CSCI316 Software and Programming Packages Installation Guide to install and Run PySpark in Jupyter Notebook on Windows (Adopted from instructions provided by Dr Guoxin Su)

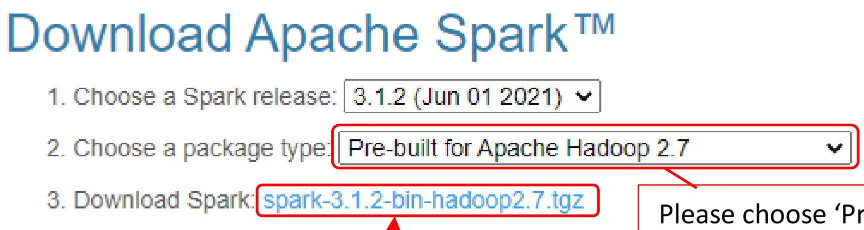
(SJ: 25 June 2021)

Installing and Integrating PySpark with Jupyter Notebook on Windows

A. Downloading PySpark, Anaconda, 7-Zip, and Java JDK.

Before setting up PySpark, we need to have the following packages installed in the System:

- **Download PySpark package**
  i. https://spark.apache.org/downloads.html

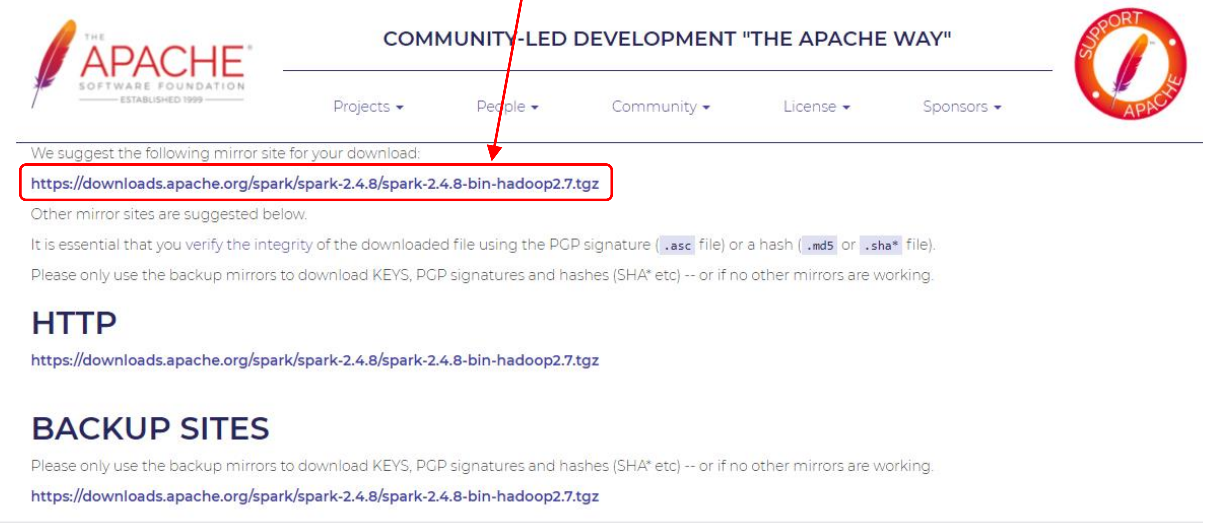


Download Apache Spark™

1. Choose a Spark release: 3.1.2 (Jun 01 2021)
2. Choose a package type: Pre-built for Apache Hadoop 2.7
3. Download Spark: spark-3.1.2-bin-hadoop2.7.tgz

Please choose 'Pre-built for Apache Hadoop 2.7'.

  ii. For the package type, please choose **'Pre-built for Apache Hadoop 2.7'**.
  iii. Click the label **'spark-3.1.2-bin-hadoop2.7.tgz'** to start downloading the package.
  iv. In the next screen, choose the mirror site that you want to download the package from, and double-click on the link to start downloading.



THE APACHE SOFTWARE FOUNDATION ESTABLISHED 1999
COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"
Projects
People
Community
License
Sponsors
SUPPORT APACHE

We suggest the following mirror site for your download:

https://downloads.apache.org/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

HTTP

https://downloads.apache.org/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz

BACKUP SITES

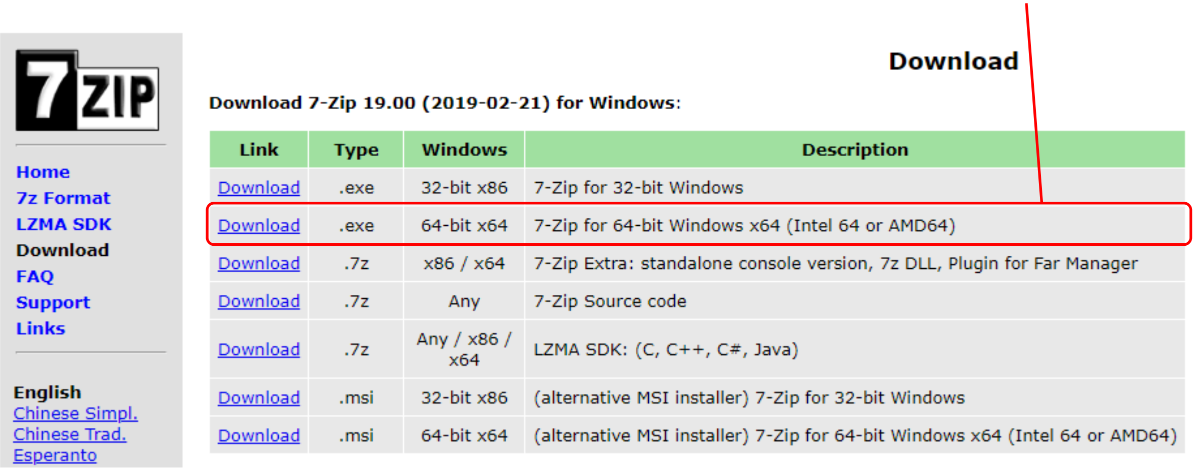
Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

https://downloads.apache.org/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz

  v. Save the downloaded package to a working directory.

- **Installing 7-zip**
  i. If you already have 7-zip installed on your Windows, you can skip this step, otherwise, download the 7-zip installer and install the 7-zip application.
  ii. https://www.7-zip.org/download.html

- Choose the version of installer that fits your system and click the 'Download' link to start downloading the installer. For example, for my system, I choose the **64-bit x64** version.



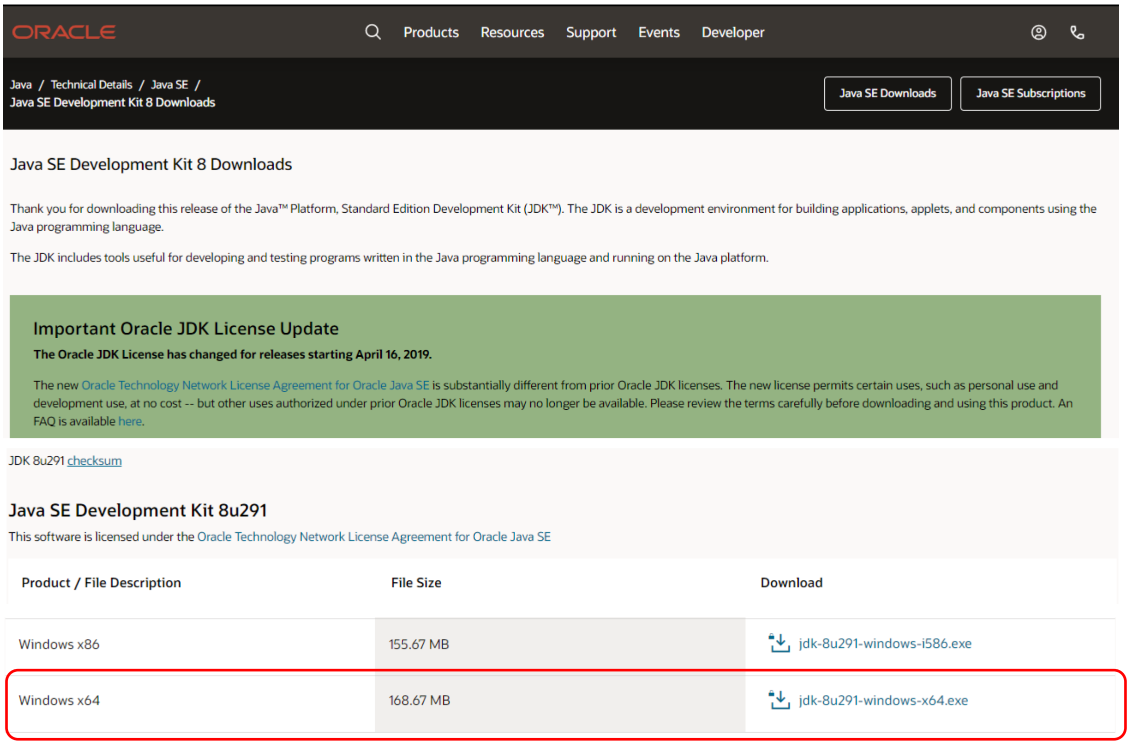
Download

Download 7-Zip 19.00 (2019-02-21) for Windows:

Link	Type	Windows	Description
Download	.exe	32-bit x86	7-Zip for 32-bit Windows
Download	.exe	64-bit x64	7-Zip for 64-bit Windows x64 (Intel 64 or AMD64)
Download	.7z	x86 / x64	7-Zip Extra: standalone console version, 7z DLL, Plugin for Far Manager
Download	.7z	Any	7-Zip Source code
Download	.7z	Any / x86 / x64	LZMA SDK: (C, C++, C#, Java)
Download	.msi	32-bit x86	(alternative MSI installer) 7-Zip for 32-bit Windows
Download	.msi	64-bit x64	(alternative MSI installer) 7-Zip for 64-bit Windows x64 (Intel 64 or AMD64)

- Once the downloading is completed, and the installer is checked, execute the installer to install the 7-Zip application. You can follow all the default setting unless you want to install the 7-Zip differently.

- **Installing Java JDK**
  i. If your system has no Java installed or the Java version is 7.x or less, please download and install Java from Oracle: https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html
  ii. For my system, I downloaded and install Java SE Development Kit (Java JDK 8) (Note: So far this version works, I have tried installing the latest JDK and it did not work.)



Java SE Development Kit 8 Downloads

Thank you for downloading this release of the Java™ Platform, Standard Edition Development Kit (JDK™). The JDK is a development environment for building applications, applets, and components using the Java programming language.

The JDK includes tools useful for developing and testing programs written in the Java programming language and running on the Java platform.

Important Oracle JDK License Update

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

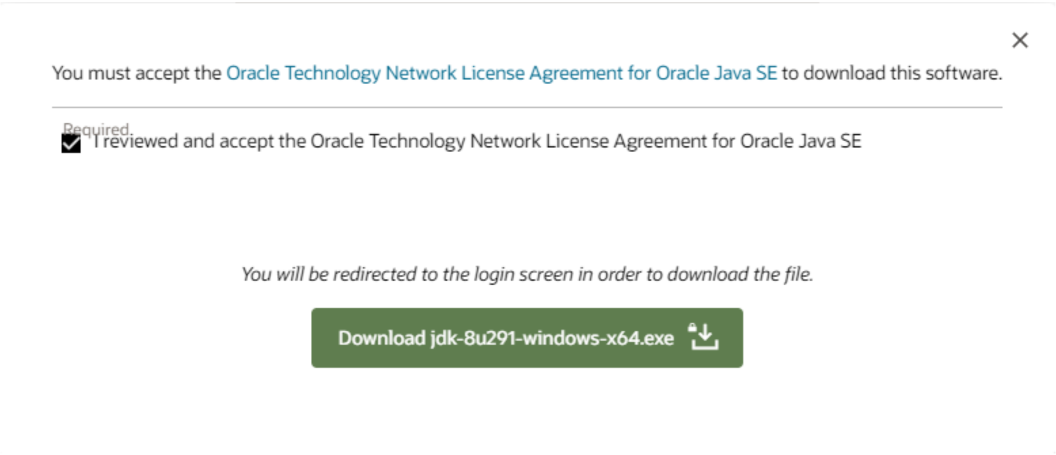
The new Oracle Technology Network License Agreement for Oracle Java SE is substantially different from prior Oracle JDK licenses. The new license permits certain uses, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle JDK licenses may no longer be available.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

JDK 8u291 checksum

Java SE Development Kit 8u291

This software is licensed under the Oracle Technology Network License Agreement for Oracle Java SE

Product / File Description	File Size	Download
Windows x86	155.67 MB	jdk-8u291-windows-i586.exe
Windows x64	168.67 MB	jdk-8u291-windows-x64.exe



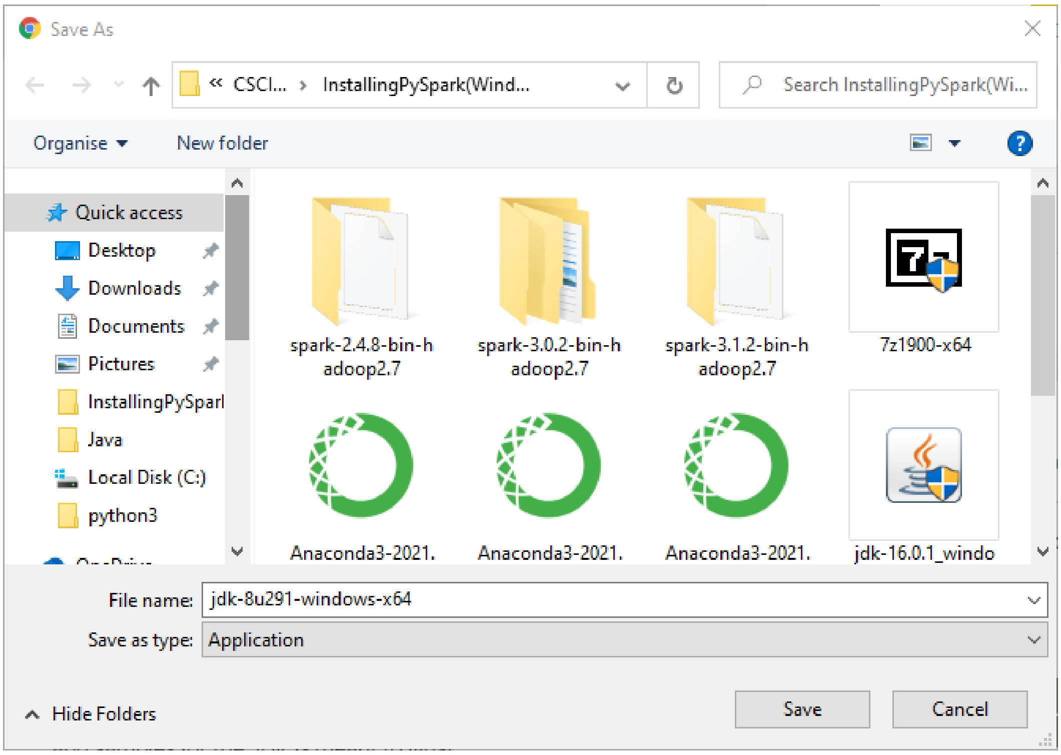
You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

Required.

☑ I reviewed and accept the Oracle Technology Network License Agreement for Oracle Java SE

You will be redirected to the login screen in order to download the file.

- Save to a directory of your choice.

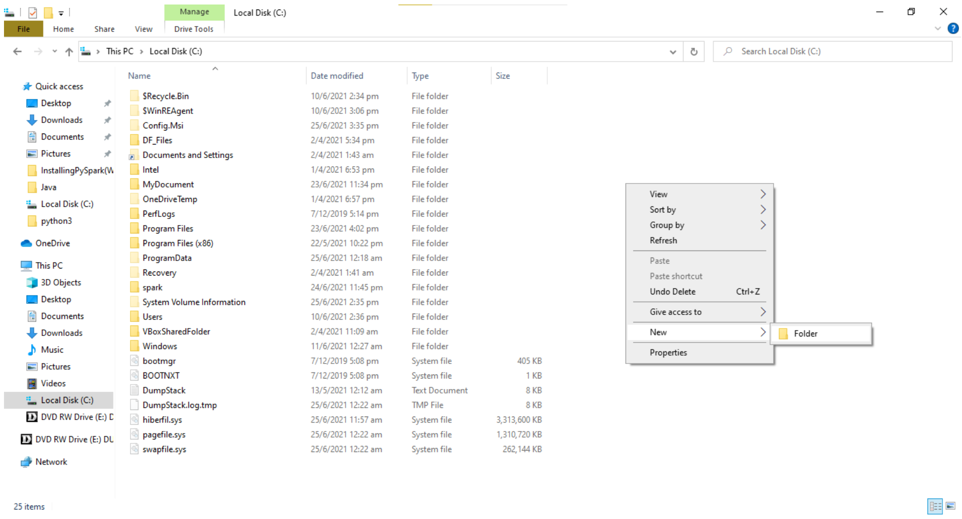


File name: jdk-8u291-windows-x64

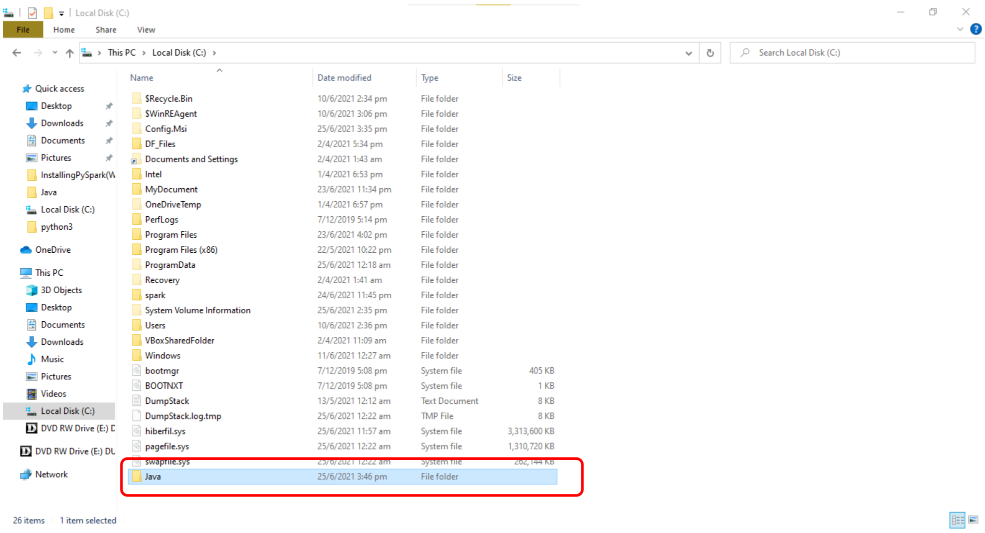
Save as type: Application

- Next, you can proceed to install the Java JDK, but please do not use the default setting from the installer.

- Installing Java JDK:
  i. Create a new folder in your local drive; it can be in ‘C:\’ or ‘D:\’. For me, I install it in my ‘C:\’ drive, and name the folder ‘Java’. Note: Please use a name that does not have a space in between.

