



Roberto Diniz

Olivia Smith
August 29th, 2020

PYTHON +1

Installation of Pyspark (All operating systems)

This tutorial will demonstrate the installation of Pyspark and how to manage the environment variables in Windows, Linux, and Mac Operating System.

Pyspark = Python + Apache Spark

Apache Spark is a new and open-source framework used in the big data industry for real-time processing and batch processing. It supports different languages, like Python, Scala, Java, and R.

Apache Spark is initially written in a Java Virtual Machine(JVM) language called Scala, whereas Pyspark is like a Python API which contains a library called Py4J. This allows dynamic interaction with JVM objects.



EXPLORE DATACAMP'S PYTHON
COURSE LIBRARY

Explore Now

Windows Installation

The installation which is going to be shown is for the Windows Operating System. It consists of the installation of Java with the environment variable and Apache Spark with the environment variable.

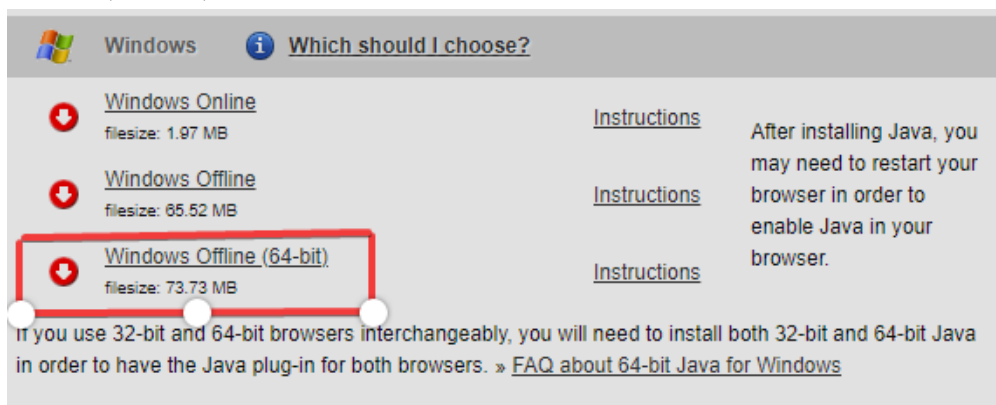
The recommended pre-requisite installation is Python, which is done from [here](#).



1. Go to [Download Java JDK](#).

Visit Oracle's website for the download of the Java Development Kit(JDK).

2. Move to download section consisting of operating system Windows, and in my case, it's Windows Offline(64-bit). The installer file will be downloaded.



