5. The system creates a child RDD in relation to the first one. Then, specify how many lines you want to print from the value **y**:

```
y.take(11).foreach(println)
```

```
scala> y.take(11).foreach(println)
01 swodniW rof selbairaV tnemnorivnE krapS ehcapA

EMOH_KRAPs :emaN elbairaV
7.2poodah-nib-5.4.2-krapS\krapS\C :eulaV elbairaV
nib\%EMOH_KRAPs% :htaP

EMOH_POODAH :emaN elbairaV
poodah\C :eulaV elbairaV
nib\%EMOH_POODAH% :htaP

moc.panxineohp
```

The output prints 11 lines of the *pnptest* file in the reverse order.

When done, exit the shell using **ctrl-d**.

## Conclusion

You should now have a working installation of Apache Spark on Windows 10 with all dependencies installed. Get started running an instance of Spark in your Windows environment.



## Next you should also read

✉ DevOps And Development,

Databases

### How to Install Elasticsearch, Logstash, and Kibana (ELK Stack) on CentOS 8

May 6, 2020

---

Need to install the ELK stack to manage server log files on your CentOS 8? Follow this step-by-step guide and...

[READ MORE](#)

✉ DevOps And Development,

Databases

### Cassandra vs MongoDB – What are the Differences?

April 27, 2020

---

Learn about the difference between Cassandra and MongoDB. These NoSQL databases have some similarities, but...

[READ MORE](#)

✉ DevOps And Development,

Databases

### How to Install Elasticsearch on Ubuntu 18.04

April 23, 2020

---

Elasticsearch is an open-source engine that enhances searching, storing and analyzing capabilities of your...

[READ MORE](#)

✉ DevOps And Development,

Databases

### How to Install Spark on Ubuntu

April 13, 2020

---

This Spark tutorial shows how to get started with Spark. The guide covers the procedure for installing Java,...

[READ MORE](#)

# Author

---



## Goran Jevtic

Goran combines his passions for research, writing and technology as a technical writer at phoenixNAP. Working with multiple departments and on a variety of projects, he has developed extraordinary understanding of cloud and virtualization technology trends and best practices.

---

### RECENT POSTS

[Ln Command: How to Create Symbolic Links in Linux](#)

[How to Use DD Show Progress Command in Linux?](#)

[How to Create Partitions in Linux](#)

### CATEGORIES

[SYSADMIN](#)

[VIRTUALIZATION](#)

[DEVOPS AND DEVELOPMENT](#)

[SECURITY](#)

[BACKUP AND RECOVERY](#)

[BARE METAL SERVERS](#)

[WEB SERVERS](#)

[NETWORKING](#)

[DATABASES](#)

### COMPANY

[ABOUT US](#)

[GITHUB](#)

[BLOG](#)

[RFP TEMPLATE](#)

[CAREERS](#)

### PRODUCTS

[COLOCATION](#)

[SERVERS](#)

[CLOUD SERVICES](#)

[SOLUTIONS](#)

[LOCATIONS](#)

### CONNECT

[EVENTS](#)

[PRESS](#)

[CONTACT US](#)

[Live Chat](#) [Get a Quote](#) [Support | 1-855-330-1509](#) [Sales | 1-877-588-5918](#)