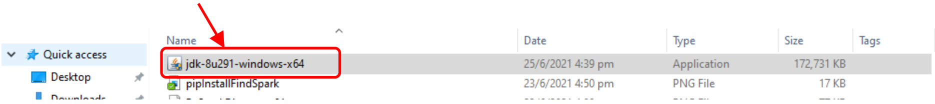


Name	Date modified	Type	Size
$Recycle.Bin	10/6/2021 2:34 pm	File folder	
$WinREAgent	10/6/2021 3:06 pm	File folder	
Config.Msi	25/6/2021 3:35 pm	File folder	
DF_Files	2/4/2021 5:34 pm	File folder	
Documents and Settings	2/4/2021 1:43 am	File folder	
Intel	1/4/2021 6:53 pm	File folder	
MyDocument	23/6/2021 11:34 pm	File folder	
OneDriveTemp	1/4/2021 6:57 pm	File folder	
PerfLogs	7/12/2019 5:14 pm	File folder	
Program Files	23/6/2021 4:02 pm	File folder	
Program Files (x86)	22/5/2021 10:22 pm	File folder	
ProgramData	25/6/2021 12:18 am	File folder	
Recovery	2/4/2021 1:41 am	File folder	
spark	24/6/2021 11:45 pm	File folder	
System Volume Information	25/6/2021 2:35 pm	File folder	
Users	10/6/2021 2:36 pm	File folder	
VBoxSharedFolder	2/4/2021 11:09 am	File folder	
Windows	11/6/2021 12:27 am	File folder	
bootmgr	7/12/2019 5:08 pm	System file	405 KB
BOOTNXT	7/12/2019 5:08 pm	System file	1 KB
DumpStack	13/5/2021 12:12 am	Text Document	8 KB
DumpStack.log.tmp	25/6/2021 12:22 am	TMP File	8 KB
hiberfil.sys	25/6/2021 11:57 am	System file	3,313,600 KB
pagefile.sys	25/6/2021 12:22 am	System file	1,310,720 KB
swapfile.sys	25/6/2021 12:22 am	System file	262,144 KB



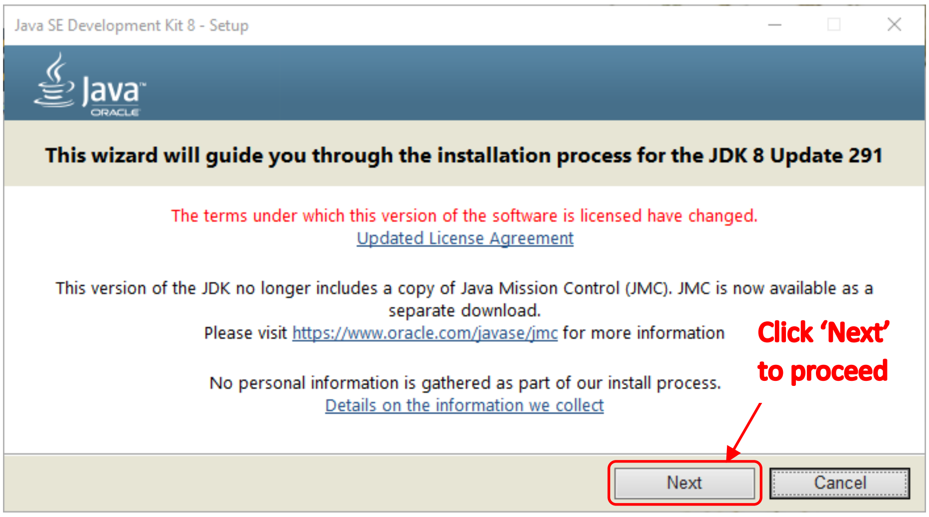
Name	Date modified	Type	Size
$Recycle.Bin	10/6/2021 2:34 pm	File folder	
$WinREAgent	10/6/2021 3:06 pm	File folder	
Config.Msi	25/6/2021 3:35 pm	File folder	
DF_Files	2/4/2021 5:34 pm	File folder	
Documents and Settings	2/4/2021 1:43 am	File folder	
Intel	1/4/2021 6:53 pm	File folder	
MyDocument	23/6/2021 11:34 pm	File folder	
OneDriveTemp	1/4/2021 6:57 pm	File folder	
PerfLogs	7/12/2019 5:14 pm	File folder	
Program Files	23/6/2021 4:02 pm	File folder	
Program Files (x86)	22/5/2021 10:22 pm	File folder	
ProgramData	25/6/2021 12:18 am	File folder	
Recovery	2/4/2021 1:41 am	File folder	
spark	24/6/2021 11:45 pm	File folder	
System Volume Information	25/6/2021 2:35 pm	File folder	
Users	10/6/2021 2:36 pm	File folder	
VBoxSharedFolder	2/4/2021 11:09 am	File folder	
Windows	11/6/2021 12:27 am	File folder	
bootmgr	7/12/2019 5:08 pm	System file	405 KB
BOOTNXT	7/12/2019 5:08 pm	System file	1 KB
DumpStack	13/5/2021 12:12 am	Text Document	8 KB
DumpStack.log.tmp	25/6/2021 12:22 am	TMP File	8 KB
hiberfil.sys	25/6/2021 11:57 am	System file	3,313,600 KB
pagefile.sys	25/6/2021 12:22 am	System file	1,310,720 KB
swapfile.sys	25/6/2021 12:22 am	System file	262,144 KB
Java	25/6/2021 3:46 pm	File folder	

ii. Execute your java JDK installer. From the directory where you save the Java JDK installer, **double-click** on the installer to start the installation.



Name	Date	Type	Size	Tags
jdk-8u291-windows-x64	25/6/2021 4:39 pm	Application	172,731 KB	
pipInstallFindSpark	23/6/2021 4:50 pm	PNG File	17 KB	

iii. In the pop-up menu, click ‘Next’ to proceed.



Java SE Development Kit 8 - Setup

This wizard will guide you through the installation process for the JDK 8 Update 291

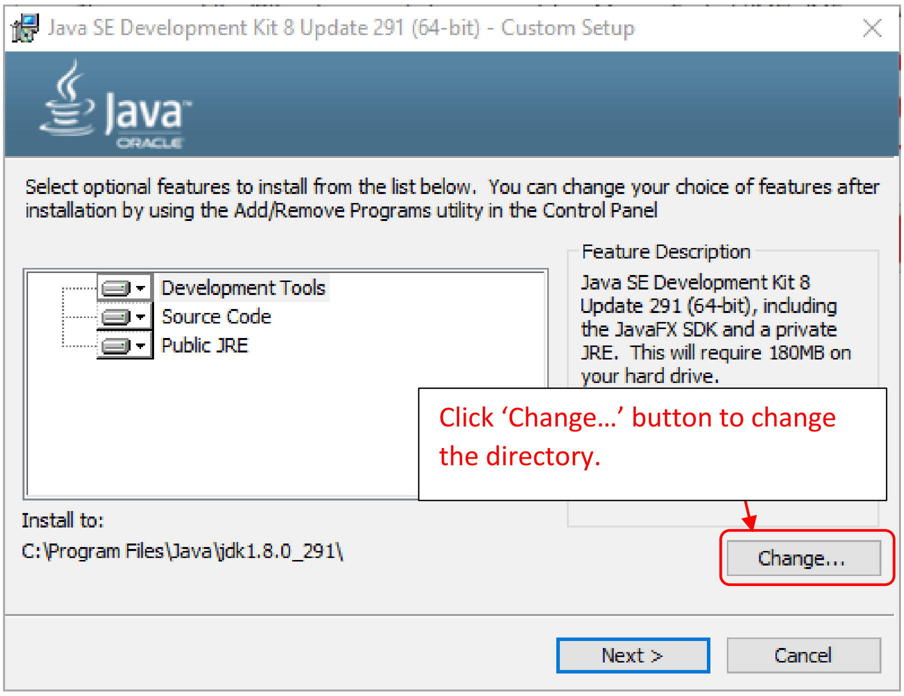
The terms under which this version of the software is licensed have changed.
Updated License Agreement

This version of the JDK no longer includes a copy of Java Mission Control (JMC). JMC is now available as a separate download.
Please visit https://www.oracle.com/javase/jmc for more information

No personal information is gathered as part of our install process.
Details on the information we collect

Click ‘Next’ to proceed

iv. Click the ‘Change…’ button to change the directory where you intend to install your Java JDK.



Java SE Development Kit 8 Update 291 (64-bit) - Custom Setup

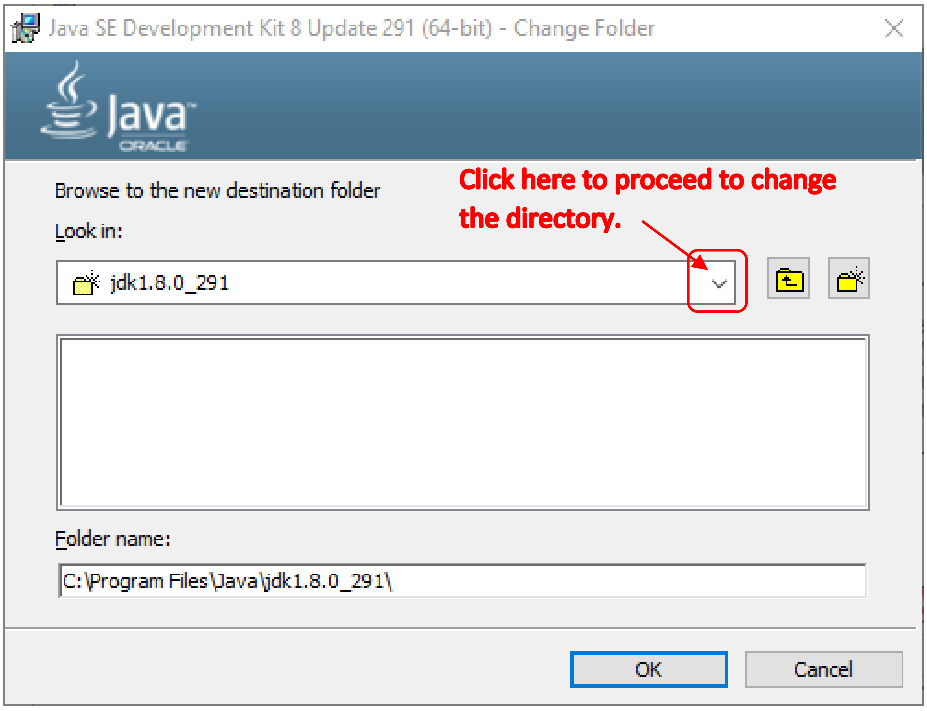
Select optional features to install from the list below. You can change your choice of features after installation by using the Add/Remove Programs utility in the Control Panel

Feature Description
Java SE Development Kit 8 Update 291 (64-bit), including the JavaFX SDK and a private JRE. This will require 180MB on your hard drive.

Install to:
C:\Program Files\Java\jdk1.8.0_291\

Click ‘Change…’ button to change the directory.

v. In the next window, click on the ⌄ icon to proceed to change the directory.



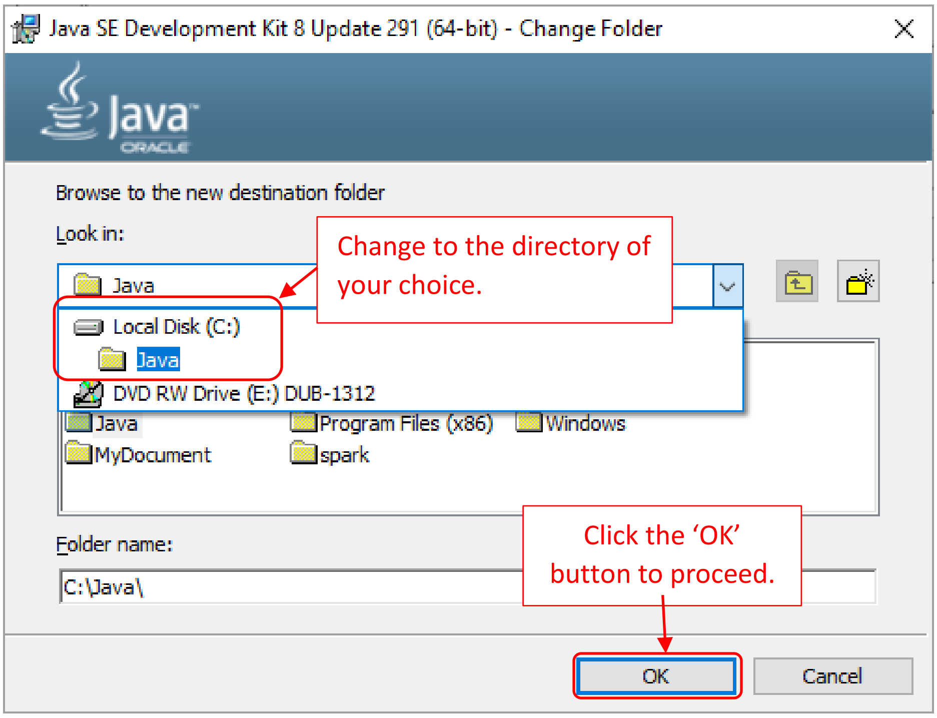
Java SE Development Kit 8 Update 291 (64-bit) - Change Folder

Browse to the new destination folder

Look in: jdk1.8.0_291

Folder name: C:\Program Files\Java\jdk1.8.0_291\

vi. Change to the directory where you want to install you Java JDK and click the 'OK' button.



Java SE Development Kit 8 Update 291 (64-bit) - Change Folder

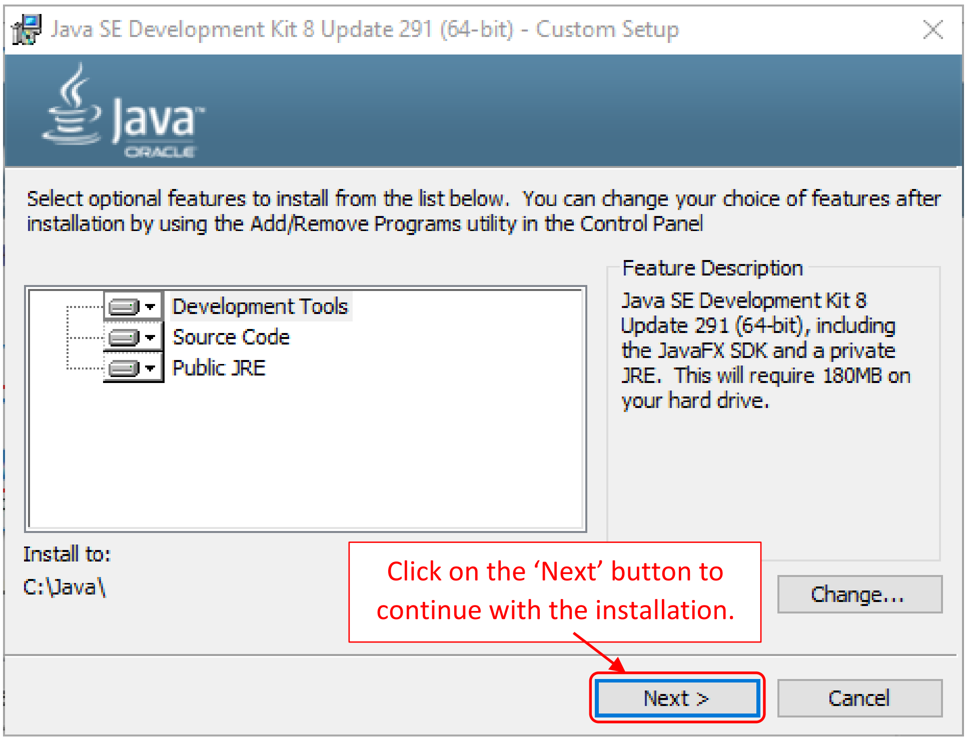
Browse to the new destination folder

Look in: Java

Folder name: C:\Java\

Local Disk (C:)
Java
DVD RW Drive (E:) DUB-1312
Java
Program Files (x86)
Windows
MyDocument
spark
Change to the directory of your choice.
Click the 'OK' button to proceed.
OK
Cancel

vii. Click on the ‘Next’ button to continue the installation.



Java SE Development Kit 8 Update 291 (64-bit) - Custom Setup

Select optional features to install from the list below. You can change your choice of features after installation by using the Add/Remove Programs utility in the Control Panel

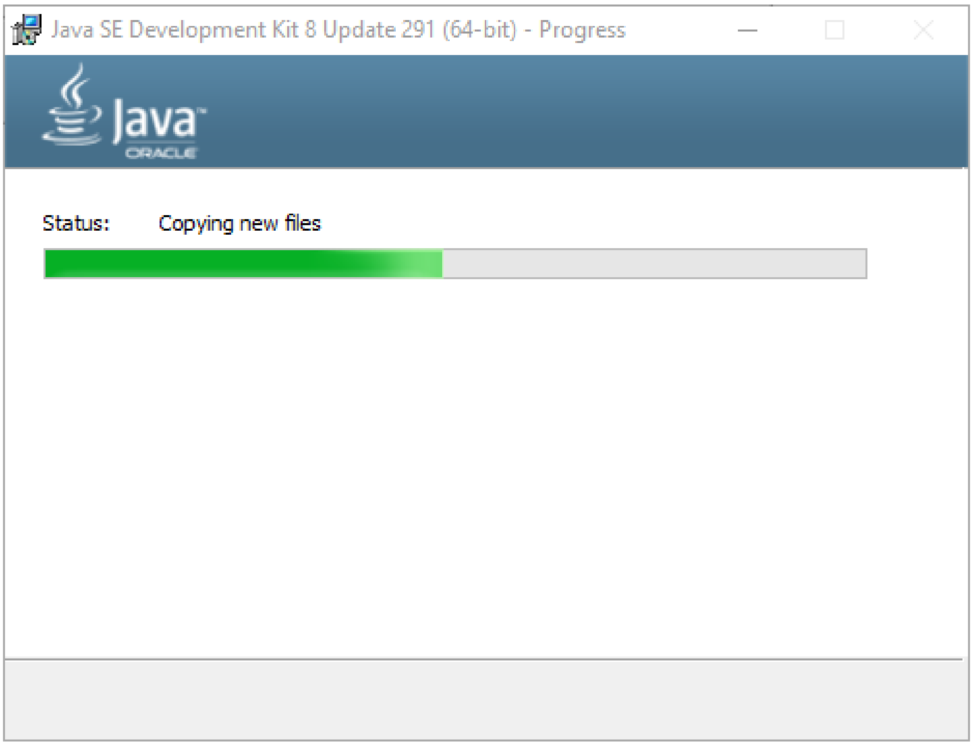
Feature Description

Java SE Development Kit 8 Update 291 (64-bit), including the JavaFX SDK and a private JRE. This will require 180MB on your hard drive.

Install to:

C:\Java\

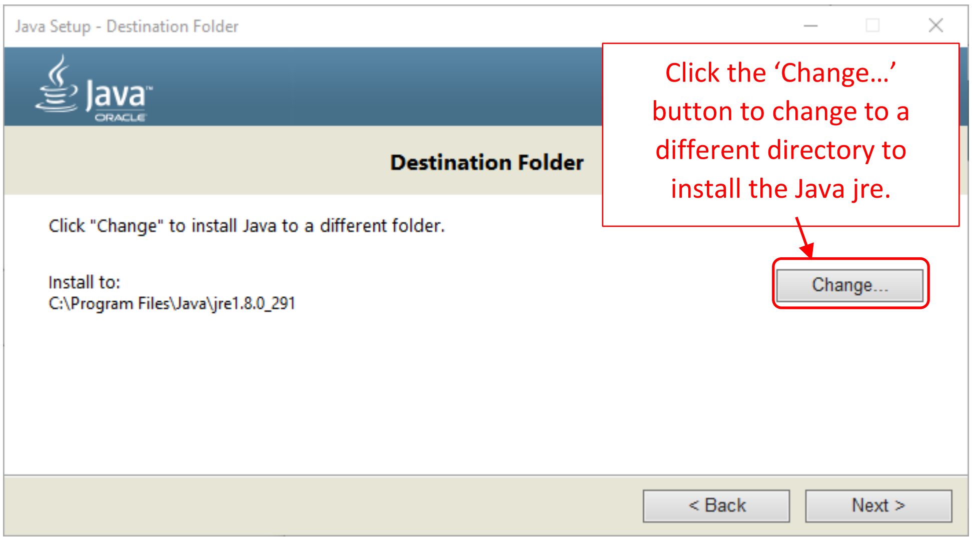
Click on the ‘Next’ button to continue with the installation.



Java SE Development Kit 8 Update 291 (64-bit) - Progress

Status: Copying new files

viii. In the next window, similarly, click the ‘Change…’ button to change to a different directory to install the Java jre.