3. Open the installer file, and the download begins.

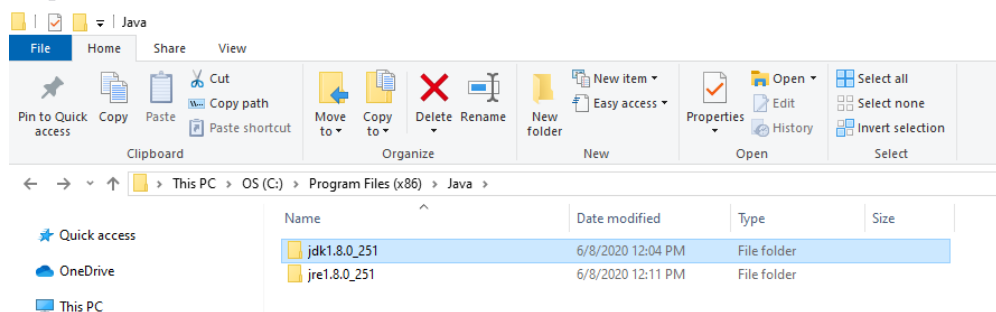


4. Go to "Command Prompt" and type "java -version" to know the version and know whether it is installed or not.

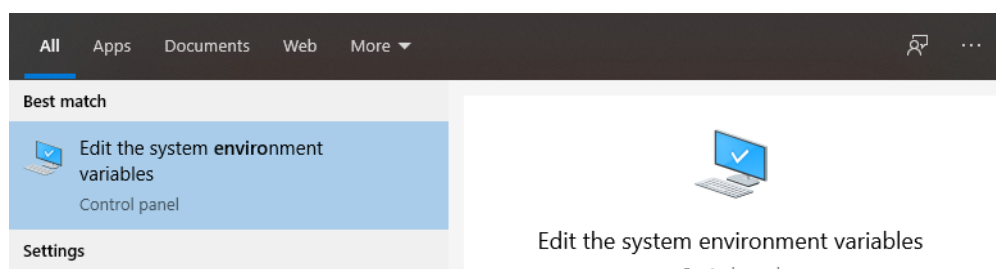


```
(C) 2019 Microsoft Corporation. All Rights Reserved.  
  
C:\Users\Dell>java -version  
java version "1.8.0_251"  
Java(TM) SE Runtime Environment (build 1.8.0_251-b08)  
Java HotSpot(TM) Client VM (build 25.251-b08, mixed mode)
```