Java Setup - Destination Folder

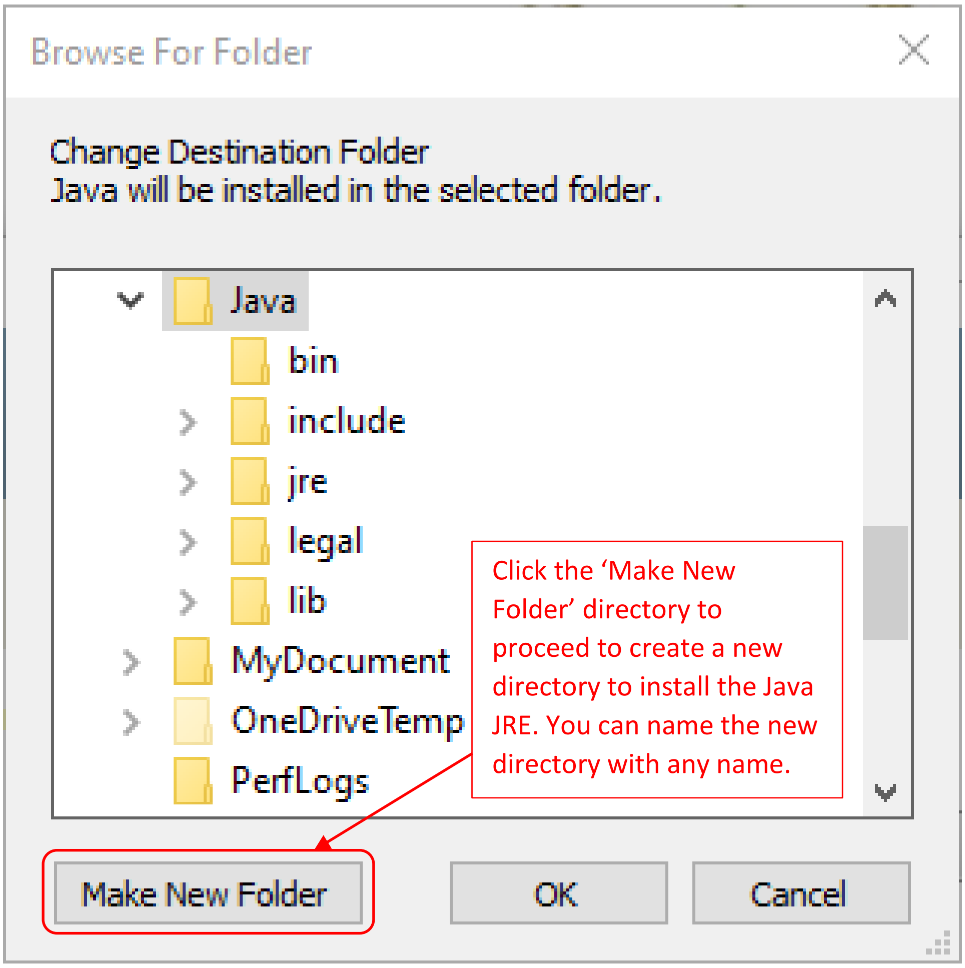
Destination Folder

Click "Change" to install Java to a different folder.

Install to:
C:\Program Files\Java\jre1.8.0_291

Click the ‘Change…’ button to change to a different directory to install the Java jre.

ix. Change to the directory where you want to install you Java JRE. You may want to create (make) a new directory for that.

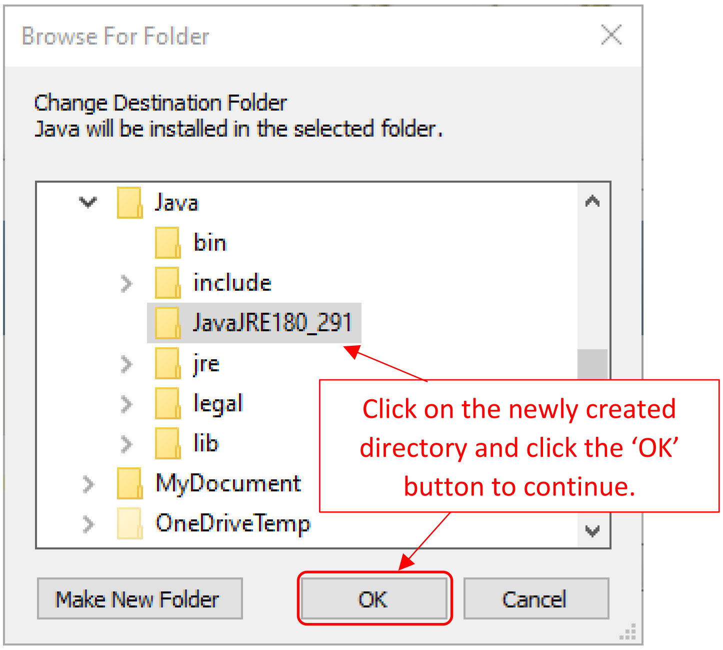


Browse For Folder

Change Destination Folder
Java will be installed in the selected folder.

Click the ‘Make New Folder’ directory to proceed to create a new directory to install the Java JRE. You can name the new directory with any name.

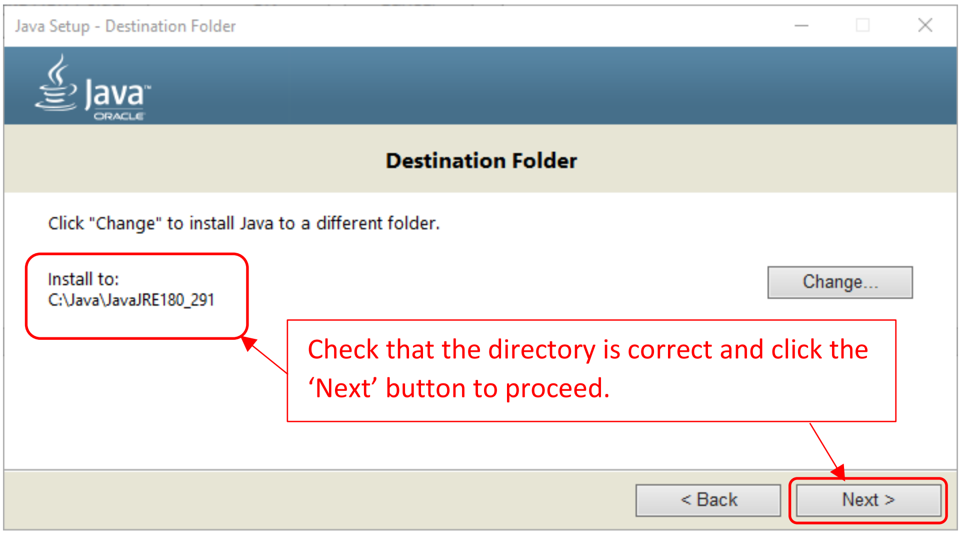
x. Click on the newly created directory (in my example, JavaJRE180_291) and click the ‘OK’ button to continue.



Change Destination Folder
Java will be installed in the selected folder.

Click on the newly created directory and click the ‘OK’ button to continue.

xi. Check that the directory is correct and click the ‘Next’ button to proceed with the installation.

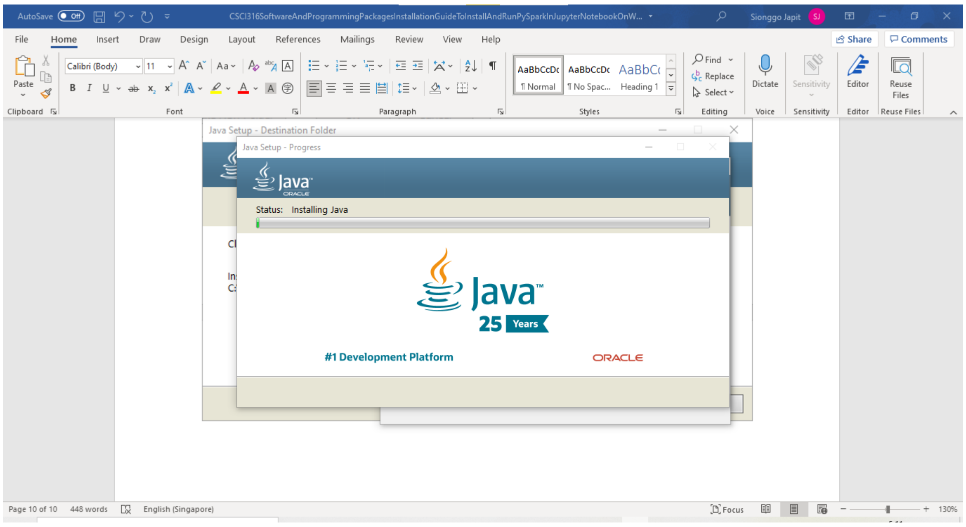


Click "Change" to install Java to a different folder.

Install to:
C:\Java\JavaJRE180_291

Check that the directory is correct and click the ‘Next’ button to proceed.

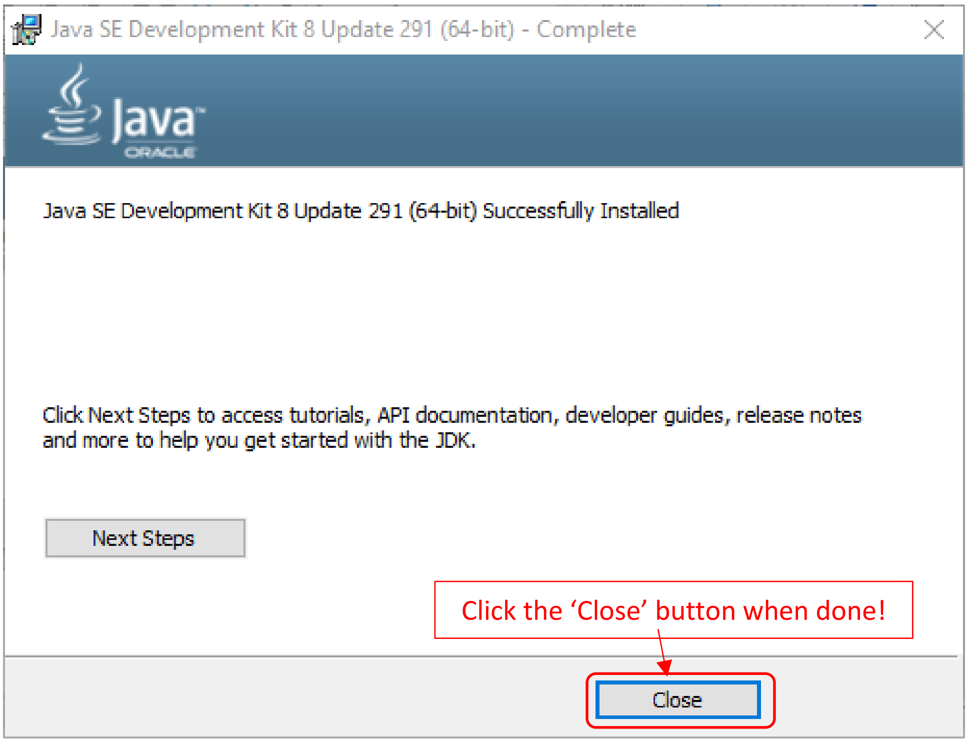
xii. Wait for the installation to complete….



Java Setup - Progress

Status: Installing Java

xiii. Click the ‘Close’ button when done.



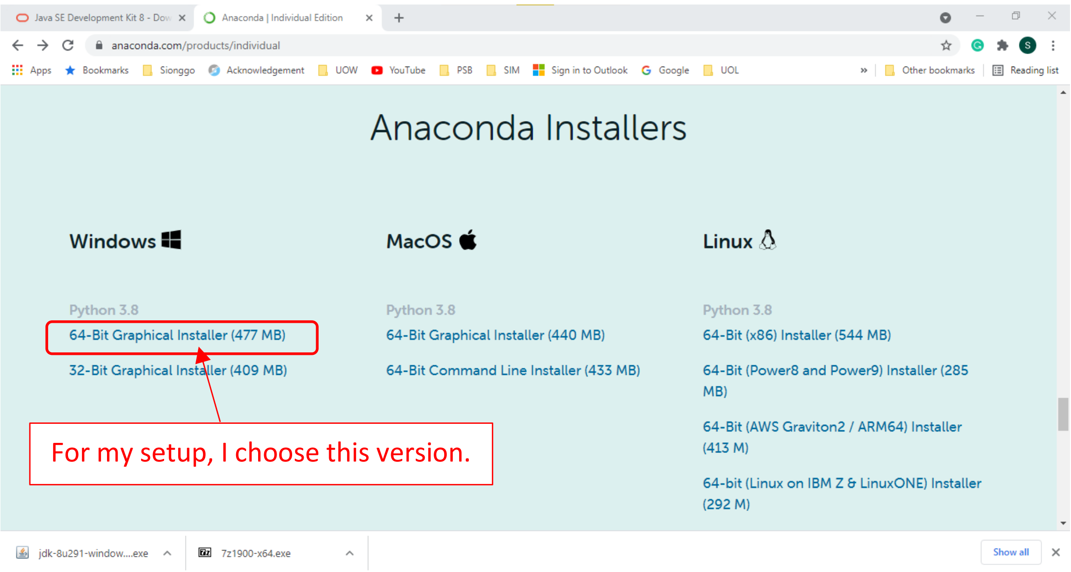
Java SE Development Kit 8 Update 291 (64-bit) - Complete

Java SE Development Kit 8 Update 291 (64-bit) Successfully Installed

Click Next Steps to access tutorials, API documentation, developer guides, release notes and more to help you get started with the JDK.

Click the ‘Close’ button when done!

- **Installing Python and Jupyter Notebook**
  i. You can get both Python and Jupyter by installing the Python 3.x version of Anaconda distribution.
  ii. https://www.anaconda.com/products/individual
  iii. Scroll all the way down to the bottom of the page and choose the version of the installer that suit your system to start the download.

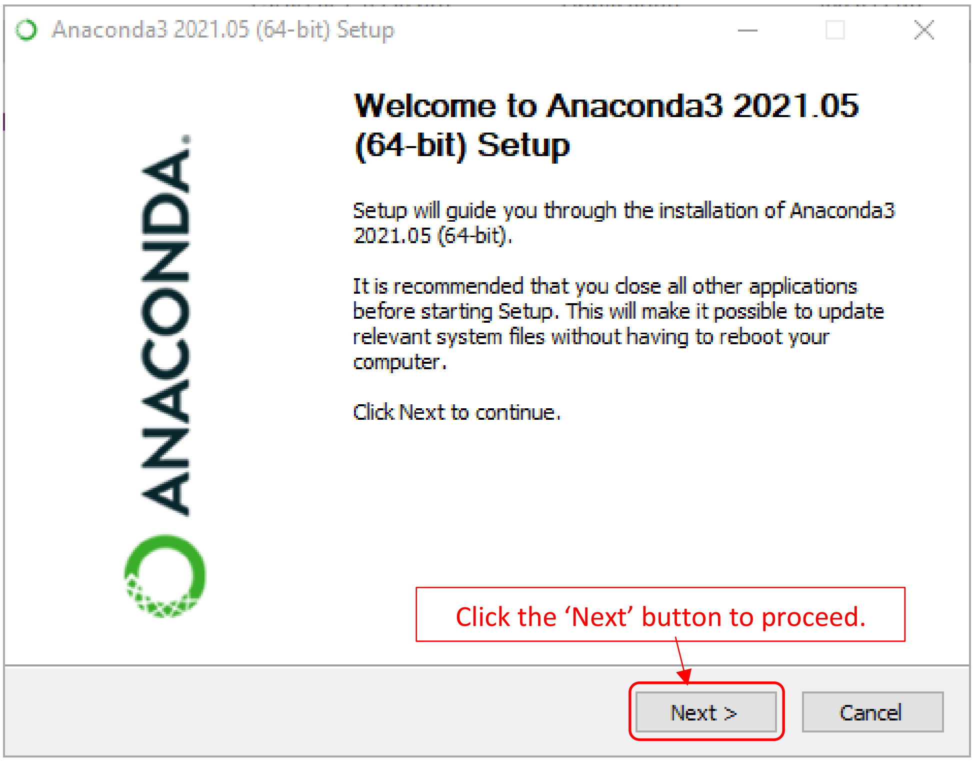


Anaconda Installers

Windows	MacOS	Linux
Python 3.8	Python 3.8	Python 3.8
64-Bit Graphical Installer (477 MB)	64-Bit Graphical Installer (440 MB)	64-Bit (x86) Installer (544 MB)
32-Bit Graphical Installer (409 MB)	64-Bit Command Line Installer (433 MB)	64-Bit (Power8 and Power9) Installer (285 MB)
		64-Bit (AWS Graviton2 / ARM64) Installer (413 M)
		64-bit (Linux on IBM Z & LinuxONE) Installer (292 M)

For my setup, I choose this version.

  iv. Click the 'Download' button to start the download.
  v. Installing Anaconda:
     - After the downloading is complete, you can proceed to install the Anaconda distribution.



Welcome to Anaconda3 2021.05 (64-bit) Setup

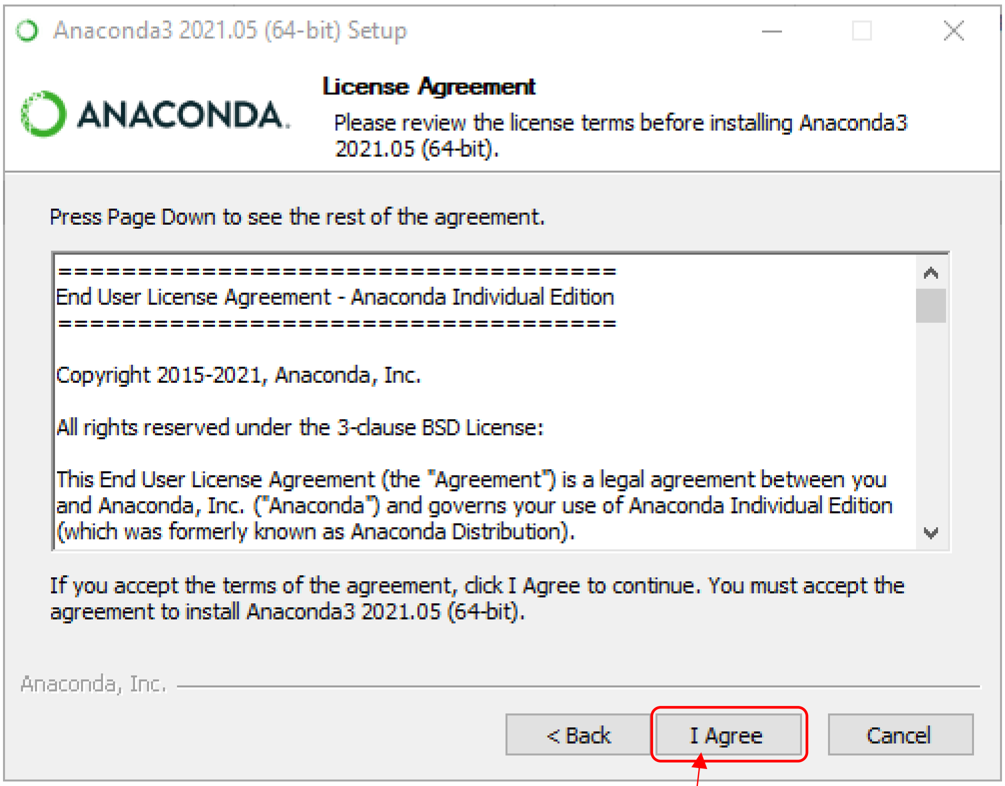
Setup will guide you through the installation of Anaconda3 2021.05 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Click the 'Next' button to proceed.

- Read the ‘License Agreement’ and when finished reading, click the ‘I Agree’ button to proceed with the installation if you agree with the terms and condition specified in the License Agreement. 😊.



Anaconda3 2021.05 (64-bit) Setup

License Agreement

Please review the license terms before installing Anaconda3 2021.05 (64-bit).