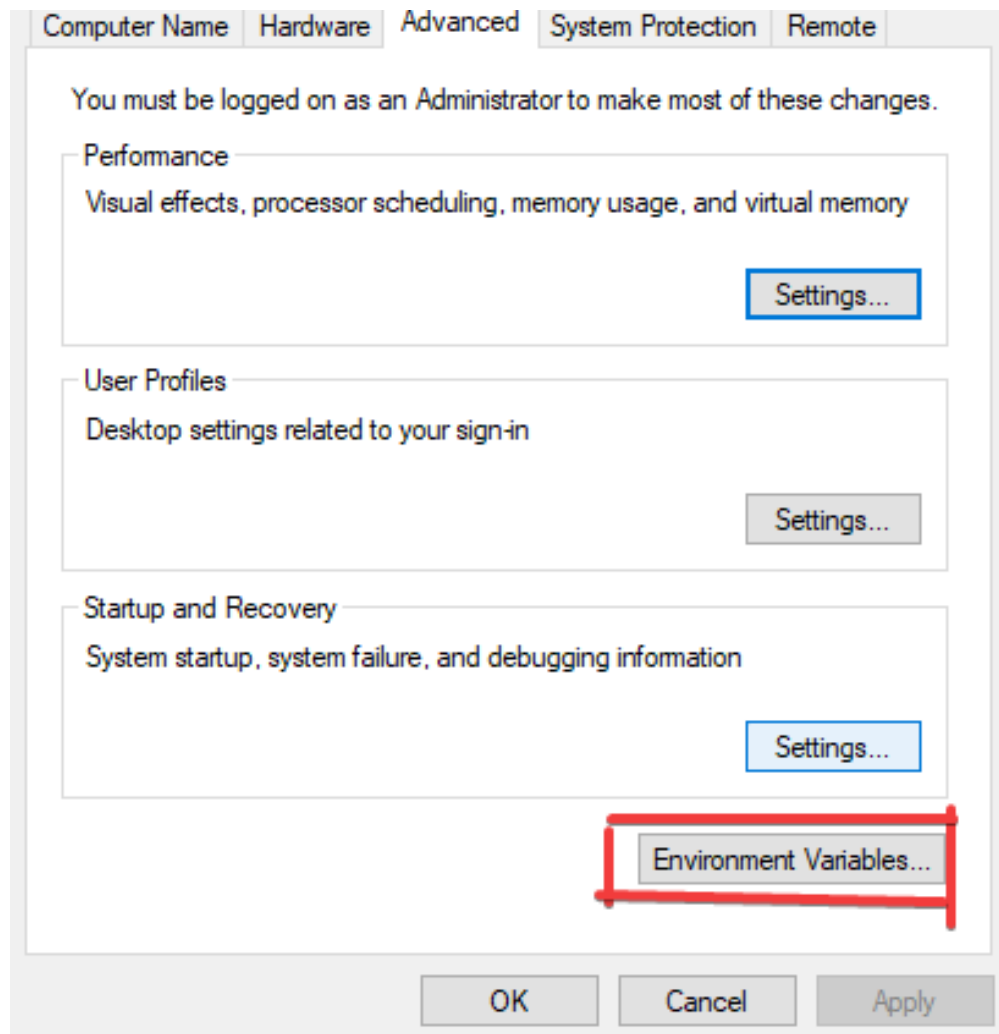
5. Add the Java path



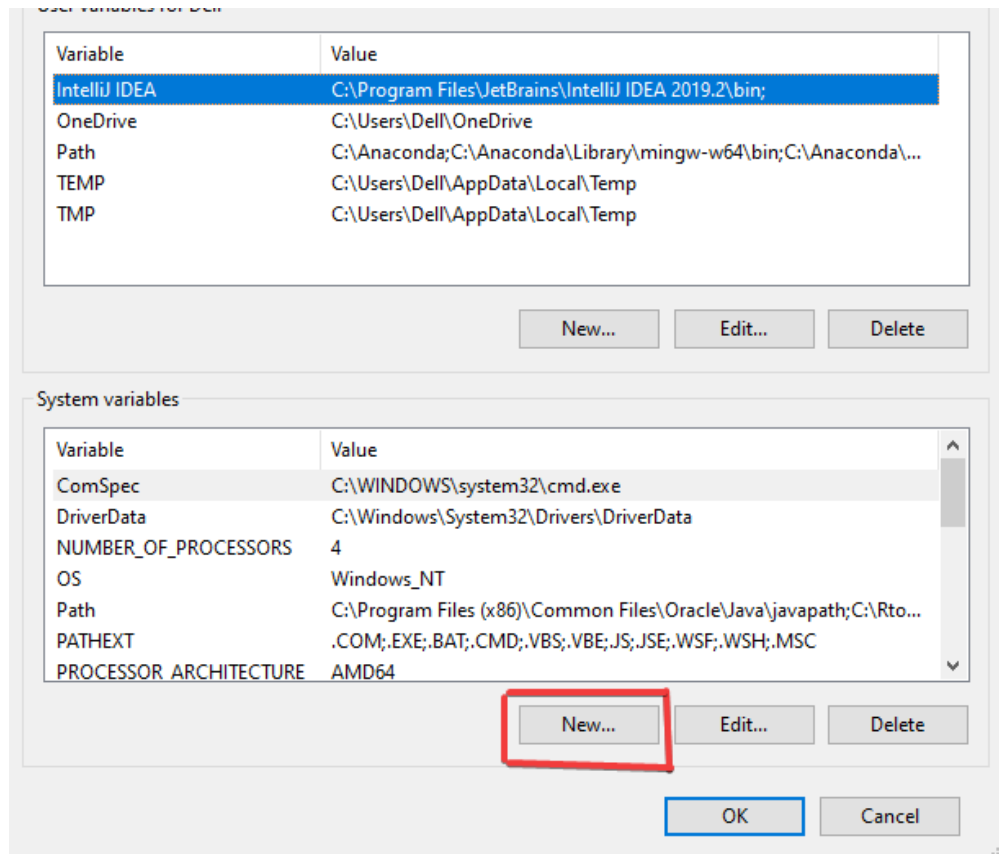
6. Go to the search bar and "EDIT THE ENVIRONMENT VARIABLES."



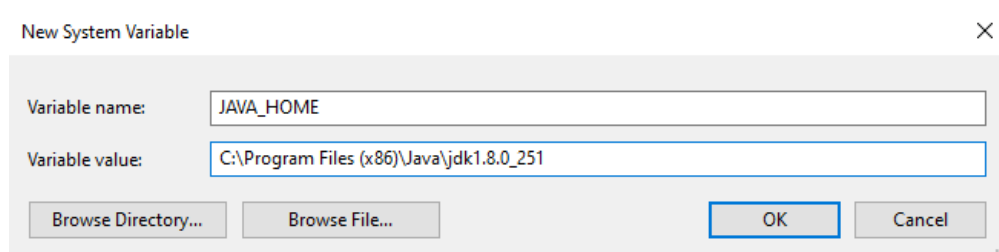
7. Click into the "Environment Variables"



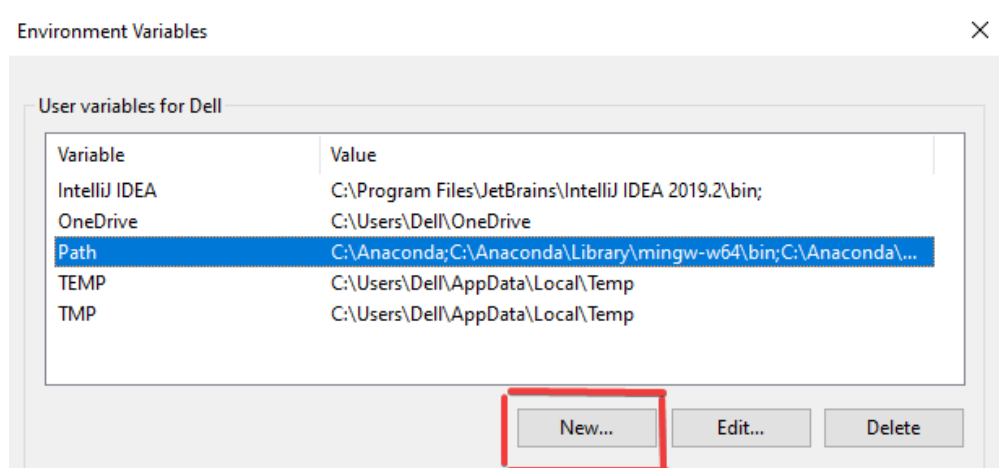
8. Click into "New" to create your new Environment variable.



9. Use Variable Name as 'JAVA_HOME' and your Variable Value as 'C:\Program Files (x86)\Java\jdk1.8.0_251'. This is your location of the Java file. Click 'OK' after you've finished the process.

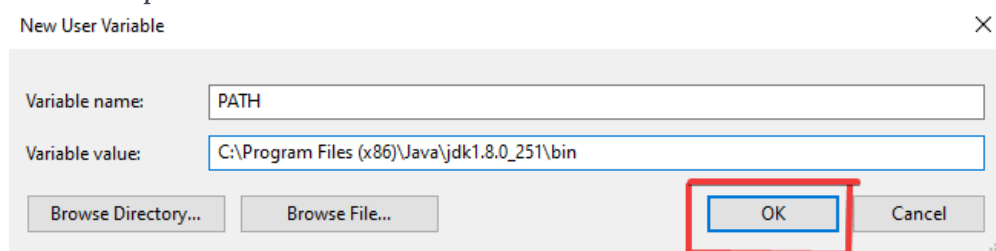


10. Let's add the User variable and select 'Path' and click 'New' to create it.

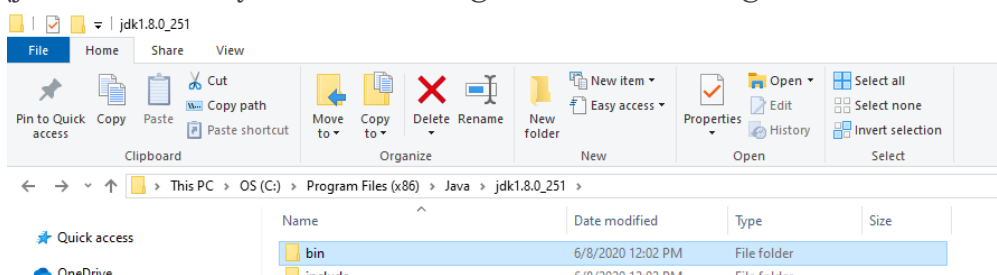




you've finished the process.

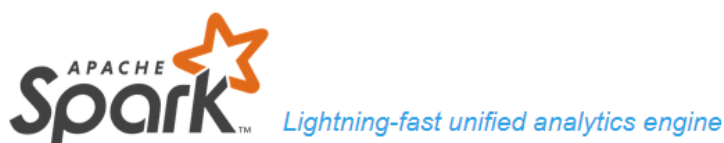


Note: You can locate your Java file by going to C drive, which is C:\Program Files (x86)\Java\jdk1.8.0_251' if you've not changed location during the download.



Installing Pyspark

1. Head over to the [Spark homepage](#).
2. Select the Spark release and package type as following and download the .tgz file.



[Download](#) [Libraries](#) [Documentation](#) [Examples](#) [Community](#) [Developers](#)

Download Apache Spark™

1. Choose a Spark release: [2.4.6 \(Jun 05 2020\)](#)
2. Choose a package type: [Pre-built for Apache Hadoop 2.7](#)
3. Download Spark: [spark-2.4.6-bin-hadoop2.7.tgz](#)
4. Verify this release using the 2.4.6 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.



COMMUNITY-LED DEVELOPMENT "THE APACH

Projects ▾

People ▾

Community ▾

License ▾

We suggest the following mirror site for your download:

<https://downloads.apache.org/spark/spark-2.4.6/spark-2.4.6-bin-hadoop2.7.tgz>

You can make a new folder called 'spark' in the C directory and extract the given file by using 'Winrar', which will be helpful afterward.