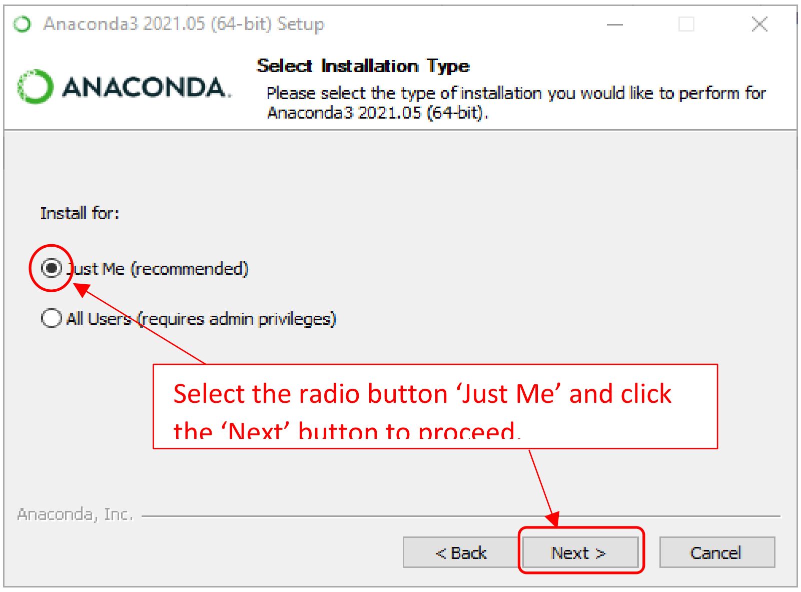
Press Page Down to see the rest of the agreement.

```
===================================
End User License Agreement - Anaconda Individual Edition
===================================

Copyright 2015-2021, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you
and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Individual Edition
(which was formerly known as Anaconda Distribution).
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Anaconda3 2021.05 (64-bit).

Click the ‘I Agree’ button to proceed with the installation if you agree with the terms and condition specified in the License Agreement.

- Select the radio button ‘Just Me’ and click the ‘Next’ button to proceed.



Anaconda3 2021.05 (64-bit) Setup

Select Installation Type

Please select the type of installation you would like to perform for Anaconda3 2021.05 (64-bit).

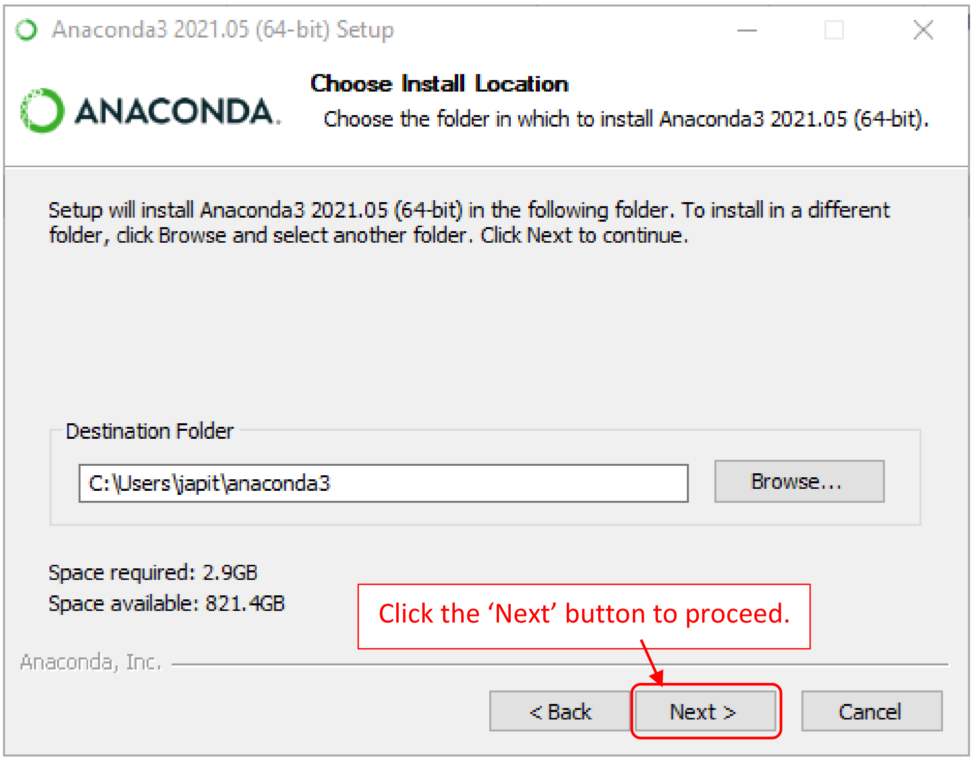
Install for:

● Just Me (recommended)

○ All Users (requires admin privileges)

Select the radio button ‘Just Me’ and click the ‘Next’ button to proceed.

- You can accept the proposed destination folder. You can also change to a different directory if you want to. Click the 'Next' button to proceed.



Anaconda3 2021.05 (64-bit) Setup

Choose Install Location

Choose the folder in which to install Anaconda3 2021.05 (64-bit).

Setup will install Anaconda3 2021.05 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

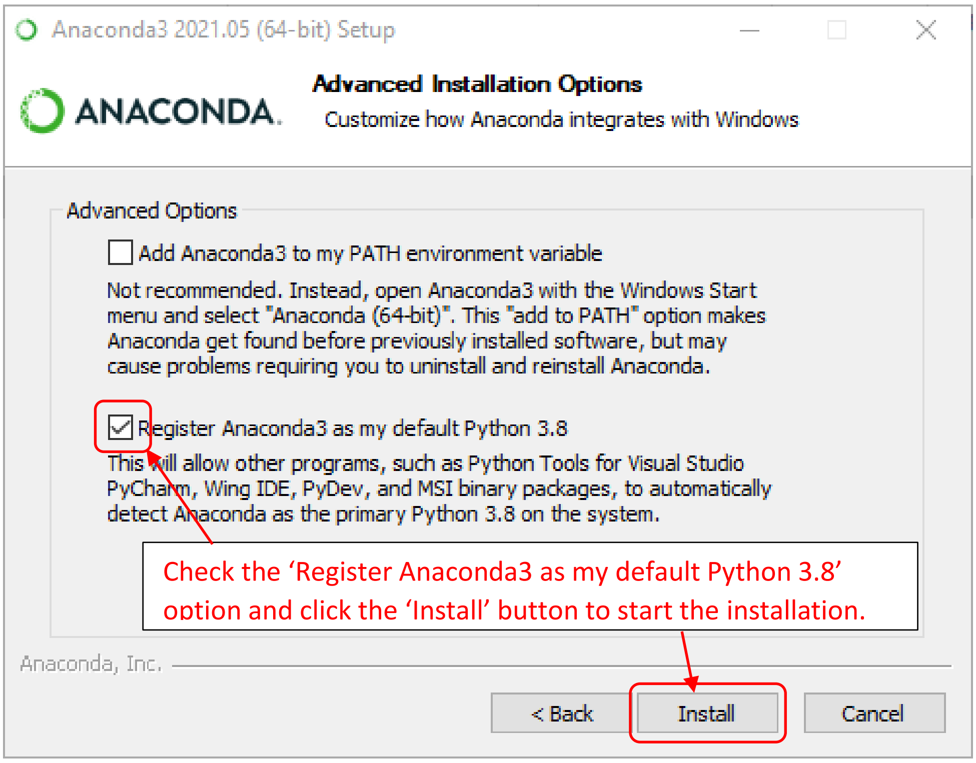
C:\Users\japit\anaconda3

Space required: 2.9GB
Space available: 821.4GB

Anaconda, Inc.

Click the 'Next' button to proceed.

- Check the 'Register Anaconda3 as my default Python 3.8' option and click the 'Install' button to start the installation. This installation will take a while. Just relax....



Anaconda3 2021.05 (64-bit) Setup

Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

□ Add Anaconda3 to my PATH environment variable

Not recommended. Instead, open Anaconda3 with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda3 as my default Python 3.8

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.8 on the system.

Check the 'Register Anaconda3 as my default Python 3.8' option and click the 'Install' button to start the installation.

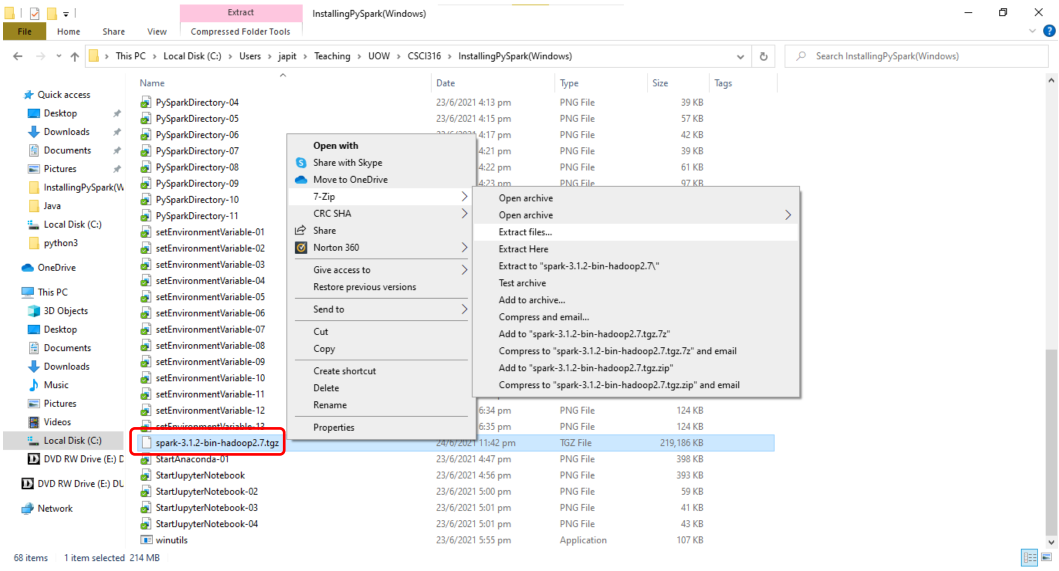
Anaconda, Inc.

B. Installing PySpark

- While waiting for the Anaconda to finish the installation, we can proceed to un-zip the PySpark package for the PySpark setup.

  i. Unpack the spark package ‘spark-3.1.2-bin-hadoop2.7.tgz’ that was downloaded earlier. I save the package in my drive C:\Users\japit\Teaching\UOW\CSCI316\InstallingPySpark(Windows)\ spark-3.1.2-bin-hadoop2.7.tgz.

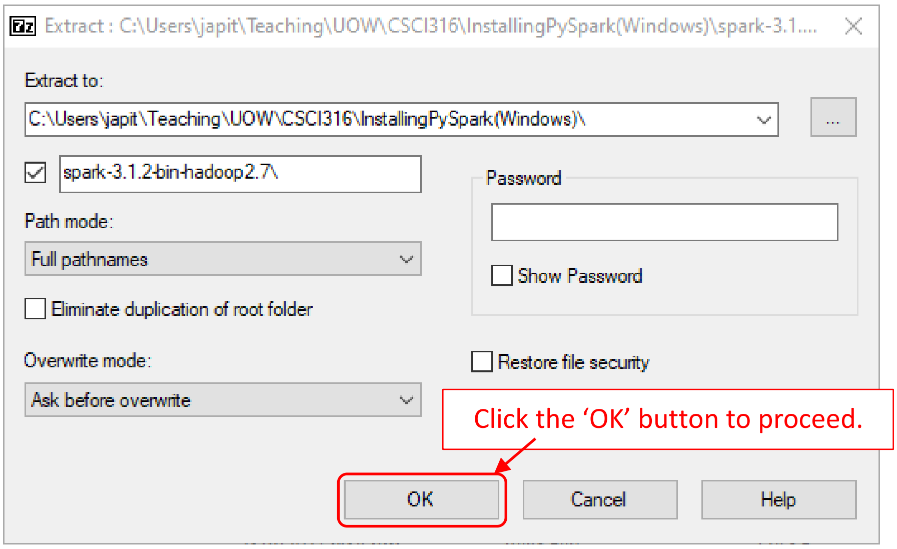
  This package is rather special. It has been packed (zipped) two times. Hence, you need to unpack (unzip) the package two times.



Name	Date	Type	Size	Tags
PySparkDirectory-04	23/6/2021 4:13 pm	PNG File	39 KB	
PySparkDirectory-05	23/6/2021 4:15 pm	PNG File	57 KB	
PySparkDirectory-06	4:17 pm	PNG File	42 KB	
PySparkDirectory-07	4:21 pm	PNG File	39 KB	
PySparkDirectory-08	4:22 pm	PNG File	61 KB	
PySparkDirectory-09	4:23 pm	PNG File	97 KB	
PySparkDirectory-10				
PySparkDirectory-11				
setEnvironmentVariable-01				
setEnvironmentVariable-02				
setEnvironmentVariable-03				
setEnvironmentVariable-04				
setEnvironmentVariable-05				
setEnvironmentVariable-06				
setEnvironmentVariable-07				
setEnvironmentVariable-08				
setEnvironmentVariable-09				
setEnvironmentVariable-10				
setEnvironmentVariable-11	6:34 pm	PNG File	124 KB	
setEnvironmentVariable-12	6:35 pm	PNG File	124 KB	
setEnvironmentVariable-13				
spark-3.1.2-bin-hadoop2.7.tgz	24/6/2021 11:42 pm	TGZ File	219,186 KB	
StartAnaconda-01	23/6/2021 4:47 pm	PNG File	398 KB	
StartJupyterNotebook	23/6/2021 4:56 pm	PNG File	393 KB	
StartJupyterNotebook-02	23/6/2021 5:00 pm	PNG File	59 KB	
StartJupyterNotebook-03	23/6/2021 5:01 pm	PNG File	41 KB	
StartJupyterNotebook-04	23/6/2021 5:01 pm	PNG File	43 KB	
winutils	23/6/2021 5:55 pm	Application	107 KB	

68 items 1 item selected 214 MB

Click the ‘OK’ button to proceed.


Extract : C:\Users\japit\Teaching\UOW\CSCI316\InstallingPySpark(Windows)\spark-3.1....
...
OK
Cancel
Help

Extract to:
C:\Users\japit\Teaching\UOW\CSCI316\InstallingPySpark(Windows)\

☑ spark-3.1.2-bin-hadoop2.7\

Password

□ Show Password

Path mode:
Full pathnames

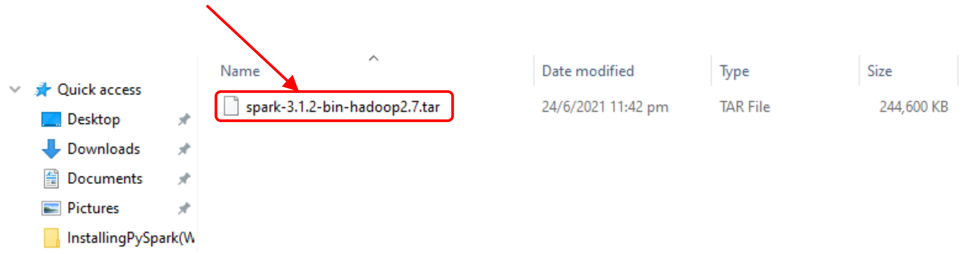
□ Eliminate duplication of root folder

Overwrite mode:
Ask before overwrite

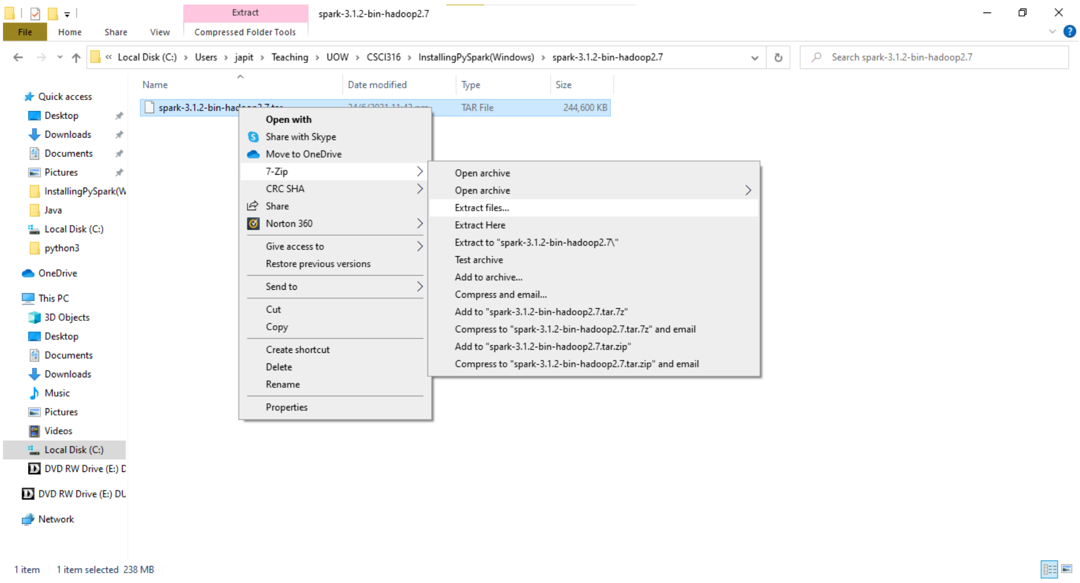
□ Restore file security

Click the ‘OK’ button to proceed.

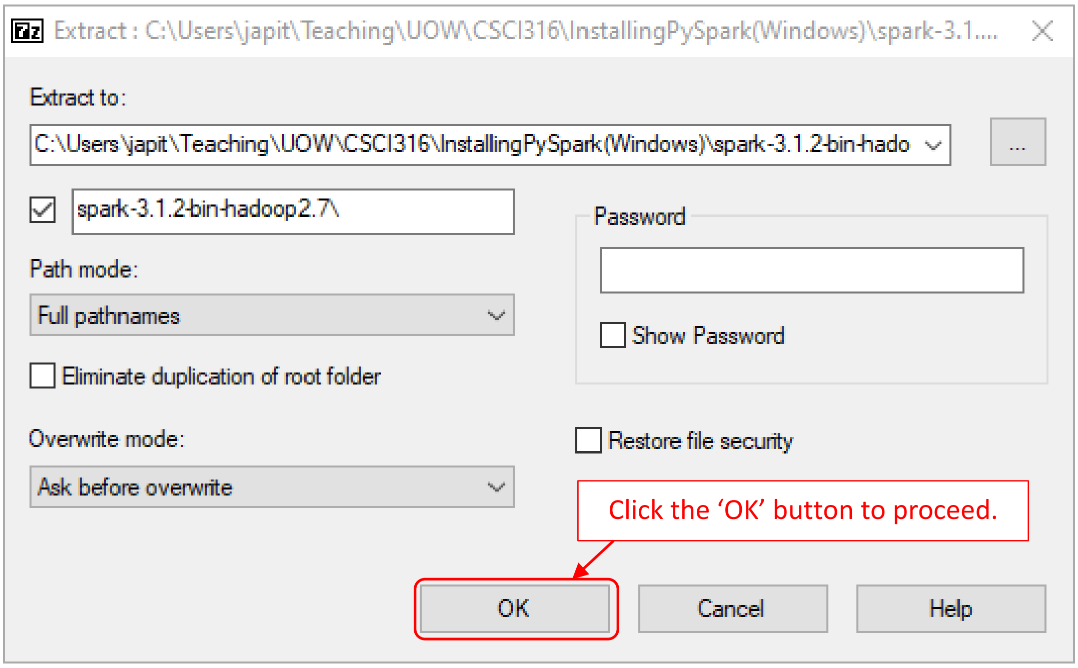
After the unpacking, navigate to the folder of the unpacked file, and do another unpacking process.



Name	Date modified	Type	Size
spark-3.1.2-bin-hadoop2.7.tar	24/6/2021 11:42 pm	TAR File	244,600 KB



Click the ‘OK’ button to proceed with the unpacking process.



Extract : C:\Users\japit\Teaching\UOW\CSCI316\InstallingPySpark(Windows)\spark-3.1....

Extract to: C:\Users\japit\Teaching\UOW\CSCI316\InstallingPySpark(Windows)\spark-3.1.2-bin-hado

☑ spark-3.1.2-bin-hadoop2.7\

Path mode: Full pathnames

□ Eliminate duplication of root folder

Overwrite mode: Ask before overwrite

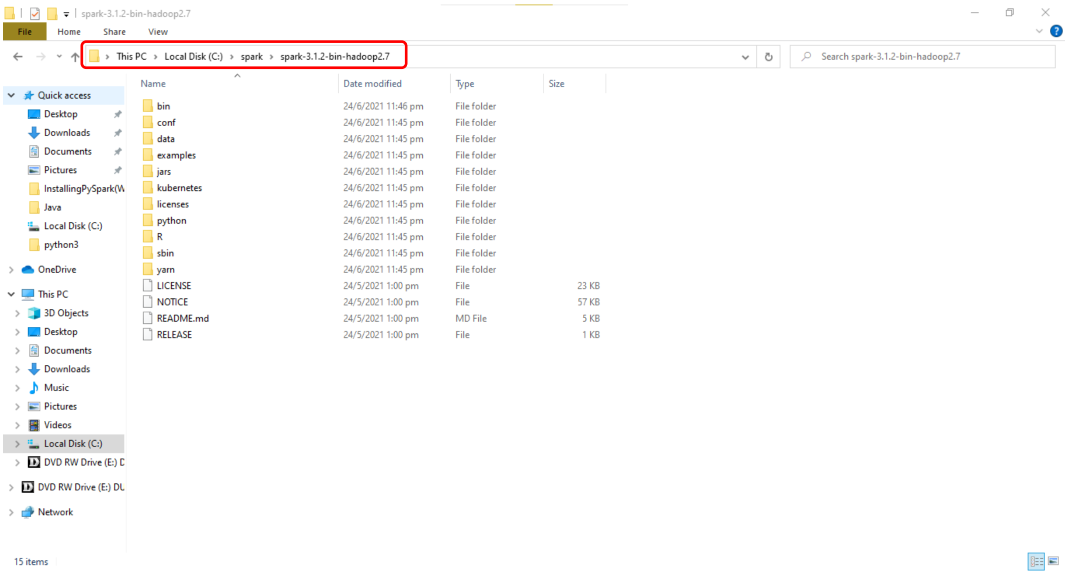
Password

□ Show Password

□ Restore file security

Click the ‘OK’ button to proceed.

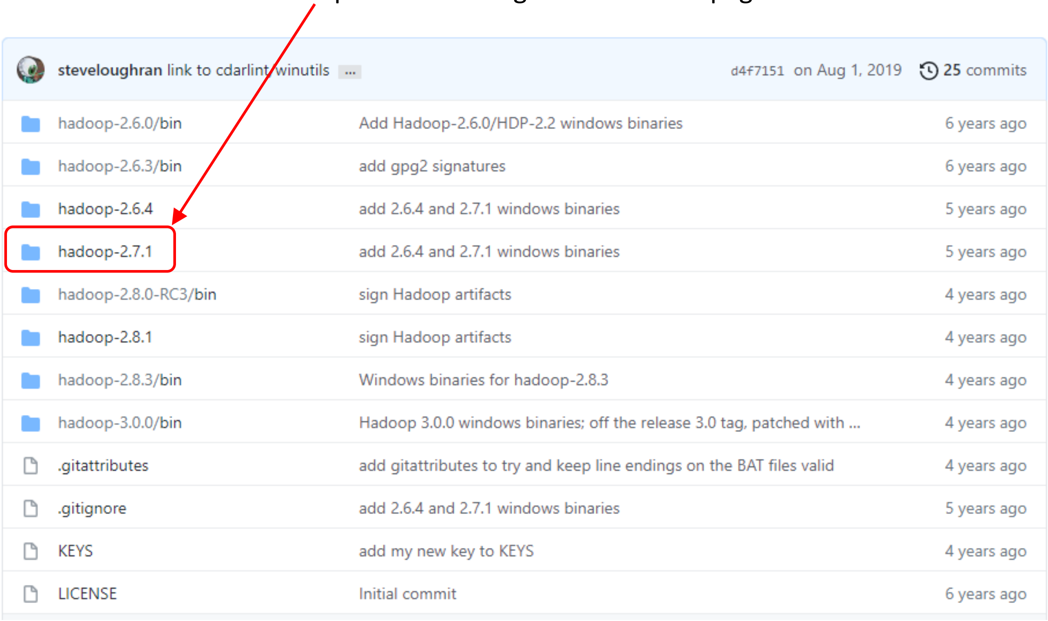
ii. Put the unpacked package to a directory of your choice. I put mine under C:\spark\spark-3.1.2-bin-hadoop2.7.



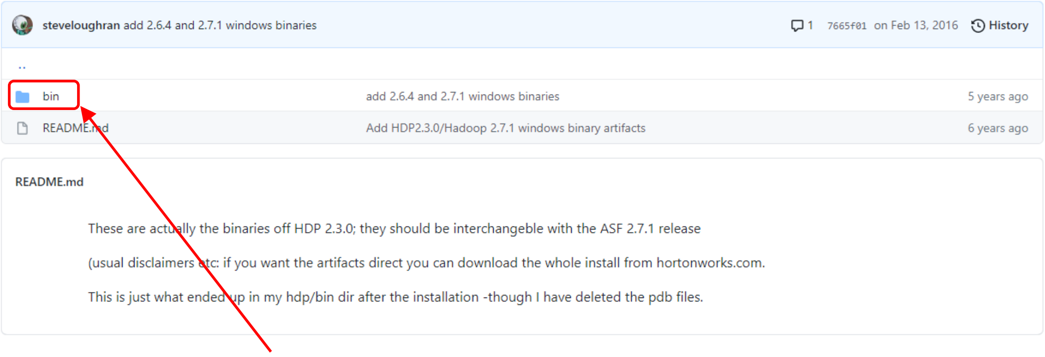
Name	Date modified	Type	Size
bin	24/6/2021 11:46 pm	File folder	
conf	24/6/2021 11:45 pm	File folder	
data	24/6/2021 11:45 pm	File folder	
examples	24/6/2021 11:45 pm	File folder	
jars	24/6/2021 11:45 pm	File folder	
kubernetes	24/6/2021 11:45 pm	File folder	
licenses	24/6/2021 11:45 pm	File folder	
python	24/6/2021 11:45 pm	File folder	
R	24/6/2021 11:45 pm	File folder	
sbin	24/6/2021 11:45 pm	File folder	
yarn	24/6/2021 11:45 pm	File folder	
LICENSE	24/5/2021 1:00 pm	File	23 KB
NOTICE	24/5/2021 1:00 pm	File	57 KB
README.md	24/5/2021 1:00 pm	MD File	5 KB
RELEASE	24/5/2021 1:00 pm	File	1 KB

iii. ‘winutils’ is a collection of useful TCL commands that access some part of the Win32 API. This enables the user to use Windows specific services. Winutils is required when installing Hadoop on Windows environment. Winutils can be downloaded from Steve Loughran’s GitHub repo.

- https://github.com/steveloughran/winutils/
- Go to the corresponding Hadoop version in the Spark distribution and find winutils.exe under /bin. Note that when we download the PySpark package in the earlier step, we choose the ‘Pre-built for Apache Hadoop 2.7’.
- Click the link ‘hadoop-2.7.1’ to navigate to the next page.



hadoop-2.6.0/bin	Add Hadoop-2.6.0/HDP-2.2 windows binaries	6 years ago
hadoop-2.6.3/bin	add gpg2 signatures	6 years ago
hadoop-2.6.4	add 2.6.4 and 2.7.1 windows binaries	5 years ago
hadoop-2.7.1	add 2.6.4 and 2.7.1 windows binaries	5 years ago
hadoop-2.8.0-RC3/bin	sign Hadoop artifacts	4 years ago
hadoop-2.8.1	sign Hadoop artifacts	4 years ago
hadoop-2.8.3/bin	Windows binaries for hadoop-2.8.3	4 years ago
hadoop-3.0.0/bin	Hadoop 3.0.0 windows binaries; off the release 3.0 tag, patched with ...	4 years ago
.gitattributes	add gitattributes to try and keep line endings on the BAT files valid	4 years ago
.gitignore	add 2.6.4 and 2.7.1 windows binaries	5 years ago
KEYS	add my new key to KEYS	4 years ago
LICENSE	Initial commit	6 years ago



steveloughran add 2.6.4 and 2.7.1 windows binaries 1 7665f01 on Feb 13, 2016 History

..

bin	add 2.6.4 and 2.7.1 windows binaries	5 years ago
README.md	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago

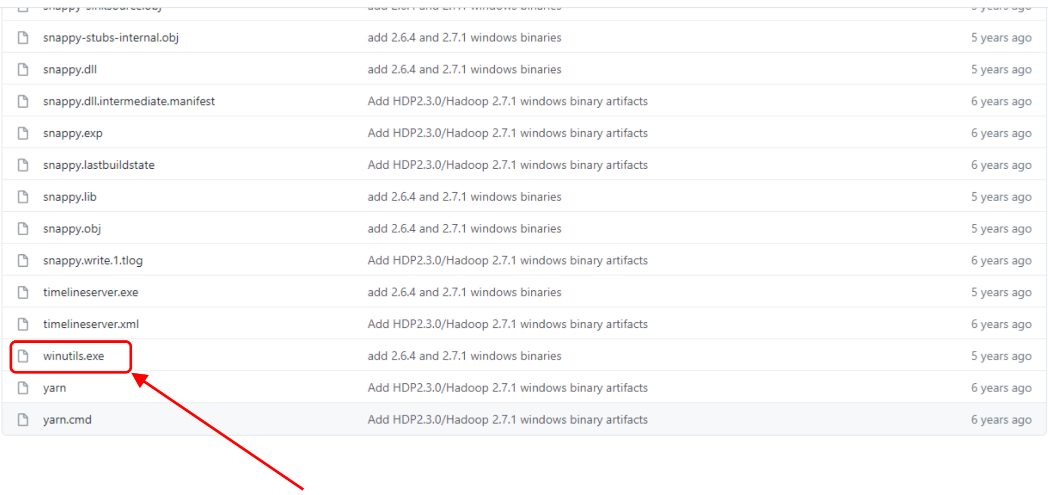
README.md

These are actually the binaries off HDP 2.3.0; they should be interchangeble with the ASF 2.7.1 release