Download and setup winutils.exe

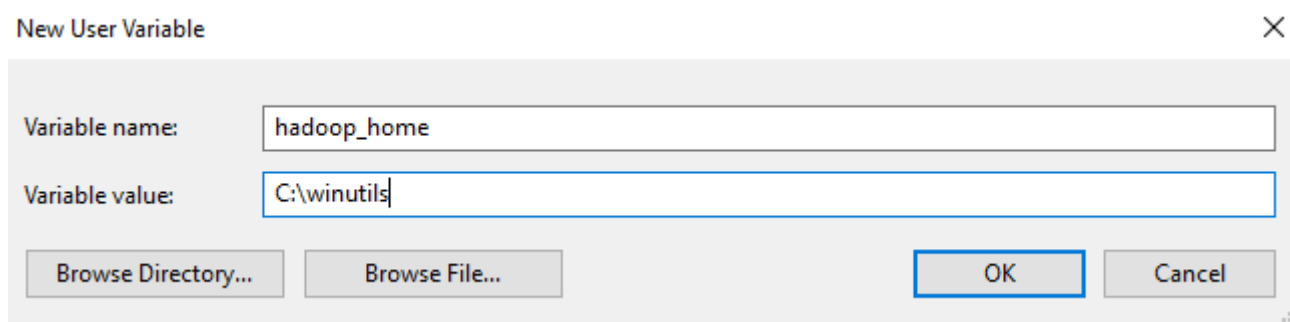
Go to [Winutils](#) choose your previously downloaded Hadoop version, then download the winutils.exe file by going inside 'bin'. The link to my Hadoop version is:

<https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>

Make a new folder called 'winutils' and inside of it create again a new folder called 'bin'. Then put the file recently download 'winutils' inside it.

Environment variables

1. Let's create a new environment where variable name as "hadoop_home" and variable value to be the location of winutils, which is "C:\winutils" and click "OK".



2. For spark, also let's create a new environment where the variable name is "Spark_home" and the variable value to be the location of spark, which is "C:\spark" and click "OK".



Variable name:

Variable value:

3. Finally, double click the 'path' and change the following as done below where a new path is created "%Spark_Home%\bin" is added and click "OK".

Edit environment variable

C:\Anaconda
C:\Anaconda\Library\mingw-w64\bin
C:\Anaconda\Library\usr\bin
C:\Anaconda\Library\bin
C:\Anaconda\Scripts
C:\Users\Dell\AppData\Local\Programs\Python\Python37-32\Scripts\
C:\Users\Dell\AppData\Local\Programs\Python\Python37-32\
C:\Program Files\MySQL\MySQL Shell 8.0\bin\
%IntelliJ IDEA%
C:\Users\Dell\AppData\Roaming\npm
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Program Files (x86)\heroku\bin
%Spark_Home%\bin

Finalizing Pyspark Installation

1. Open Command Prompt and type the following command.

```
Microsoft Windows [Version 10.0.18362.900]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\Dell>pyspark
```

2. Once everything is successfully done, the following message is obtained.



```
Setting default log level to 'WARN'.
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      _/ _ \| | | | _/_/
     / _ \| | | | |/_/
    / ___ \| | | | |/_/
   /___\ \| | | | |/_/
  /___\ \| | | | |/_/
 /___\ \| | | | |/_/
/_/___\ \| | | | |/_/

version 2.4.6

Using Python version 3.7.4 (default, Aug  9 2019 18:34:13)
SparkSession available as 'spark'.
>>> _
```

Linux Installation

The installation which is going to be shown is for the **Linux** Operating System. It consists of the installation of Java with the environment variable along with Apache Spark and the environment variable.





The recommended pre-requisite installation is Python, which is done from [here](#).