(usual disclaimers etc: if you want the artifacts direct you can download the whole install from hortonworks.com.

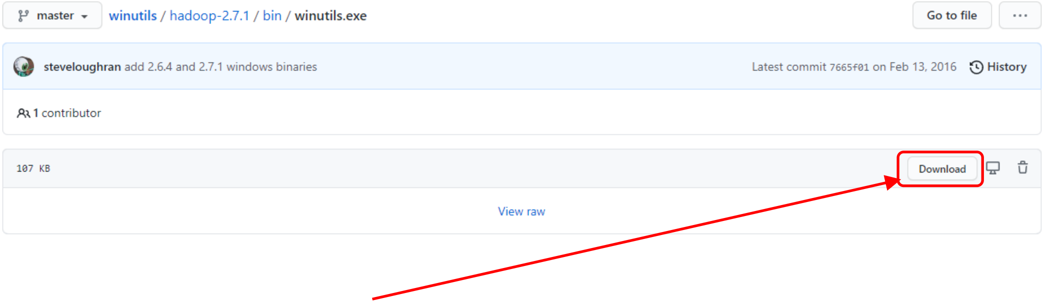
This is just what ended up in my hdp/bin dir after the installation -though I have deleted the pdb files.

- ○ Click the link ‘bin’ to navigate to the next page.



snappy-stubs-internal.obj	add 2.6.4 and 2.7.1 windows binaries	5 years ago
snappy.dll	add 2.6.4 and 2.7.1 windows binaries	5 years ago
snappy.dll.intermediate.manifest	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
snappy.exp	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
snappy.lastbuildstate	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
snappy.lib	add 2.6.4 and 2.7.1 windows binaries	5 years ago
snappy.obj	add 2.6.4 and 2.7.1 windows binaries	5 years ago
snappy.write.1.tlog	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
timelineserver.exe	add 2.6.4 and 2.7.1 windows binaries	5 years ago
timelineserver.xml	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
winutils.exe	add 2.6.4 and 2.7.1 windows binaries	5 years ago
yarn	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago
yarn.cmd	Add HDP2.3.0/Hadoop 2.7.1 windows binary artifacts	6 years ago

- ○ Click the link ‘winutils.exe’ to navigate to the next page.



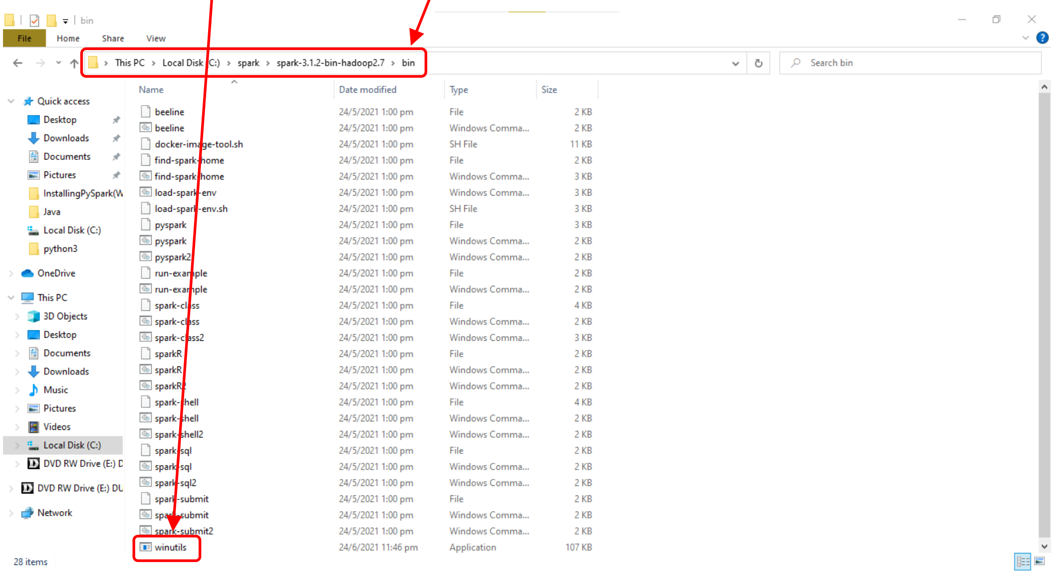
steveloughran add 2.6.4 and 2.7.1 windows binaries Latest commit 7665f01 on Feb 13, 2016 History

1 contributor

107 KB

- ○ Click the button ‘Download’ to start the download.
- ○ The downloaded ‘winutils.exe’ need to be placed in the bin directory of the spark folders ‘C:\spark\spark-3.1.2-bin-hadoop2.7.’.

- The ‘winutils.exe’ is placed here.

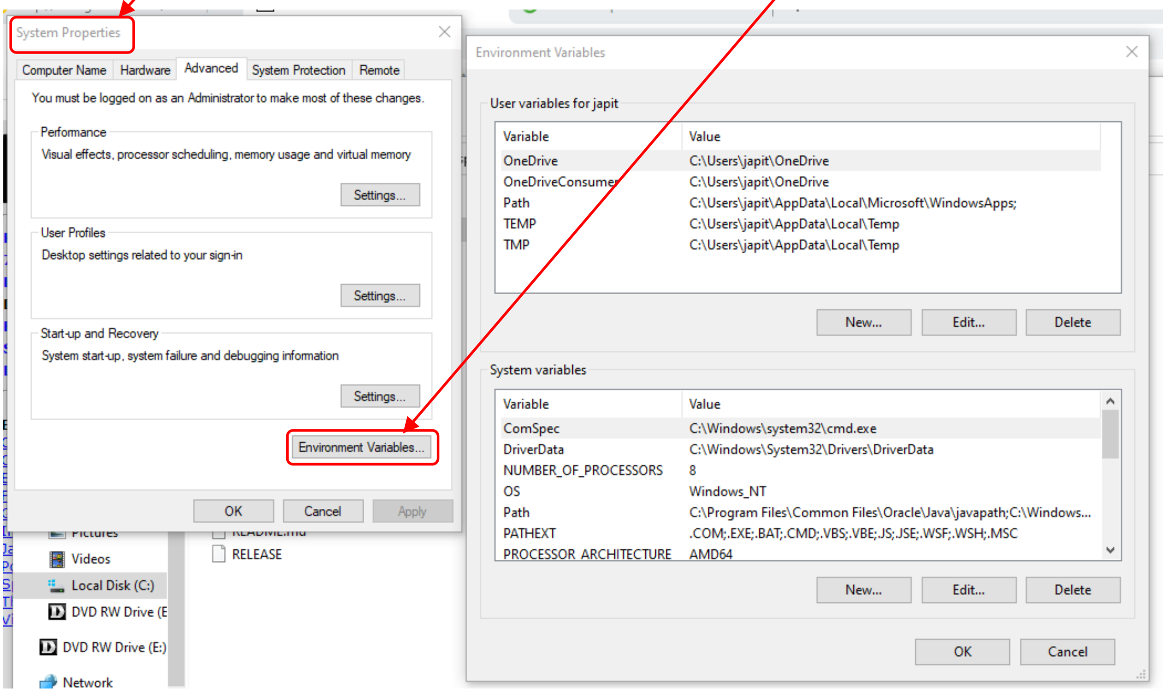


This PC › Local Disk (C:) › spark › spark-3.1.2-bin-hadoop2.7 › bin

Name	Date modified	Type	Size
beeline	24/5/2021 1:00 pm	File	2 KB
beeline	24/5/2021 1:00 pm	Windows Comma...	2 KB
docker-image-tool.sh	24/5/2021 1:00 pm	SH File	11 KB
find-spark-home	24/5/2021 1:00 pm	File	2 KB
find-spark-home	24/5/2021 1:00 pm	Windows Comma...	3 KB
load-spark-env	24/5/2021 1:00 pm	Windows Comma...	3 KB
load-spark-env.sh	24/5/2021 1:00 pm	SH File	3 KB
pyspark	24/5/2021 1:00 pm	File	3 KB
pyspark	24/5/2021 1:00 pm	Windows Comma...	2 KB
pyspark2	24/5/2021 1:00 pm	Windows Comma...	2 KB
run-example	24/5/2021 1:00 pm	File	2 KB
run-example	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-class	24/5/2021 1:00 pm	File	4 KB
spark-class	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-class2	24/5/2021 1:00 pm	Windows Comma...	3 KB
sparkR	24/5/2021 1:00 pm	File	2 KB
sparkR	24/5/2021 1:00 pm	Windows Comma...	2 KB
sparkR2	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-shell	24/5/2021 1:00 pm	File	4 KB
spark-shell	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-shell2	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-sql	24/5/2021 1:00 pm	File	2 KB
spark-sql	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-sql2	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-submit	24/5/2021 1:00 pm	File	2 KB
spark-submit	24/5/2021 1:00 pm	Windows Comma...	2 KB
spark-submit2	24/5/2021 1:00 pm	Windows Comma...	2 KB
winutils	24/6/2021 11:46 pm	Application	107 KB

28 items

- **Setting of user’s environment variable**
- In order for the system to know where the various components (packages) are installed, we need to add the path (directory) of each component to the user’s environment variables. These variables let Windows find where the files are when we start the PySpark kernel.
- You can find the environment variable settings by typing “environment…” in the search box. When the System Properties windows is shown, click on the ‘Environment Variables…’ button to proceed.



System Properties

Computer Name | Hardware | Advanced | System Protection | Remote

You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage and virtual memory

User Profiles

Desktop settings related to your sign-in

Start-up and Recovery

System start-up, system failure and debugging information

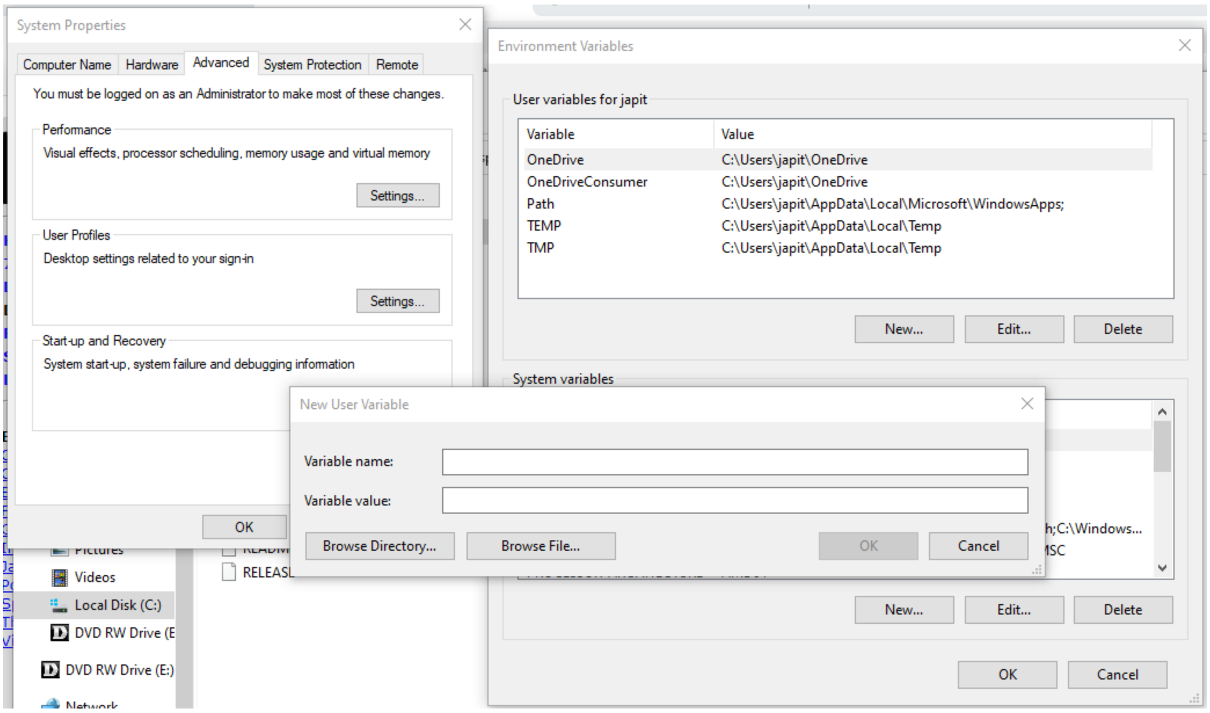
Environment Variables

User variables for japit

Variable	Value
OneDrive	C:\Users\japit\OneDrive
OneDriveConsumer	C:\Users\japit\OneDrive
Path	C:\Users\japit\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\japit\AppData\Local\Temp
TMP	C:\Users\japit\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Windows...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64



System Properties

You must be logged on as an Administrator to make most of these changes.

Performance: Visual effects, processor scheduling, memory usage and virtual memory

User Profiles: Desktop settings related to your sign-in

Start-up and Recovery: System start-up, system failure and debugging information

Environment Variables

User variables for japit

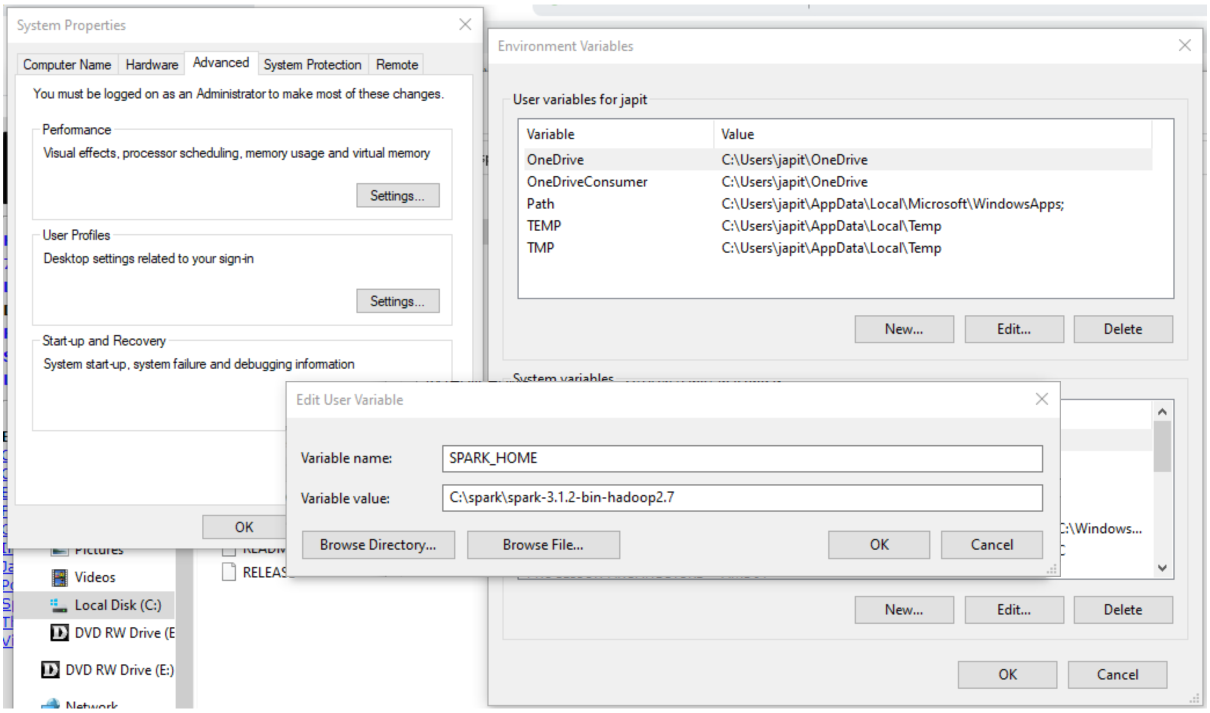
Variable	Value
OneDrive	C:\Users\japit\OneDrive
OneDriveConsumer	C:\Users\japit\OneDrive
Path	C:\Users\japit\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\japit\AppData\Local\Temp
TMP	C:\Users\japit\AppData\Local\Temp

System variables

New User Variable

Variable name:

Variable value:



System Properties

You must be logged on as an Administrator to make most of these changes.

Performance: Visual effects, processor scheduling, memory usage and virtual memory

User Profiles: Desktop settings related to your sign-in

Start-up and Recovery: System start-up, system failure and debugging information

Environment Variables

User variables for japit

Variable	Value
OneDrive	C:\Users\japit\OneDrive
OneDriveConsumer	C:\Users\japit\OneDrive
Path	C:\Users\japit\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\japit\AppData\Local\Temp
TMP	C:\Users\japit\AppData\Local\Temp

System variables

Edit User Variable

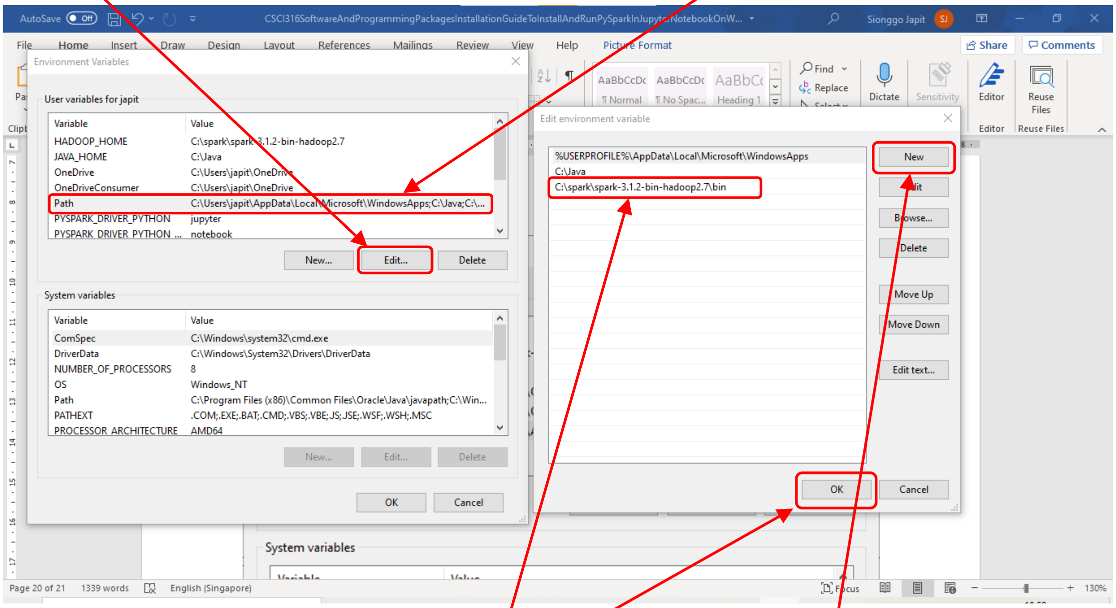
Variable name: SPARK_HOME

Variable value: C:\spark\spark-3.1.2-bin-hadoop2.7

- Do the same for the rest of the components shown below. Note that the values shown here are according to my setup (the directories where I place the packages.) You need to follow according to your setup.

Name	Values
SPARK_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
HADOOP_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
PYSPARK_DRIVER_PYTHON	Jupyter
PYSPARK_DRIVER_PYTHON_OPTS	notebook

- You also need to set the path where you unpacked your spark to the system's path variable. In the same environment variable settings window, look for the 'Path' variable and click the 'Edit…' button to proceed.



User variables for japit

Variable	Value
HADOOP_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
JAVA_HOME	C:\Java
OneDrive	C:\Users\japit\OneDrive
OneDriveConsumer	C:\Users\japit\OneDrive
Path	C:\Users\japit\AppData\Local\Microsoft\WindowsApps;C:\Java;C:\...
PYSPARK_DRIVER_PYTHON	jupyter
PYSPARK_DRIVER_PYTHON ...	notebook

System variables

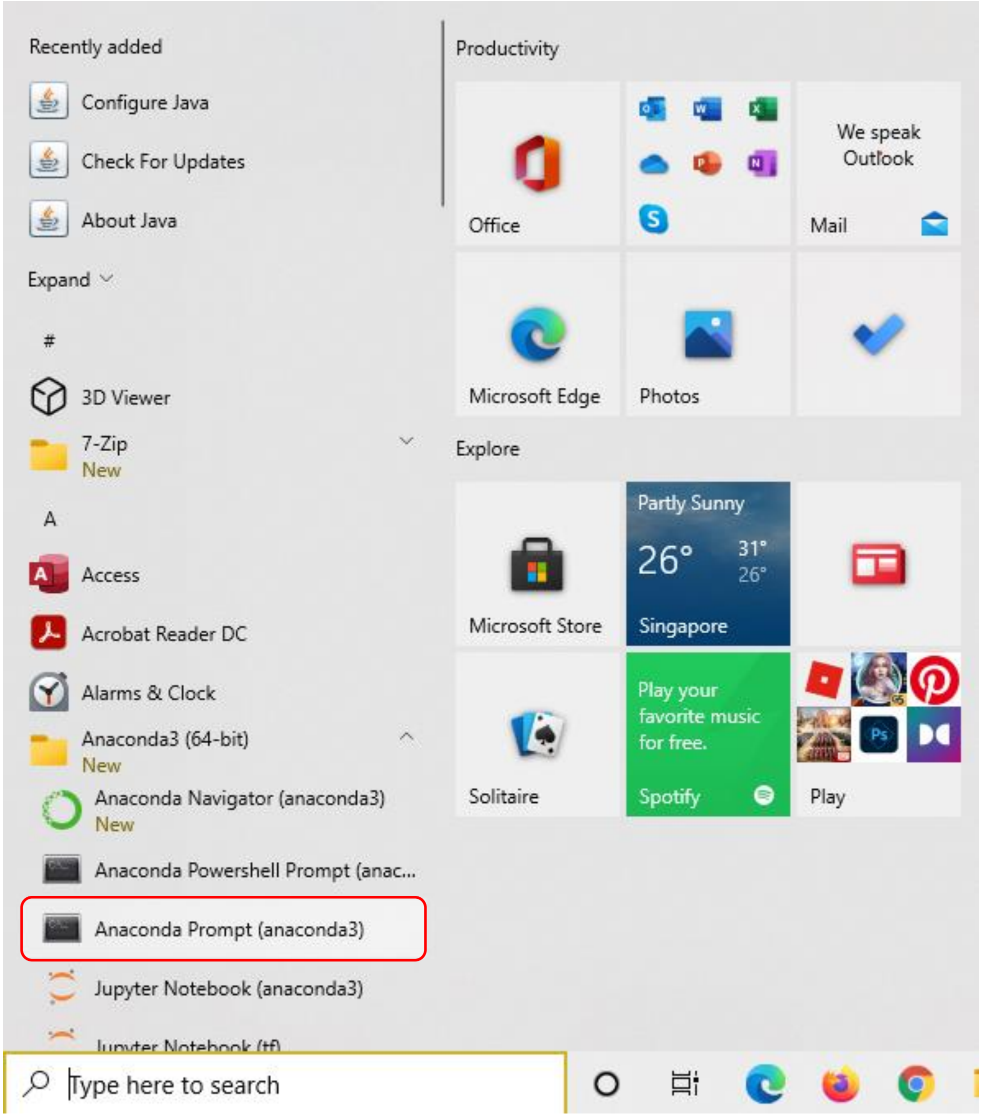
Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Java
C:\spark\spark-3.1.2-bin-hadoop2.7\bin

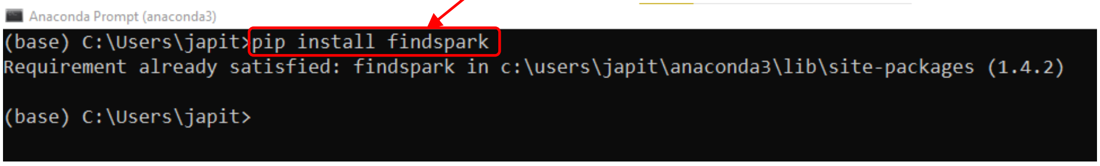
- In the next 'Edit environment variables' pop-up window, click the 'New' button, and enter/add the path where you unpacked your spark packages follows by a '\bin'. For example, I unpacked my spark in C:\spark\spark-3.1.2-bin-hadoop2.7. In this path, there is one directory where the binaries are stored. I will add that directory to the path, hence, my path will be 'C:\spark\spark-3.1.2-bin-hadoop2.7\bin'.

  Click the 'OK' button to save your entry.

- By now, I suppose the installation of Anaconda may have been completed. Check that the Anaconda installation is indeed completed.
- Next, we need install findspark, pyspark, and other required libraries for CSCI316. We will use Anaconda to do so:
  i. Open the Anaconda prompt.



  ii. Installing findspark
  - At the prompt, type **'pip install findspark'** followed with a 'return' key.

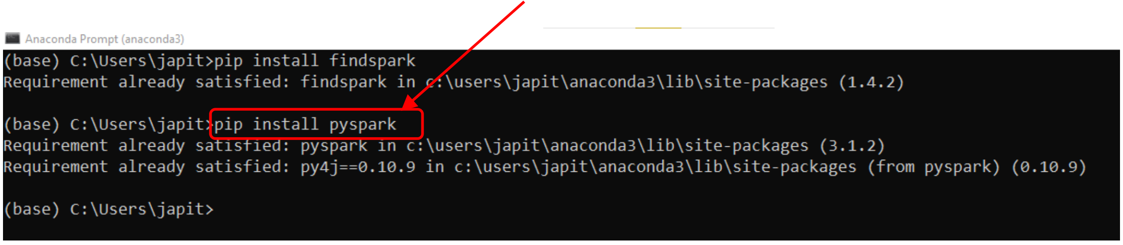