Java Installation

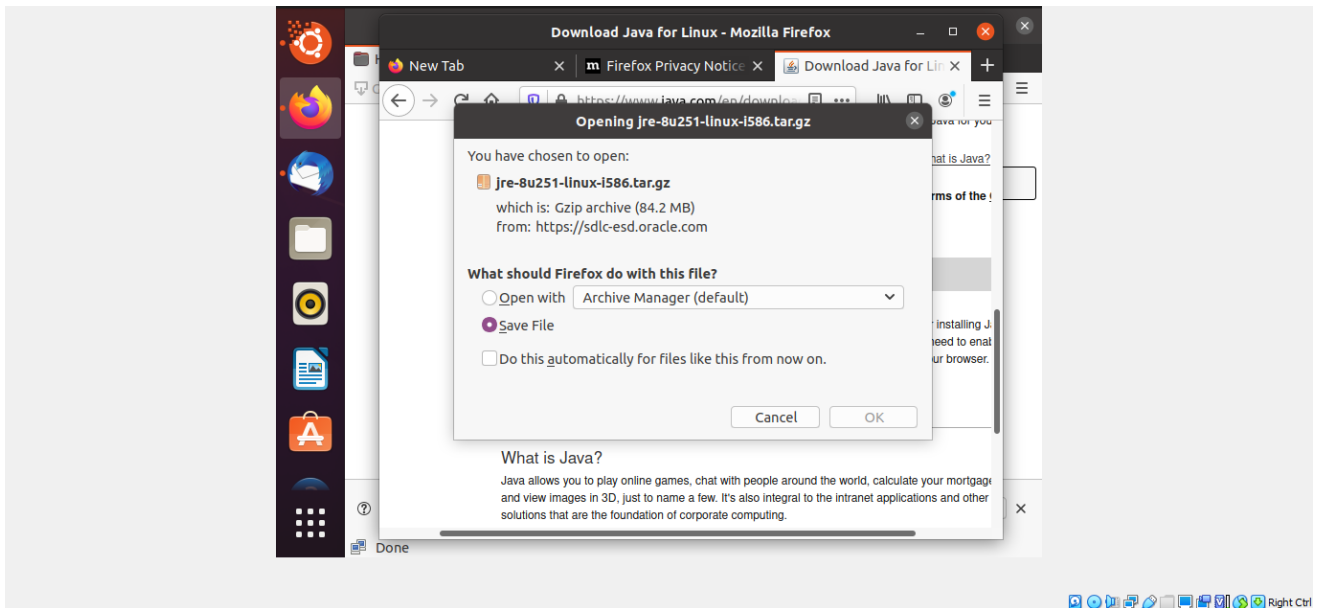
1. Go to [Download Java JDK](#).

Visit Oracle's website for the download of the Java Development Kit (JDK).

2. Move to the download section consisting of the operating system **Linux** and download it according to your system requirement.

Linux		
	Linux RPM filesize: 68.41 MB	Instructions
	Linux filesize: 84.22 MB	Instructions
	Linux x64 filesize: 83.49 MB	Instructions
	Linux x64 RPM filesize: 67.6 MB	Instructions

3. Save the file and click "Ok" to save in your local machine.



4. Go to your terminal and check the recently downloaded file using 'ls' command.

```
olivia@olivia-VirtualBox:~$ cd Downloads
olivia@olivia-VirtualBox:~/Downloads$ ls
jdk-11.0.7_linux-x64_bin.deb
```

5. Install the package using the following command, which will install the debian package of java, which is recently downloaded.

```
olivia@olivia-VirtualBox:~/Downloads$ ls
olivia@olivia-VirtualBox:~/Downloads$ sudo dpkg -i jdk-11.0.7_linux-x64_bin.deb
[sudo] password for olivia:
```

6. Finally, you can check your java version using 'java --version' command.

```
olivia@olivia-VirtualBox:~/Downloads$ java --version
java 11.0.7 2020-04-14 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.7+8-LTS)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.7+8-LTS, mixed mode)
```

7. For configuring environment variables, let's open the 'gedit' text editor using the following command.

```
olivia@olivia-VirtualBox:~/Downloads$ sudo gedit /etc/environment
```

8. Let's make the change by providing the following information where the 'Java' path is specified.

```
2 JAVA_HOME="/usr/lib/jvm/jdk-11.0.7"
```

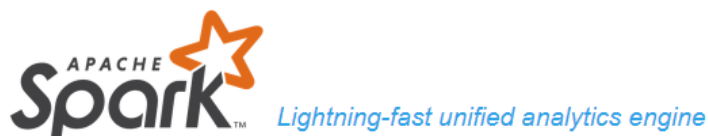


9. To make a final change, let's type the following command.

```
olivia@olivia-VirtualBox:~/Downloads$ source /etc/environment
```

Installing Spark

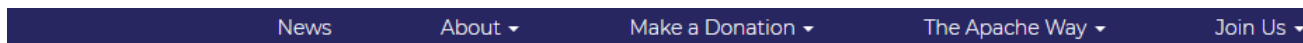
1. Head over to the [Spark homepage](#).
2. Select the Spark release and package type as following and download the .tgz file.



Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.4.6-bin-hadoop2.7.tgz](#)
4. Verify this release using the 2.4.6 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.



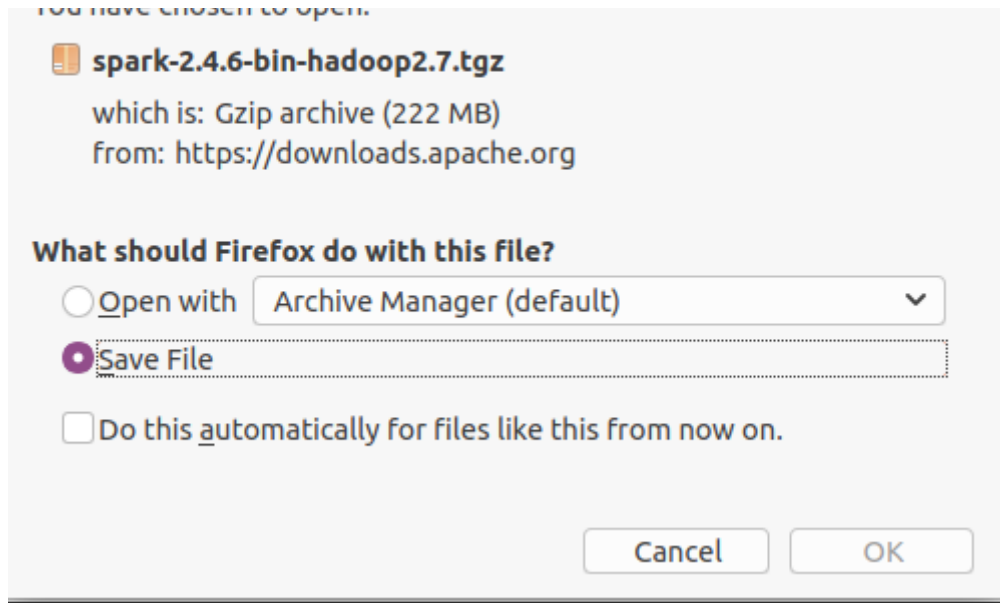
COMMUNITY-LED DEVELOPMENT "THE APACHE

[Projects](#) [People](#) [Community](#) [License](#)

We suggest the following mirror site for your download:

<https://downloads.apache.org/spark/spark-2.4.6/spark-2.4.6-bin-hadoop2.7.tgz>

3. Save the file to your local machine and click 'Ok'.



4. Open your terminal and go to the recently downloaded file.

```
olivia@olivia-VirtualBox:~/Downloads$ ls
jdk-11.0.7_linux-x64_bin.deb  spark-2.4.6-bin-hadoop2.7.tgz
```

5. Let's extract the file using the following command.

```
olivia@olivia-VirtualBox:~/Downloads$ tar -xvzf spark-2.4.6-bin-hadoop2.7.tgz
```

6. After extracting the file, the new file is created and shown using the list('ls') command.

```
olivia@olivia-VirtualBox:~/Downloads$ ls
jdk-11.0.7_linux-x64_bin.deb  spark-2.4.6-bin-hadoop2.7.tgz
spark-2.4.6-bin-hadoop2.7
```

Configuring Environment Variable in Linux

1. Let's open the 'bashrc' file using 'vim editor' by the command 'vim ~/.bashrc'.

```
olivia@olivia-VirtualBox:~$ vim ~/.bashrc
```

2. Provide the following information according to your suitable path on your computer. In my case, the following were the required path to my Spark location, Python path, and Java path. Also, first press 'Esc' and then type ":wq" to save and exit from vim.



```
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export SPARK_HOME=~/.Downloads/spark-2.4.6-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin

export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH

export PYSPARK_PYTHON=python3
export PATH=$PATH:$JAVA_HOME/jre/bin
```

119,55

Bot

3. To make a final change, save, and exit. This results in accessing the pyspark command everywhere in the directory.

```
olivia@olivia-VirtualBox:~$ source ~/.bashrc
```

4. Open pyspark using 'pyspark' command, and the final message will be shown as below.

```
olivia@olivia-VirtualBox:~$ pyspark
Welcome to
  _ _ _ _ _
 / _ _ _ \   version 2.4.
/_ _ _ _ \_
 \ _ _ _ /
  _ _ _ _
```

Mac Installation

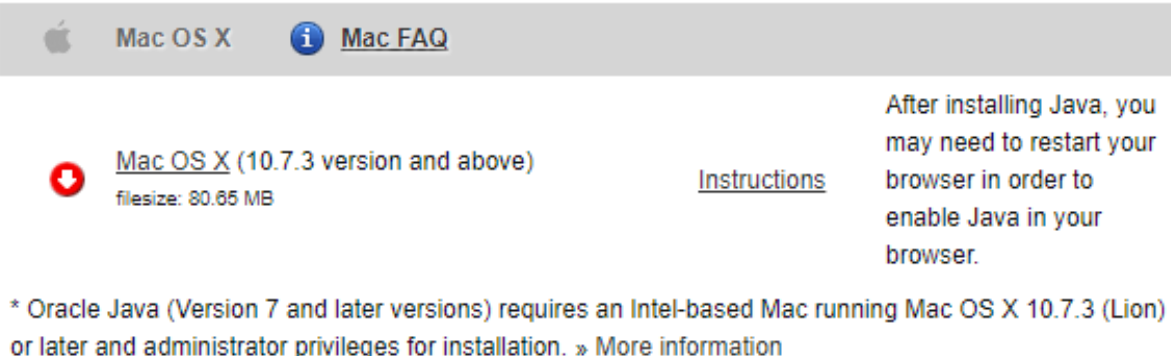
The installation which is going to be shown is for the Mac Operating System. It consists of the installation of Java with the environment variable along with Apache Spark and the environment variable.

The recommended pre-requisite installation is Python, which is done from [here](#).

Java Installation



2. Move to download section consisting of the operating system **Linux** and download according to your system requirement.



Mac OS X [Mac FAQ](#)

[Mac OS X \(10.7.3 version and above\)](#)
filesize: 80.65 MB [Instructions](#)

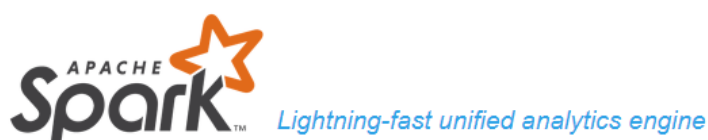
After installing Java, you may need to restart your browser in order to enable Java in your browser.

* Oracle Java (Version 7 and later versions) requires an Intel-based Mac running Mac OS X 10.7.3 (Lion) or later and administrator privileges for installation. » [More information](#)

3. The installation of Java can be confirmed by using `$java --showversion` in the Terminal.

Installing Apache Spark

1. Head over to the [Spark homepage](#).
2. Select the Spark release and package type as following and download the .tgz file.



Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.4.6-bin-hadoop2.7.tgz](#)
4. Verify this release using the 2.4.6 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.



COMMUNITY-LED DEVELOPMENT "THE APACH

Projects ▾

People ▾

Community ▾

License ▾

We suggest the following mirror site for your download:

<https://downloads.apache.org/spark/spark-2.4.6/spark-2.4.6-bin-hadoop2.7.tgz>

3. Save the file to your local machine and click 'Ok'.

4. Let's extract the file using the following command.

```
$ tar -xzf spark-2.4.6-bin-hadoop2.7.tgz
```

Configuring Environment Variable for Apache Spark and Python

You need to open the `~/.bashrc` or `~/.zshrc` file depending upon your current Mac version.

```
export SPARK_HOME="/Downloads/spark"
export PATH=$SPARK_HOME/bin:$PATH
export PYSARK_PYTHON=python3
```

Open pyspark using 'pyspark' command, and the final message will be shown as below.

```
Welcome to
  _ _ _ _ _
 _\V _V _'\ _'\
/_/_/. _\.,/_/_/_\ version 2.4.
/_/_
```

Congratulations

Congratulations, you have made it to the end of this tutorial!

In this tutorial, you've learned about the installation of Pyspark, starting the installation of Java along with Apache Spark and managing the environment variables in Windows, Linux, and Mac Operating System.



 [Subscribe to RSS](#)

[About](#) [Terms](#) [Privacy](#)