```
(base) C:\Users\japit>pip install findspark
Requirement already satisfied: findspark in c:\users\japit\anaconda3\lib\site-packages (1.4.2)

(base) C:\Users\japit>
```

Since I have installed my 'findpark' earlier, the system will indicate that the requirement has already been satisfied, otherwise, the system will install the 'findpark' for you.

iii. Installing pyspark

- ➢ At the prompt, type '**pip install pyspark**' followed with a 'return' key.

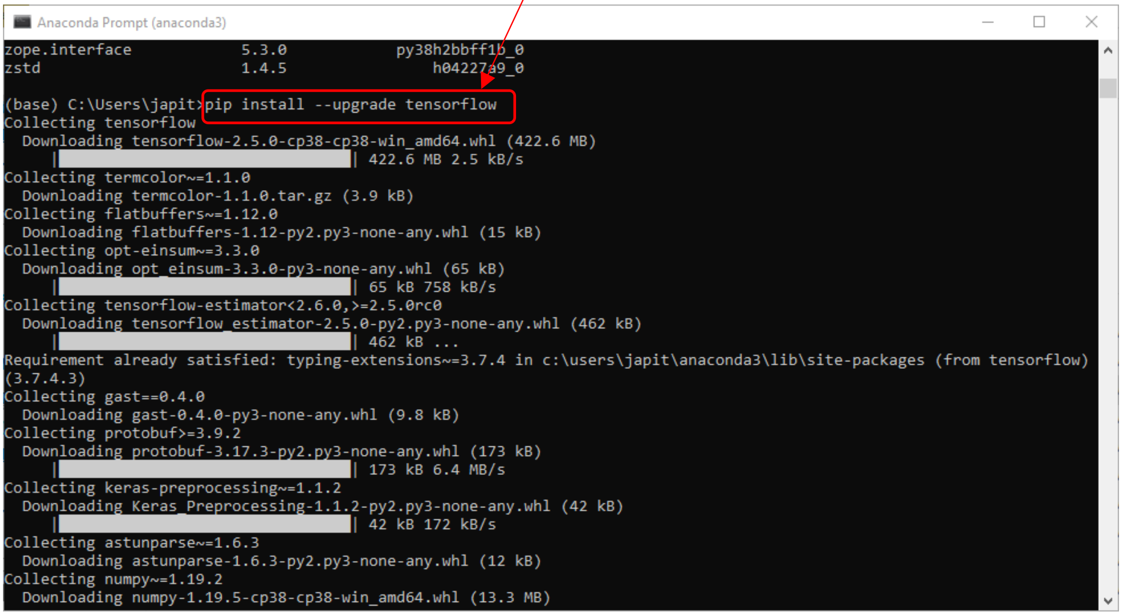


```
(base) C:\Users\japit>pip install findspark
Requirement already satisfied: findspark in c:\users\japit\anaconda3\lib\site-packages (1.4.2)

(base) C:\Users\japit>pip install pyspark
Requirement already satisfied: pyspark in c:\users\japit\anaconda3\lib\site-packages (3.1.2)
Requirement already satisfied: py4j==0.10.9 in c:\users\japit\anaconda3\lib\site-packages (from pyspark) (0.10.9)

(base) C:\Users\japit>
```

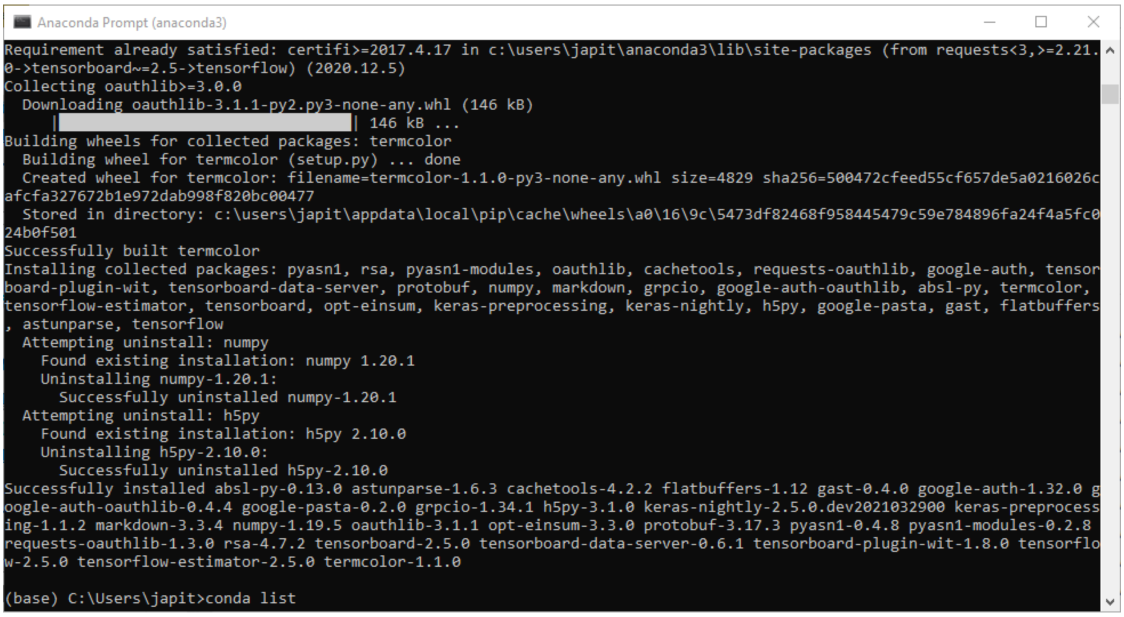
Similarly, since I have installed my 'pyspark' earlier, the system will indicate that the requirement has already been satisfied, otherwise, the system will install the 'pypark' for you.

iv. Installing tensorflow

- ➢ At the prompt, type '**pip install –upgrade tensorflow**' followed with a 'return' key.



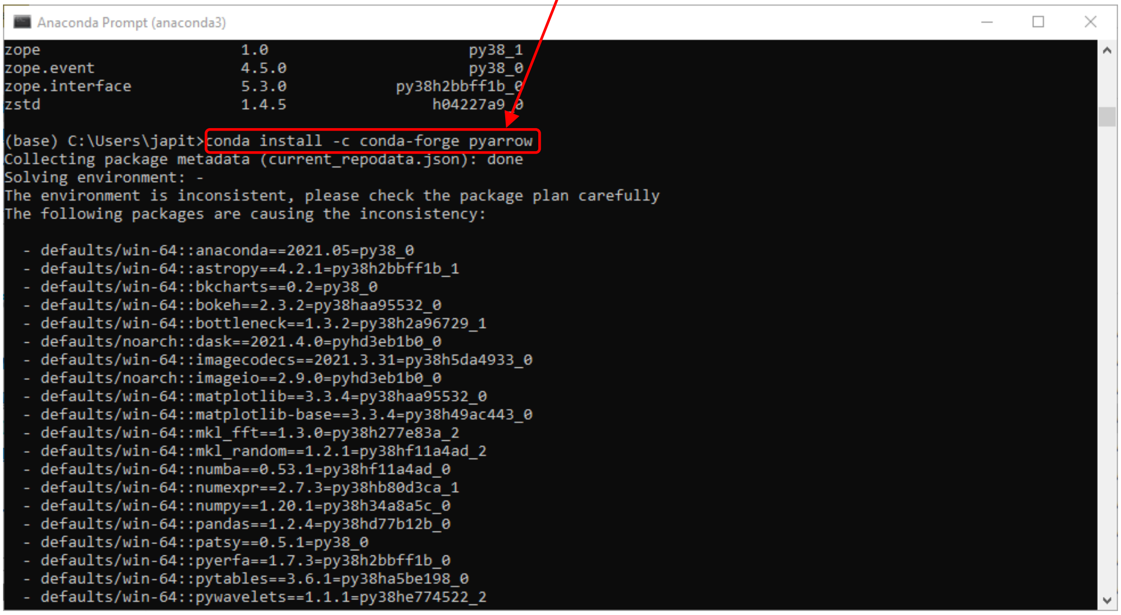
```
zope.interface            5.3.0            py38h2bbff1b_0
zstd                      1.4.5                h04227a9_0

(base) C:\Users\japit>pip install --upgrade tensorflow
Collecting tensorflow
  Downloading tensorflow-2.5.0-cp38-cp38-win_amd64.whl (422.6 MB)
     |████████████████████████████████| 422.6 MB 2.5 kB/s
Collecting termcolor~=1.1.0
  Downloading termcolor-1.1.0.tar.gz (3.9 kB)
Collecting flatbuffers~=1.12.0
  Downloading flatbuffers-1.12-py2.py3-none-any.whl (15 kB)
Collecting opt-einsum~=3.3.0
  Downloading opt_einsum-3.3.0-py3-none-any.whl (65 kB)
     |████████████████████████████████| 65 kB 758 kB/s
Collecting tensorflow-estimator<2.6.0,>=2.5.0rc0
  Downloading tensorflow_estimator-2.5.0-py2.py3-none-any.whl (462 kB)
     |████████████████████████████████| 462 kB ...
Requirement already satisfied: typing-extensions~=3.7.4 in c:\users\japit\anaconda3\lib\site-packages (from tensorflow)
(3.7.4.3)
Collecting gast==0.4.0
  Downloading gast-0.4.0-py3-none-any.whl (9.8 kB)
Collecting protobuf>=3.9.2
  Downloading protobuf-3.17.3-py2.py3-none-any.whl (173 kB)
     |████████████████████████████████| 173 kB 6.4 MB/s
Collecting keras-preprocessing~=1.1.2
  Downloading Keras_Preprocessing-1.1.2-py2.py3-none-any.whl (42 kB)
     |████████████████████████████████| 42 kB 172 kB/s
Collecting astunparse~=1.6.3
  Downloading astunparse-1.6.3-py2.py3-none-any.whl (12 kB)
Collecting numpy~=1.19.2
  Downloading numpy-1.19.5-cp38-cp38-win_amd64.whl (13.3 MB)
```



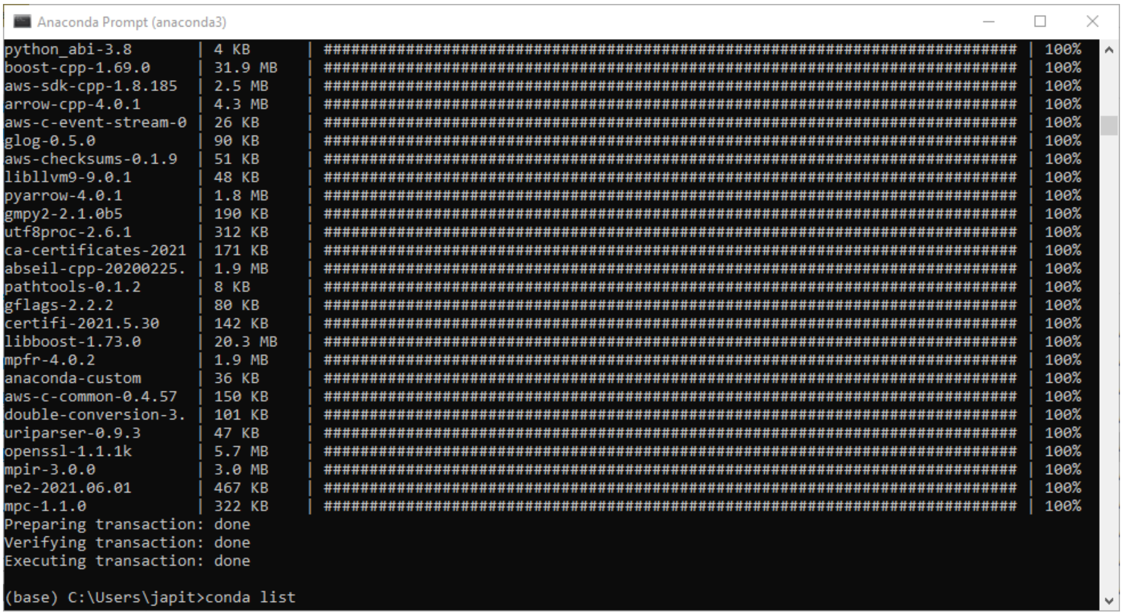
```
Requirement already satisfied: certifi>=2017.4.17 in c:\users\japit\anaconda3\lib\site-packages (from requests<3,>=2.21.
0->tensorboard~=2.5->tensorflow) (2020.12.5)
Collecting oauthlib>=3.0.0
  Downloading oauthlib-3.1.1-py2.py3-none-any.whl (146 kB)
     |████████████████████████████████| 146 kB ...
Building wheels for collected packages: termcolor
  Building wheel for termcolor (setup.py) ... done
  Created wheel for termcolor: filename=termcolor-1.1.0-py3-none-any.whl size=4829 sha256=500472cfeed55cf657de5a0216026c
afcfa327672b1e972dab998f820bc00477
  Stored in directory: c:\users\japit\appdata\local\pip\cache\wheels\a0\16\9c\5473df82468f958445479c59e784896fa24f4a5fc0
24b0f501
Successfully built termcolor
Installing collected packages: pyasn1, rsa, pyasn1-modules, oauthlib, cachetools, requests-oauthlib, google-auth, tensor
board-plugin-wit, tensorboard-data-server, protobuf, numpy, markdown, grpcio, google-auth-oauthlib, absl-py, termcolor,
tensorflow-estimator, tensorboard, opt-einsum, keras-preprocessing, keras-nightly, h5py, google-pasta, gast, flatbuffers
, astunparse, tensorflow
  Attempting uninstall: numpy
    Found existing installation: numpy 1.20.1
    Uninstalling numpy-1.20.1:
      Successfully uninstalled numpy-1.20.1
  Attempting uninstall: h5py
    Found existing installation: h5py 2.10.0
    Uninstalling h5py-2.10.0:
      Successfully uninstalled h5py-2.10.0
Successfully installed absl-py-0.13.0 astunparse-1.6.3 cachetools-4.2.2 flatbuffers-1.12 gast-0.4.0 google-auth-1.32.0 g
oogle-auth-oauthlib-0.4.4 google-pasta-0.2.0 grpcio-1.34.1 h5py-3.1.0 keras-nightly-2.5.0.dev2021032900 keras-preprocess
ing-1.1.2 markdown-3.3.4 numpy-1.19.5 oauthlib-3.1.1 opt-einsum-3.3.0 protobuf-3.17.3 pyasn1-0.4.8 pyasn1-modules-0.2.8
requests-oauthlib-1.3.0 rsa-4.7.2 tensorboard-2.5.0 tensorboard-data-server-0.6.1 tensorboard-plugin-wit-1.8.0 tensorflo
w-2.5.0 tensorflow-estimator-2.5.0 termcolor-1.1.0

(base) C:\Users\japit>conda list
```

v. Installing pyarrow

- At the prompt, type ‘**conda install -c conda-forge pyarrow**’ followed with a ‘return’ key.



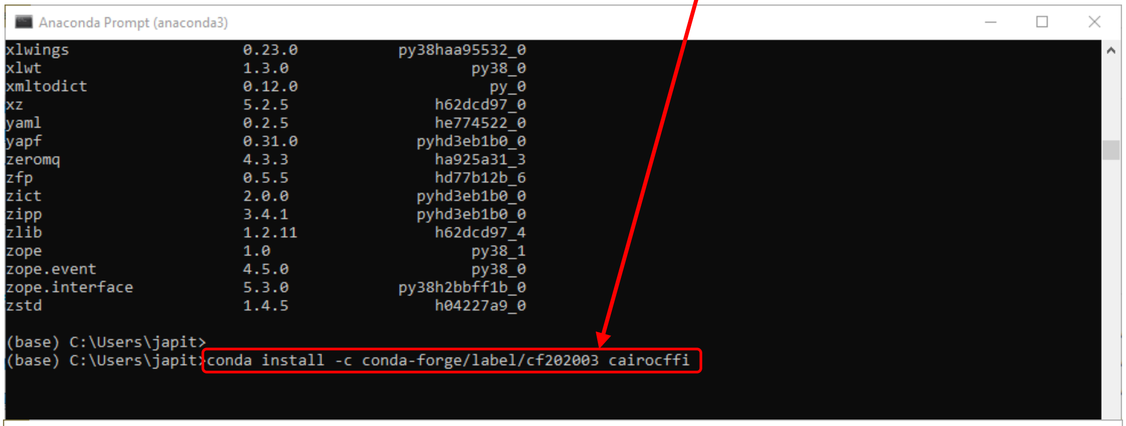
```
zope                      1.0                      py38_1
zope.event                4.5.0                    py38_0
zope.interface            5.3.0            py38h2bbff1b_0
zstd                      1.4.5                h04227a9_0

(base) C:\Users\japit>conda install -c conda-forge pyarrow
Collecting package metadata (current_repodata.json): done
Solving environment: -
The environment is inconsistent, please check the package plan carefully
The following packages are causing the inconsistency:

  - defaults/win-64::anaconda==2021.05=py38_0
  - defaults/win-64::astropy==4.2.1=py38h2bbff1b_1
  - defaults/win-64::bkcharts==0.2=py38_0
  - defaults/win-64::bokeh==2.3.2=py38haa95532_0
  - defaults/win-64::bottleneck==1.3.2=py38h2a96729_1
  - defaults/noarch::dask==2021.4.0=pyhd3eb1b0_0
  - defaults/win-64::imagecodecs==2021.3.31=py38h5da4933_0
  - defaults/noarch::imageio==2.9.0=pyhd3eb1b0_0
  - defaults/win-64::matplotlib==3.3.4=py38haa95532_0
  - defaults/win-64::matplotlib-base==3.3.4=py38h49ac443_0
  - defaults/win-64::mkl_fft==1.3.0=py38h277e83a_2
  - defaults/win-64::mkl_random==1.2.1=py38hf11a4ad_2
  - defaults/win-64::numba==0.53.1=py38hf11a4ad_0
  - defaults/win-64::numexpr==2.7.3=py38hb80d3ca_1
  - defaults/win-64::numpy==1.20.1=py38h34a8a5c_0
  - defaults/win-64::pandas==1.2.4=py38hd77b12b_0
  - defaults/win-64::patsy==0.5.1=py38_0
  - defaults/win-64::pyerfa==1.7.3=py38h2bbff1b_0
  - defaults/win-64::pytables==3.6.1=py38ha5be198_0
  - defaults/win-64::pywavelets==1.1.1=py38he774522_2
```



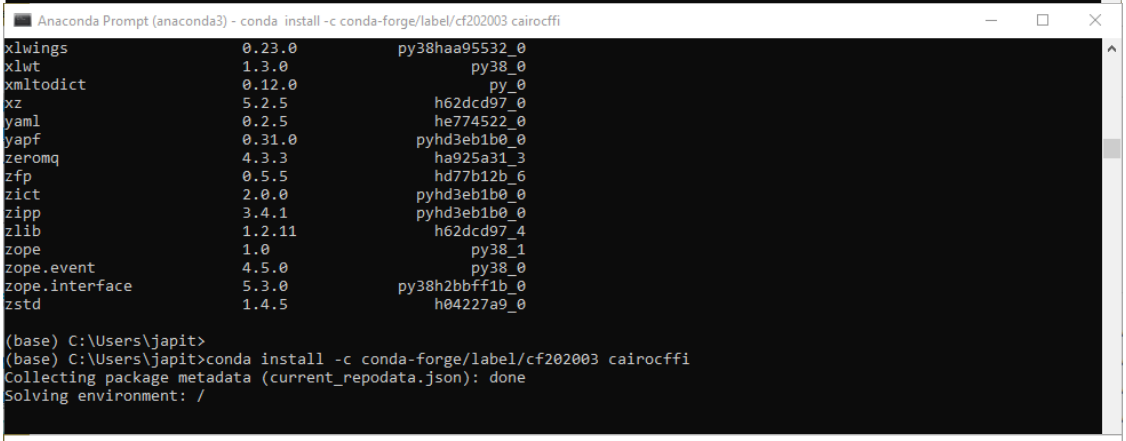
```
python_abi-3.8       | 4 KB      | ############################################################################ | 100%
boost-cpp-1.69.0     | 31.9 MB   | ############################################################################ | 100%
aws-sdk-cpp-1.8.185  | 2.5 MB    | ############################################################################ | 100%
arrow-cpp-4.0.1      | 4.3 MB    | ############################################################################ | 100%
aws-c-event-stream-0 | 26 KB     | ############################################################################ | 100%
glog-0.5.0           | 90 KB     | ############################################################################ | 100%
aws-checksums-0.1.9  | 51 KB     | ############################################################################ | 100%
libllvm9-9.0.1       | 48 KB     | ############################################################################ | 100%
pyarrow-4.0.1        | 1.8 MB    | ############################################################################ | 100%
gmpy2-2.1.0b5        | 190 KB    | ############################################################################ | 100%
utf8proc-2.6.1       | 312 KB    | ############################################################################ | 100%
ca-certificates-2021 | 171 KB    | ############################################################################ | 100%
abseil-cpp-20200225. | 1.9 MB    | ############################################################################ | 100%
pathtools-0.1.2      | 8 KB      | ############################################################################ | 100%
gflags-2.2.2         | 80 KB     | ############################################################################ | 100%
certifi-2021.5.30    | 142 KB    | ############################################################################ | 100%
libboost-1.73.0      | 20.3 MB   | ############################################################################ | 100%
mpfr-4.0.2           | 1.9 MB    | ############################################################################ | 100%
anaconda-custom      | 36 KB     | ############################################################################ | 100%
aws-c-common-0.4.57  | 150 KB    | ############################################################################ | 100%
double-conversion-3. | 101 KB    | ############################################################################ | 100%
uriparser-0.9.3      | 47 KB     | ############################################################################ | 100%
openssl-1.1.1k       | 5.7 MB    | ############################################################################ | 100%
mpir-3.0.0           | 3.0 MB    | ############################################################################ | 100%
re2-2021.06.01       | 467 KB    | ############################################################################ | 100%
mpc-1.1.0            | 322 KB    | ############################################################################ | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(base) C:\Users\japit>conda list
```

vi. Installing cairocffi

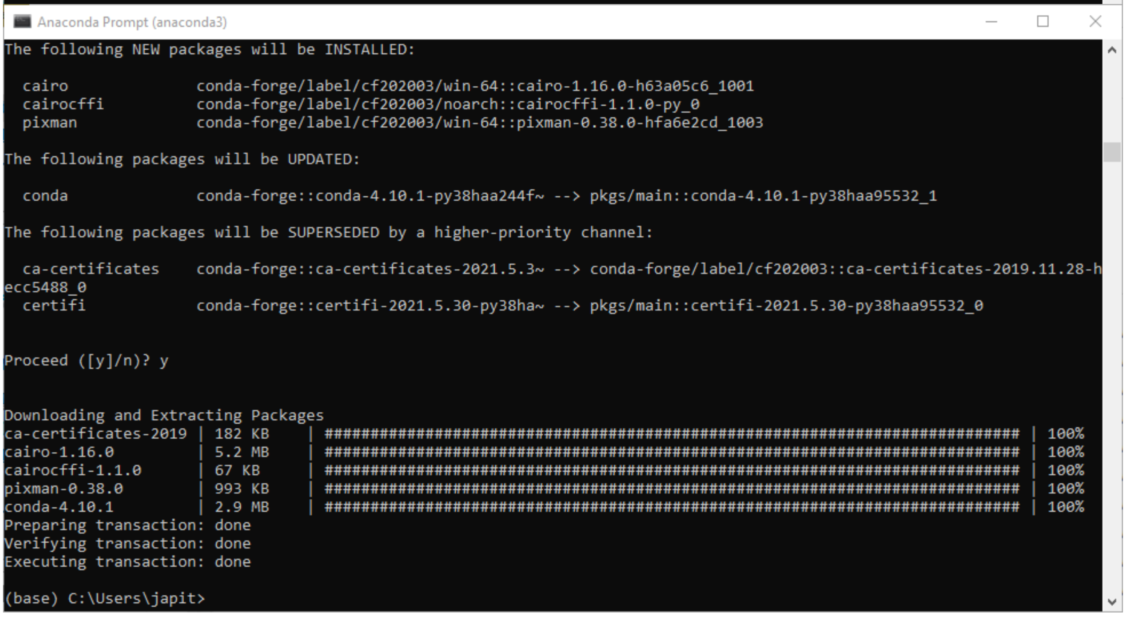
- ➢ At the prompt, type '**conda install -c conda-forge/label/cf202003 cairocffi**' followed with a 'return' key.



```
xlwings                   0.23.0           py38haa95532_0
xlwt                      1.3.0                    py38_0
xmltodict                 0.12.0                     py_0
xz                        5.2.5                h62dcd97_0
yaml                      0.2.5                he774522_0
yapf                      0.31.0             pyhd3eb1b0_0
zeromq                    4.3.3                ha925a31_3
zfp                       0.5.5                hd77b12b_6
zict                      2.0.0              pyhd3eb1b0_0
zipp                      3.4.1              pyhd3eb1b0_0
zlib                      1.2.11               h62dcd97_4
zope                      1.0                      py38_1
zope.event                4.5.0                    py38_0
zope.interface            5.3.0            py38h2bbff1b_0
zstd                      1.4.5                h04227a9_0

(base) C:\Users\japit>
(base) C:\Users\japit>conda install -c conda-forge/label/cf202003 cairocffi
```

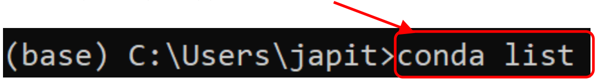

Anaconda Prompt (anaconda3) - conda install -c conda-forge/label/cf202003 cairocffi

```
xlwings                   0.23.0           py38haa95532_0
xlwt                      1.3.0                    py38_0
xmltodict                 0.12.0                     py_0
xz                        5.2.5                h62dcd97_0
yaml                      0.2.5                he774522_0
yapf                      0.31.0             pyhd3eb1b0_0
zeromq                    4.3.3                ha925a31_3
zfp                       0.5.5                hd77b12b_6
zict                      2.0.0              pyhd3eb1b0_0
zipp                      3.4.1              pyhd3eb1b0_0
zlib                      1.2.11               h62dcd97_4
zope                      1.0                      py38_1
zope.event                4.5.0                    py38_0
zope.interface            5.3.0            py38h2bbff1b_0
zstd                      1.4.5                h04227a9_0

(base) C:\Users\japit>
(base) C:\Users\japit>conda install -c conda-forge/label/cf202003 cairocffi
Collecting package metadata (current_repodata.json): done
Solving environment: /
```



```
The following NEW packages will be INSTALLED:

  cairo              conda-forge/label/cf202003/win-64::cairo-1.16.0-h63a05c6_1001
  cairocffi          conda-forge/label/cf202003/noarch::cairocffi-1.1.0-py_0
  pixman             conda-forge/label/cf202003/win-64::pixman-0.38.0-hfa6e2cd_1003

The following packages will be UPDATED:

  conda              conda-forge::conda-4.10.1-py38haa244f~ --> pkgs/main::conda-4.10.1-py38haa95532_1

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates    conda-forge::ca-certificates-2021.5.3~ --> conda-forge/label/cf202003::ca-certificates-2019.11.28-h
ecc5488_0
  certifi            conda-forge::certifi-2021.5.30-py38ha~ --> pkgs/main::certifi-2021.5.30-py38haa95532_0


Proceed ([y]/n)? y


Downloading and Extracting Packages
ca-certificates-2019 | 182 KB    | ############################################################################ | 100%
cairo-1.16.0         | 5.2 MB    | ############################################################################ | 100%
cairocffi-1.1.0      | 67 KB     | ############################################################################ | 100%
pixman-0.38.0        | 993 KB    | ############################################################################ | 100%
conda-4.10.1         | 2.9 MB    | ############################################################################ | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done

(base) C:\Users\japit>
```

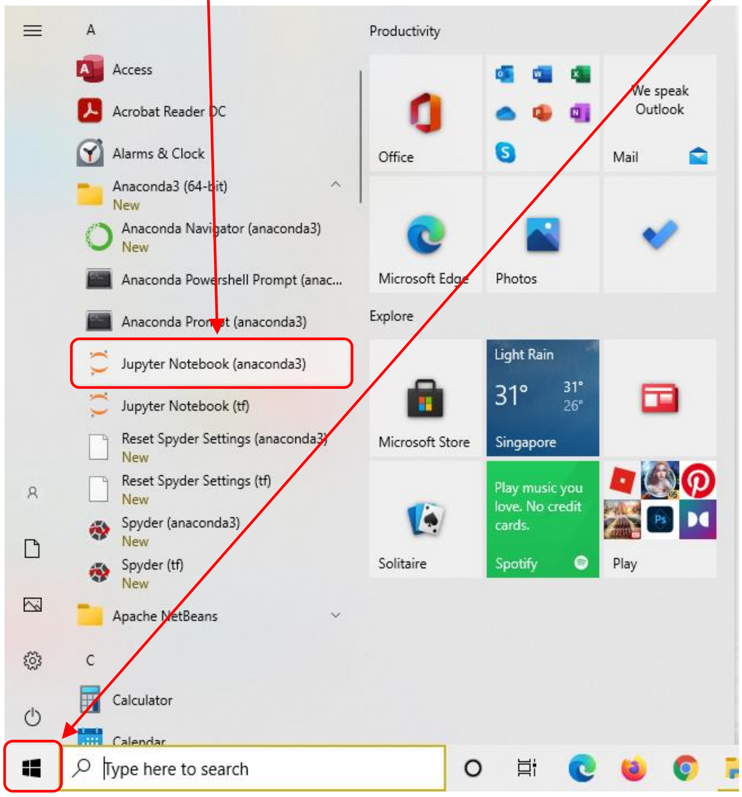
vii. We can now check/verify that all the required software/libraries are installed. To do that at the prompt, type '**conda list**' followed with a 'return' key.



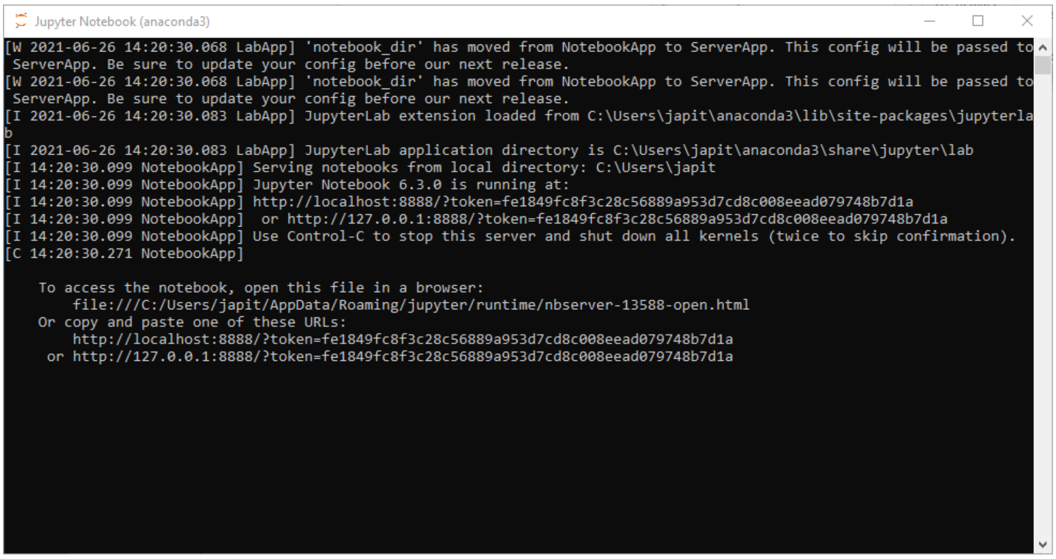
```
(base) C:\Users\japit>conda list
```

C. Running PySpark in Jupyter Notebook

i. To start Jupyter Notebook, open Jupyter Notebook via the windows '**Start**' icon. Click on the lable '**Jupyter Notebook (anaconda3)**' to start Jupyter Notebook.



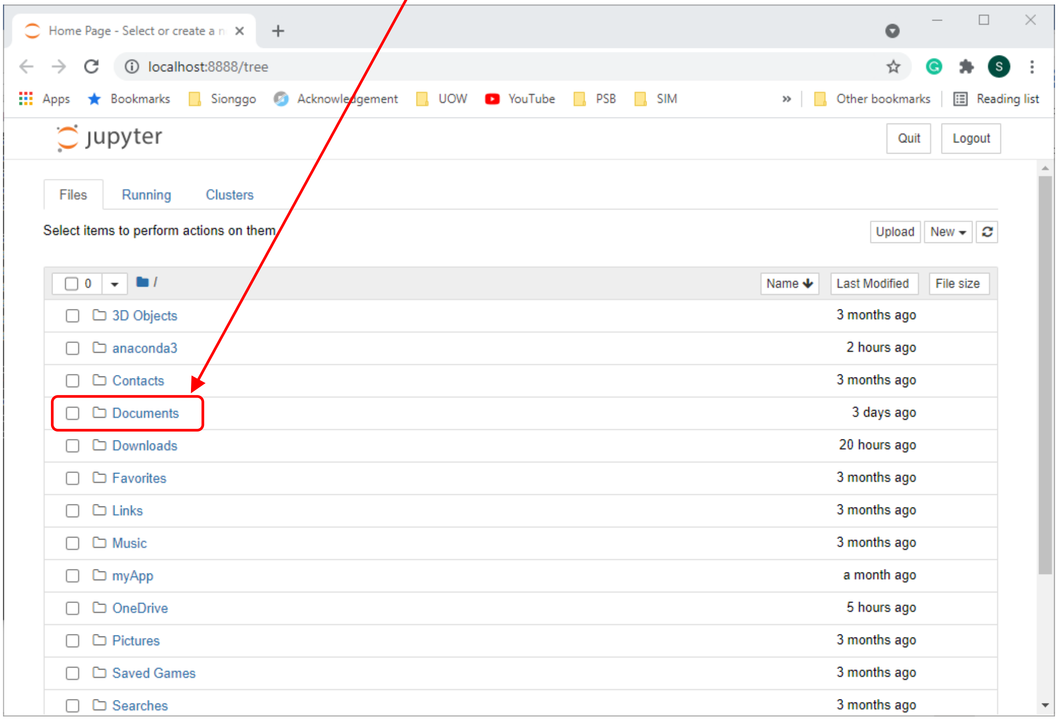
The Jupyter Notebook server is started. Leave this window stays open.



Jupyter Notebook (anaconda3)

```
[W 2021-06-26 14:20:30.068 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to
 ServerApp. Be sure to update your config before our next release.
[W 2021-06-26 14:20:30.068 LabApp] 'notebook_dir' has moved from NotebookApp to ServerApp. This config will be passed to
 ServerApp. Be sure to update your config before our next release.
[I 2021-06-26 14:20:30.083 LabApp] JupyterLab extension loaded from C:\Users\japit\anaconda3\lib\site-packages\jupyterla
b
[I 2021-06-26 14:20:30.083 LabApp] JupyterLab application directory is C:\Users\japit\anaconda3\share\jupyter\lab
[I 14:20:30.099 NotebookApp] Serving notebooks from local directory: C:\Users\japit
[I 14:20:30.099 NotebookApp] Jupyter Notebook 6.3.0 is running at:
[I 14:20:30.099 NotebookApp] http://localhost:8888/?token=fe1849fc8f3c28c56889a953d7cd8c008eead079748b7d1a
[I 14:20:30.099 NotebookApp]  or http://127.0.0.1:8888/?token=fe1849fc8f3c28c56889a953d7cd8c008eead079748b7d1a
[I 14:20:30.099 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 14:20:30.271 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///C:/Users/japit/AppData/Roaming/jupyter/runtime/nbserver-13588-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=fe1849fc8f3c28c56889a953d7cd8c008eead079748b7d1a
     or http://127.0.0.1:8888/?token=fe1849fc8f3c28c56889a953d7cd8c008eead079748b7d1a
```

A Jupyter Notebook client is open. Navigate to the desired working directory. In my setup, I have a directory named **'Documents'** created in my user's name in Windows. I will use this directory as my working directory for CSCI316.

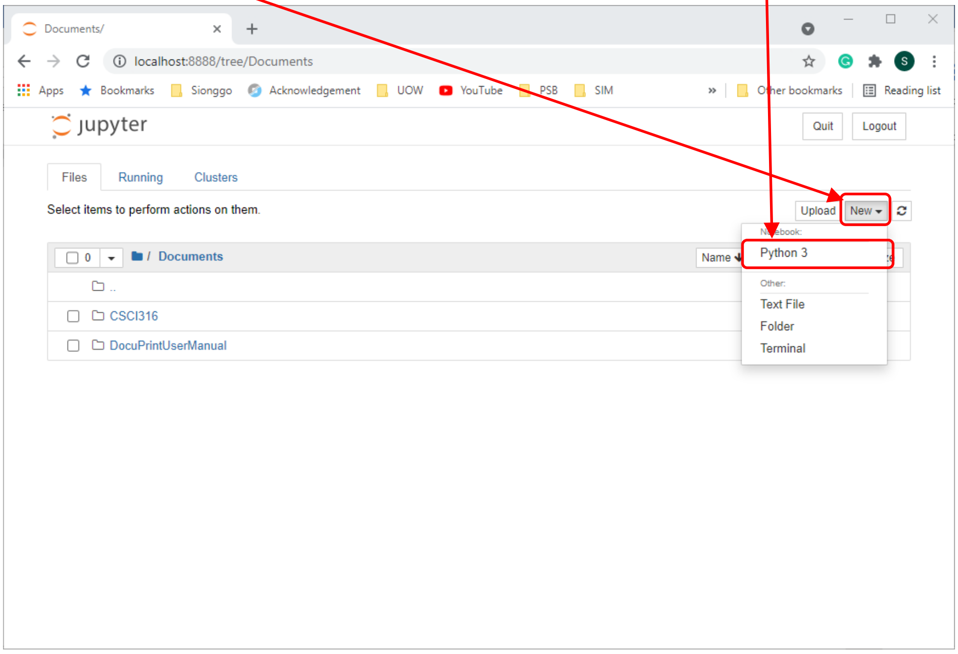


Select items to perform actions on them.

Name	Last Modified
3D Objects	3 months ago
anaconda3	2 hours ago
Contacts	3 months ago
Documents	3 days ago
Downloads	20 hours ago
Favorites	3 months ago
Links	3 months ago
Music	3 months ago
myApp	a month ago
OneDrive	5 hours ago
Pictures	3 months ago
Saved Games	3 months ago
Searches	3 months ago

If you do not have a suitable working directory and want to create a new directory, you can click on the New ▾ icon and use the option to create a new folder. Once the new folder is created, you can rename it to your choice.

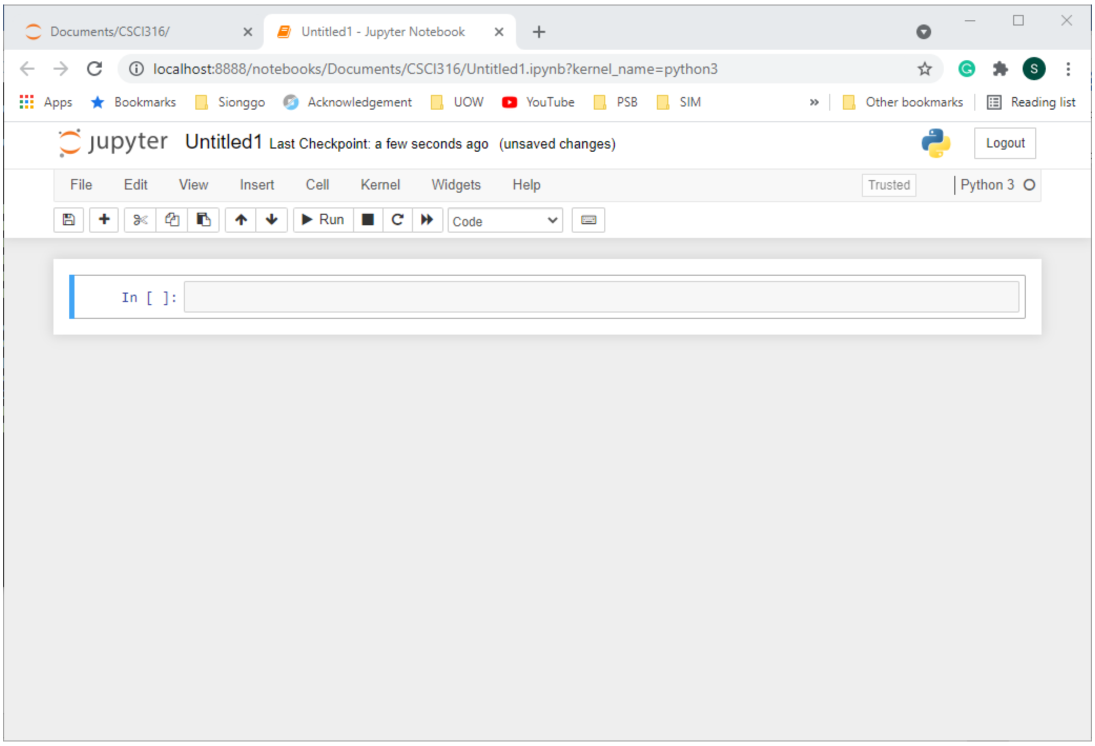
I am now in my working directory. I create a new Python Jupyter Notebook by selecting the '**New**' icon and from the drop-down option, choose '**Python 3**' option.



Select items to perform actions on them.

..
CSCI316
DocuPrintUserManual

A new Jupyter Notebook node is created.

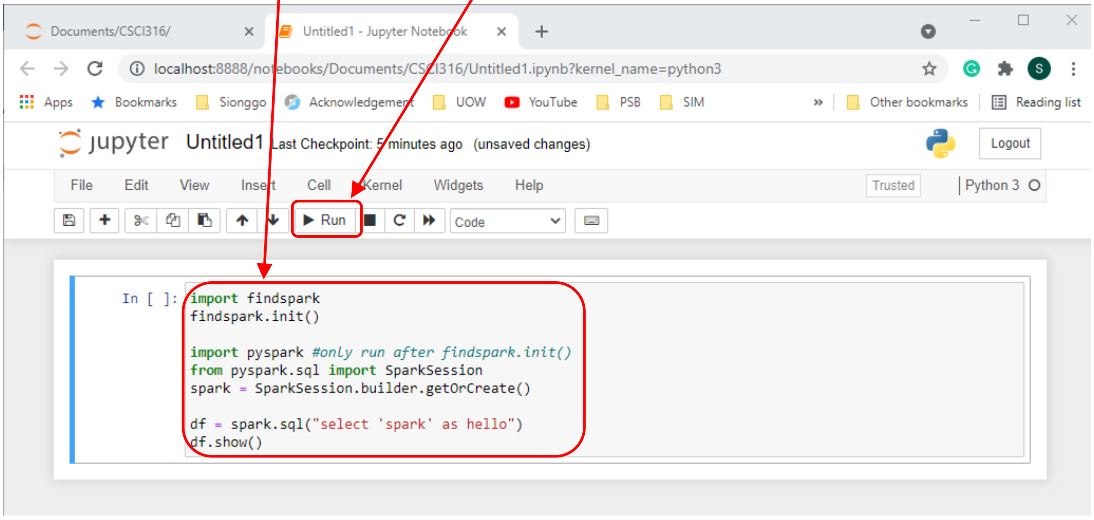


Type the following **codes** in the Jupyter notebook and **run**/execute the code.

import findspark
findspark.init()

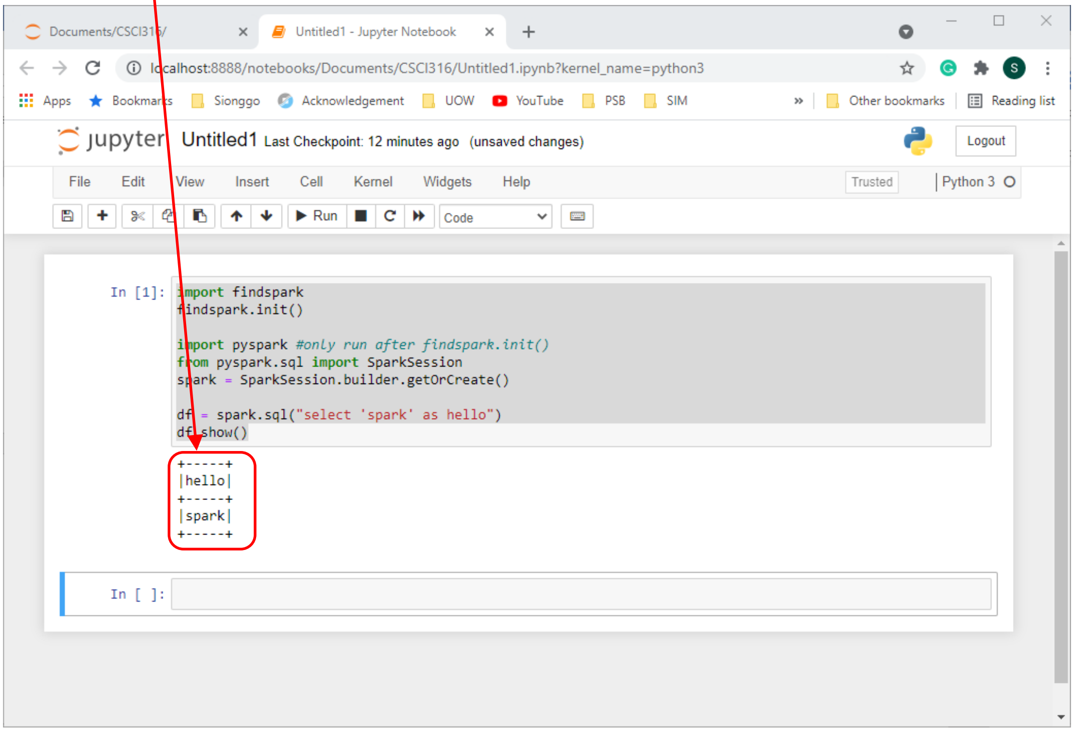
import pyspark #only run after findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

df = spark.sql("select 'spark' as hello")
df.show()



```
import findspark
findspark.init()

import pyspark #only run after findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

df = spark.sql("select 'spark' as hello")
df.show()
```

If you see this, 'Congratulation' you have successfully install PySpark in your system.



In [1]:

```
import findspark
findspark.init()

import pyspark #only run after findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

df = spark.sql("select 'spark' as hello")
df.show()
```

```
+-----+
|hello|
+-----+
|spark|
+-----+
```

In [ ]: