Chapter 1

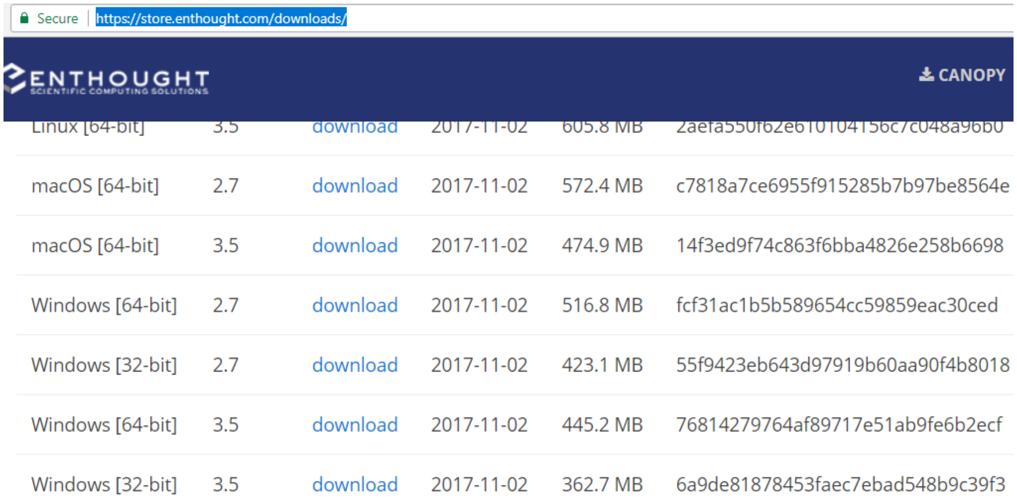
Installation of Software

1. Installation Options

In this chapter, we mainly discuss installation procedures and working with different IDEs. We would demonstrate how to install necessary software to work in Windows 10. At the end of this chapter, we would like to demonstrate the ways to work with Linux by installing Ubuntu Linux in Oracle VirtialBox for host Windows 10. We have incorporated program snippets and snapshots to work with Canopy and Anaconda for Jupyter Notebook.

1.1 Canopy for Windows 10

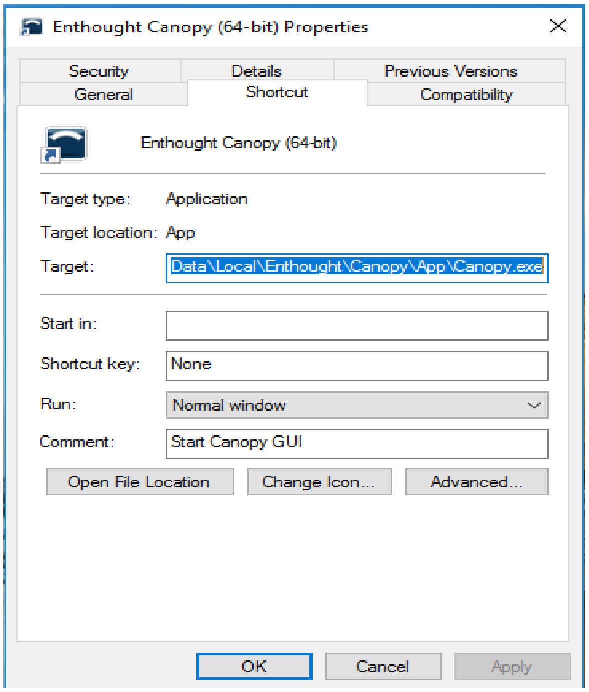
Install Python IDE: Canopy, must be for Python 3.5 or above, 64-bit (I did this for Windows 10, but compatible with Windows 7 without any issue). Download Canopy from the link https://store.enthought.com/downloads/



Secure
https://store.enthought.com/downloads/
ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS
CANOPY

Linux [64-bit]	3.5	download	2017-11-02	605.8 MB	2aefa550f62e61010415 6c7c048a96b0
macOS [64-bit]	2.7	download	2017-11-02	572.4 MB	c7818a7ce6955f915285b7b97be8564e
macOS [64-bit]	3.5	download	2017-11-02	474.9 MB	14f3ed9f74c863f6bba4826e258b6698
Windows [64-bit]	2.7	download	2017-11-02	516.8 MB	fcf31ac1b5b589654cc59859eac30ced
Windows [32-bit]	2.7	download	2017-11-02	423.1 MB	55f9423eb643d97919b60aa90f4b8018
Windows [64-bit]	3.5	download	2017-11-02	445.2 MB	76814279764af89717e51ab9fe6b2ecf
Windows [32-bit]	3.5	download	2017-11-02	362.7 MB	6a9de81878453faec7ebad548b9c39f3

Once installation is finished, you must run this as an administrator: Right click on the Canopy icon displayed on your desktop, and open the Properties from this (as in the picture). Check the box *Run this program as an administrator*.



Enthought Canopy (64-bit)

Target type: Application

Target location: App

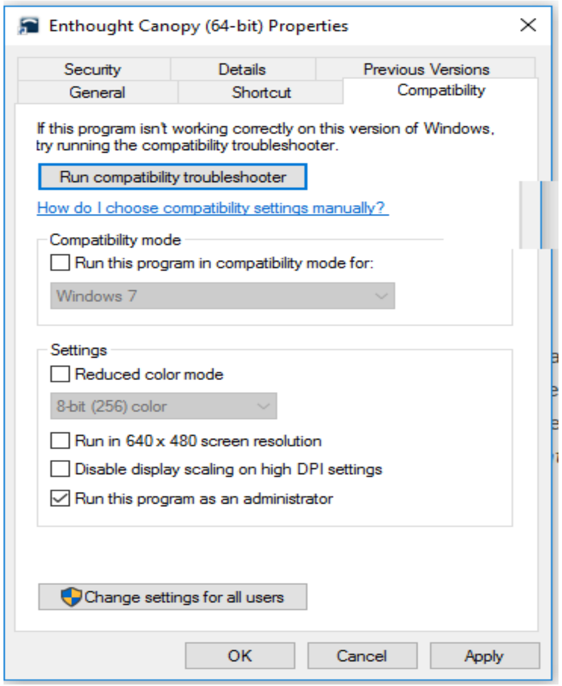
Target: Data\Local\Enthought\Canopy\App\Canopy.exe

Start in:

Shortcut key: None

Run: Normal window

Comment: Start Canopy GUI


Enthought Canopy (64-bit) Properties
Security
Details
Previous Versions
General
Shortcut
Compatibility
Run compatibility troubleshooter
How do I choose compatibility settings manually?
Change settings for all users
OK
Cancel
Apply

If this program isn't working correctly on this version of Windows, try running the compatibility troubleshooter.

Compatibility mode

□ Run this program in compatibility mode for:

Windows 7

Settings

□ Reduced color mode

8-bit (256) color

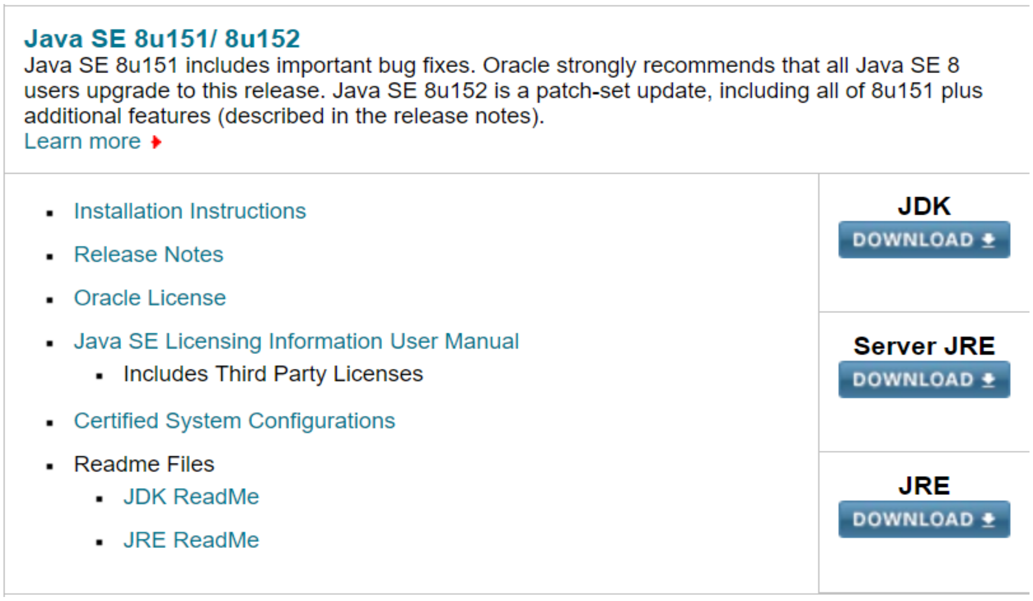
□ Run in 640 x 480 screen resolution

□ Disable display scaling on high DPI settings

☑ Run this program as an administrator

Install JDK: Download and install Java Development Kit, JDK 8. Please do not use JDK 9 as Apache Spark is still not compatible to it. See the JDK in the downloading site http://www.oracle.com/technetwork/java/javase/downloads/index.html

Scroll down to the page and find Java SE 8u151/8u152, and choose to JDK to go to the download page.



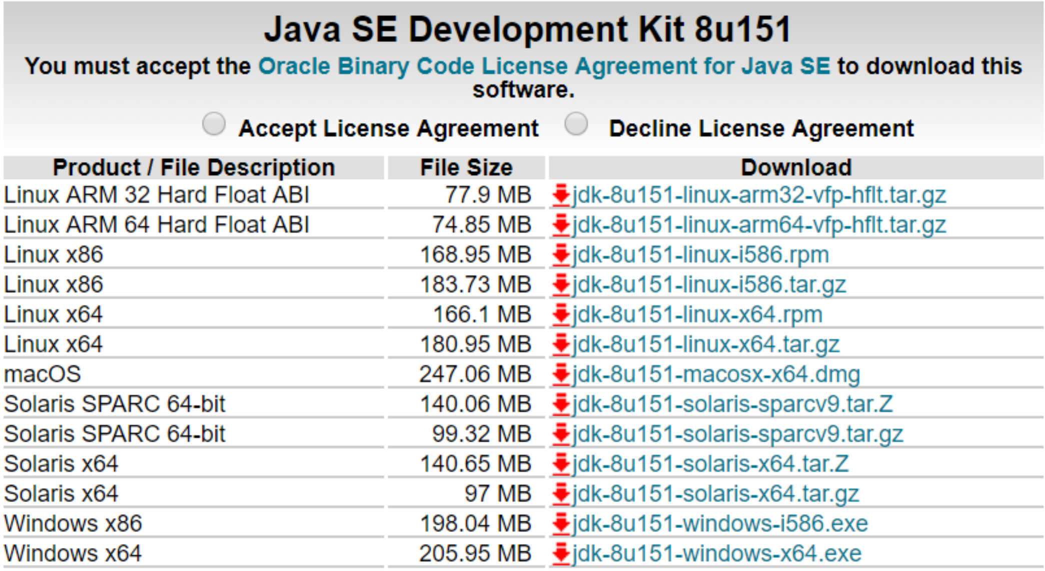
Java SE 8u151/ 8u152

Java SE 8u151 includes important bug fixes. Oracle strongly recommends that all Java SE 8 users upgrade to this release. Java SE 8u152 is a patch-set update, including all of 8u151 plus additional features (described in the release notes).

Learn more

- Installation Instructions
- Release Notes
- Oracle License
- Java SE Licensing Information User Manual
  - Includes Third Party Licenses
- Certified System Configurations
- Readme Files
  - JDK ReadMe
  - JRE ReadMe

Select the option JDK to start download agreement as in the next picture, and *Accept License Agreement* to start downloading (I choose to Windows x64, you may have different choice).



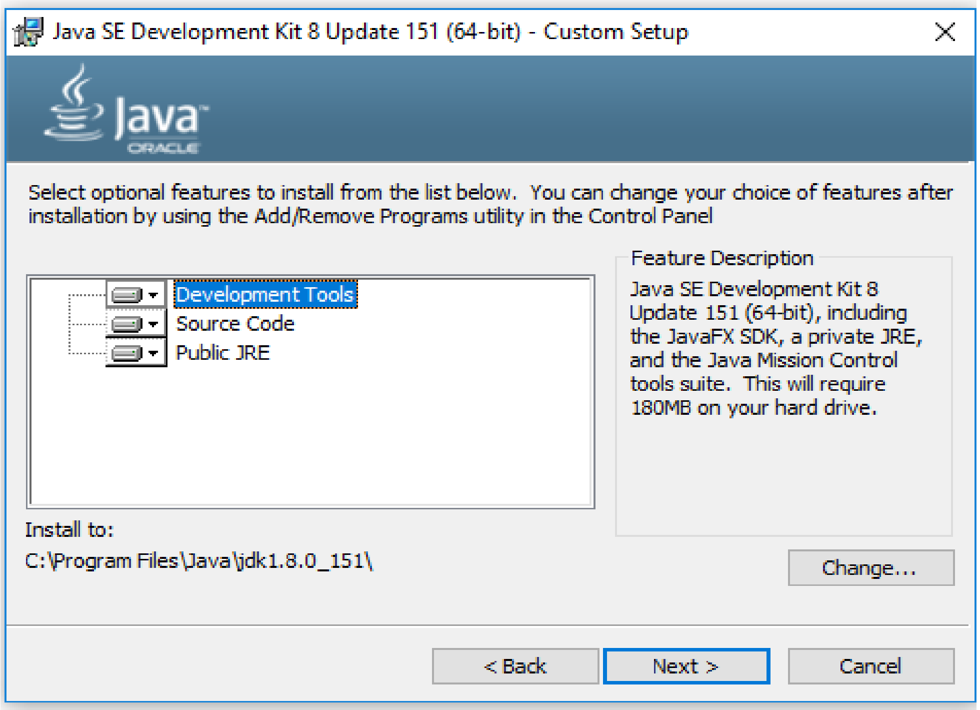
Java SE Development Kit 8u151

You must accept the Oracle Binary Code License Agreement for Java SE to download this software.

○ Accept License Agreement ○ Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	77.9 MB	jdk-8u151-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	74.85 MB	jdk-8u151-linux-arm64-vfp-hflt.tar.gz
Linux x86	168.95 MB	jdk-8u151-linux-i586.rpm
Linux x86	183.73 MB	jdk-8u151-linux-i586.tar.gz
Linux x64	166.1 MB	jdk-8u151-linux-x64.rpm
Linux x64	180.95 MB	jdk-8u151-linux-x64.tar.gz
macOS	247.06 MB	jdk-8u151-macosx-x64.dmg
Solaris SPARC 64-bit	140.06 MB	jdk-8u151-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	99.32 MB	jdk-8u151-solaris-sparcv9.tar.gz
Solaris x64	140.65 MB	jdk-8u151-solaris-x64.tar.Z
Solaris x64	97 MB	jdk-8u151-solaris-x64.tar.gz
Windows x86	198.04 MB	jdk-8u151-windows-i586.exe
Windows x64	205.95 MB	jdk-8u151-windows-x64.exe

Install JDK from the downloaded file. Be careful about the path, default installation goes to C:\Program Files\Java\ and you should redirect the installation to C:\jdk\. To do that click on *Change* as in the pictures below.



Java SE Development Kit 8 Update 151 (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

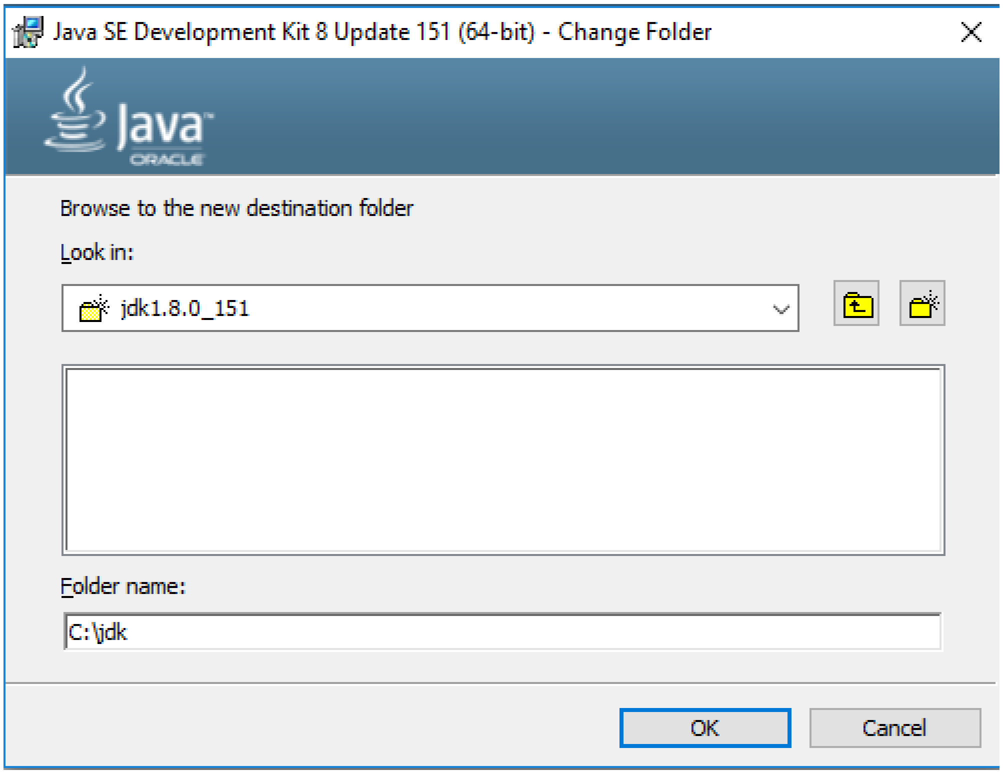
- Development Tools
- Source Code
- Public JRE

Feature Description

Java SE Development Kit 8 Update 151 (64-bit), including the JavaFX SDK, a private JRE, and the Java Mission Control tools suite. This will require 180MB on your hard drive.

Install to:

C:\Program Files\Java\jdk1.8.0_151\

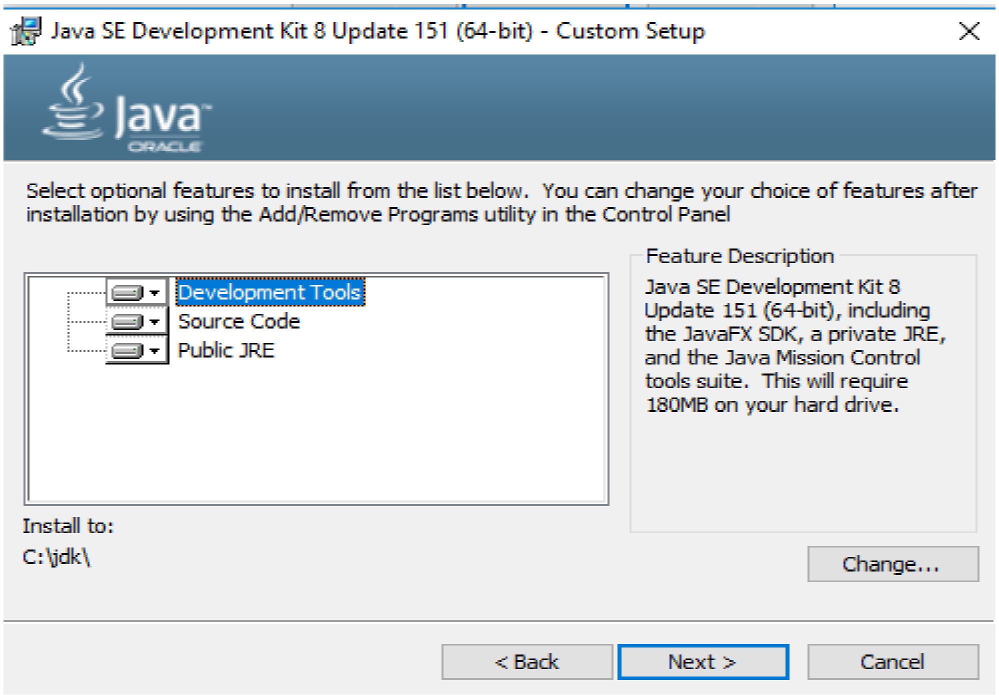


Java SE Development Kit 8 Update 151 (64-bit) - Change Folder

Browse to the new destination folder

Look in: jdk1.8.0_151

Folder name: C:\jdk



Java SE Development Kit 8 Update 151 (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

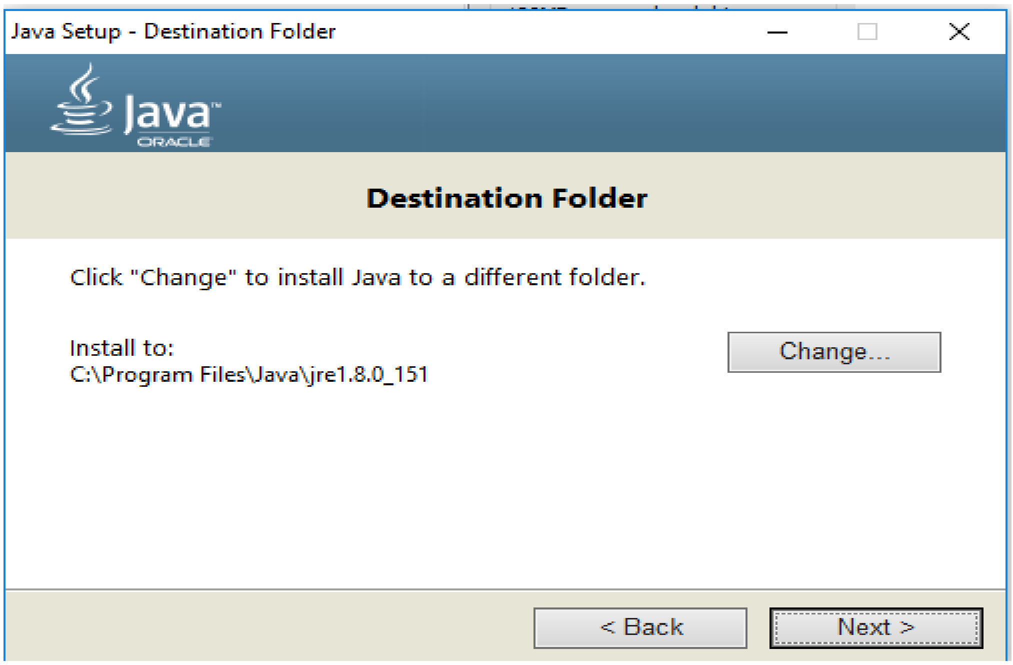
- Development Tools
- Source Code
- Public JRE

Feature Description

Java SE Development Kit 8 Update 151 (64-bit), including the JavaFX SDK, a private JRE, and the Java Mission Control tools suite. This will require 180MB on your hard drive.

Install to:
C:\jdk\

Install this by clicking on *Next*. Then install run Java runtime as can be allerted in the picture below, you should change the destination folder for this installation too.

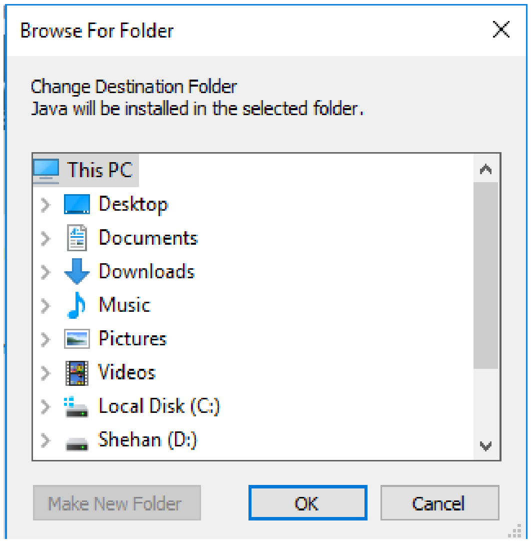


Java Setup - Destination Folder

Destination Folder

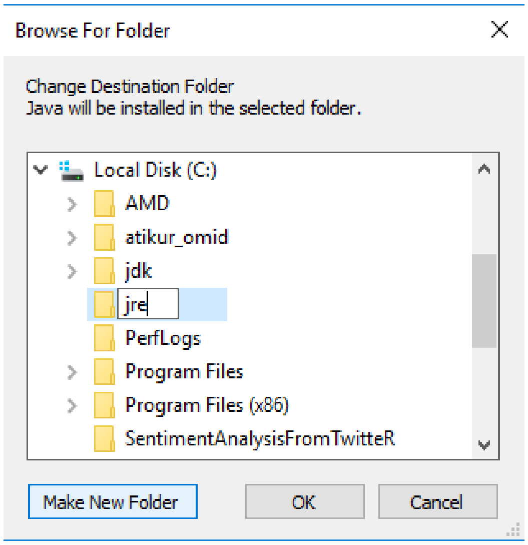
Click "Change" to install Java to a different folder.

Install to:
C:\Program Files\Java\jre1.8.0_151



Browse For Folder

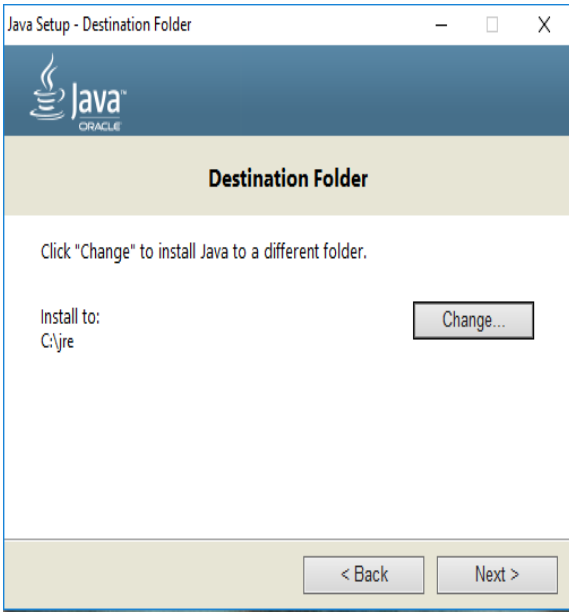
Change Destination Folder
Java will be installed in the selected folder.



Browse For Folder

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

Select the C drive and *Make New Folder*, name the new folder to *jre* and press the OK button. Installation begins as you press the *Next* button (as shown below). See the message *Java SE Development Kit 8 Update 151 (64-bit) Successfully Installed* and then *Close* the window. Installation complete!!!

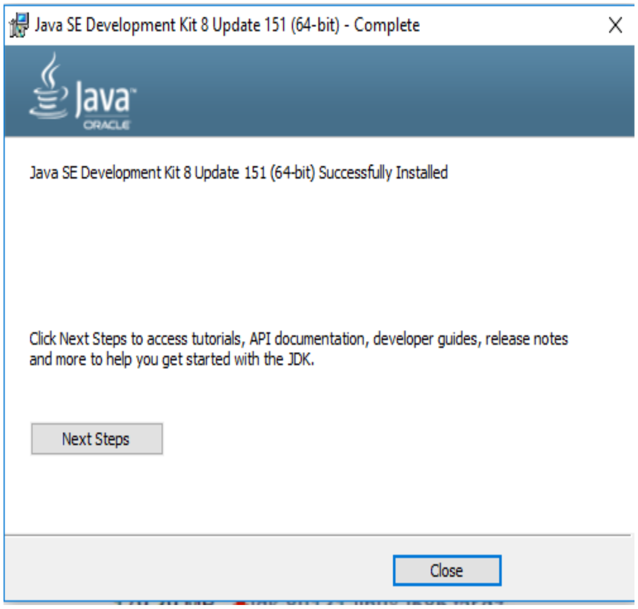


Java Setup - Destination Folder

Destination Folder

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

Install to:
C:\jre

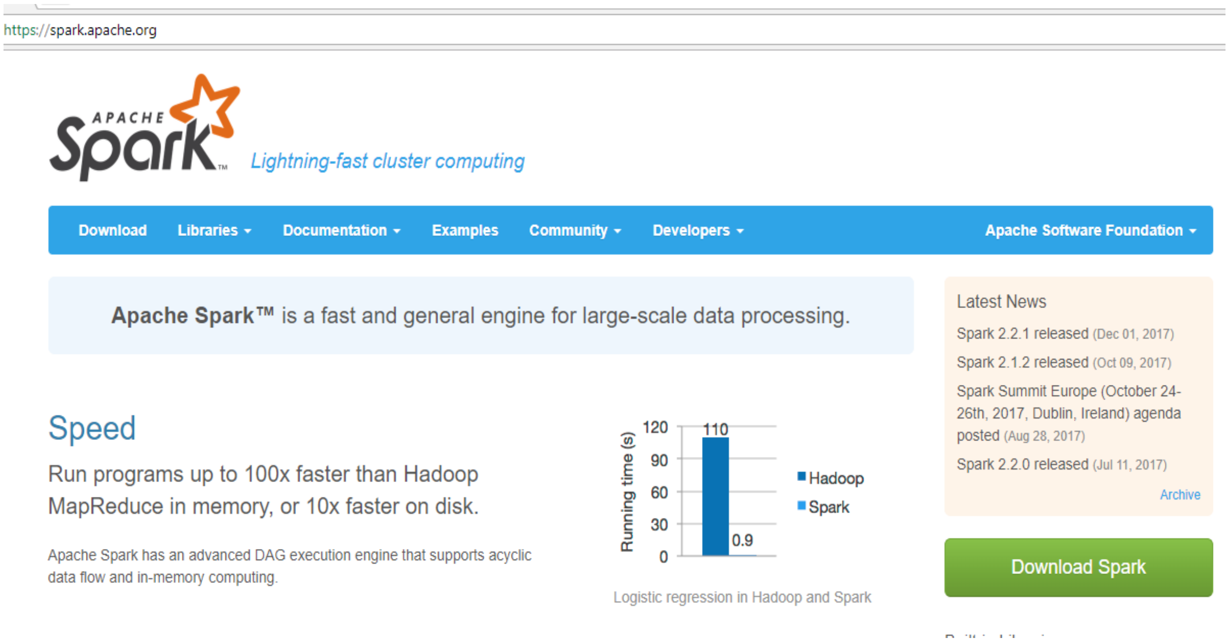


Java SE Development Kit 8 Update 151 (64-bit) - Complete

Java SE Development Kit 8 Update 151 (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.

Install Apache Spark: We have already installed Canopy for Python and JDK. Now we are ready to install Apache Spark. Download from https://spark.apache.org/ through a click on *Download Spark* button (as in the picture).



Apache Spark™ is a fast and general engine for large-scale data processing.

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.

Logistic regression in Hadoop and Spark

Latest News

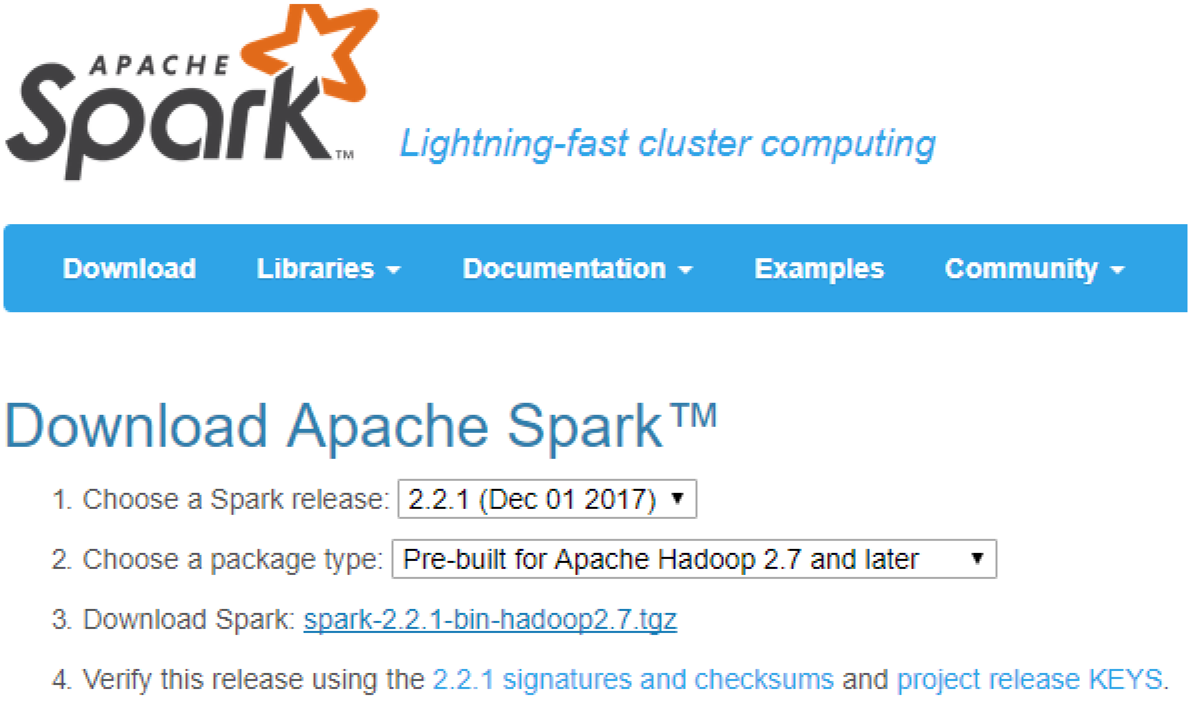
Spark 2.2.1 released (Dec 01, 2017)

Spark 2.1.2 released (Oct 09, 2017)

Spark Summit Europe (October 24-26th, 2017, Dublin, Ireland) agenda posted (Aug 28, 2017)

Spark 2.2.0 released (Jul 11, 2017)

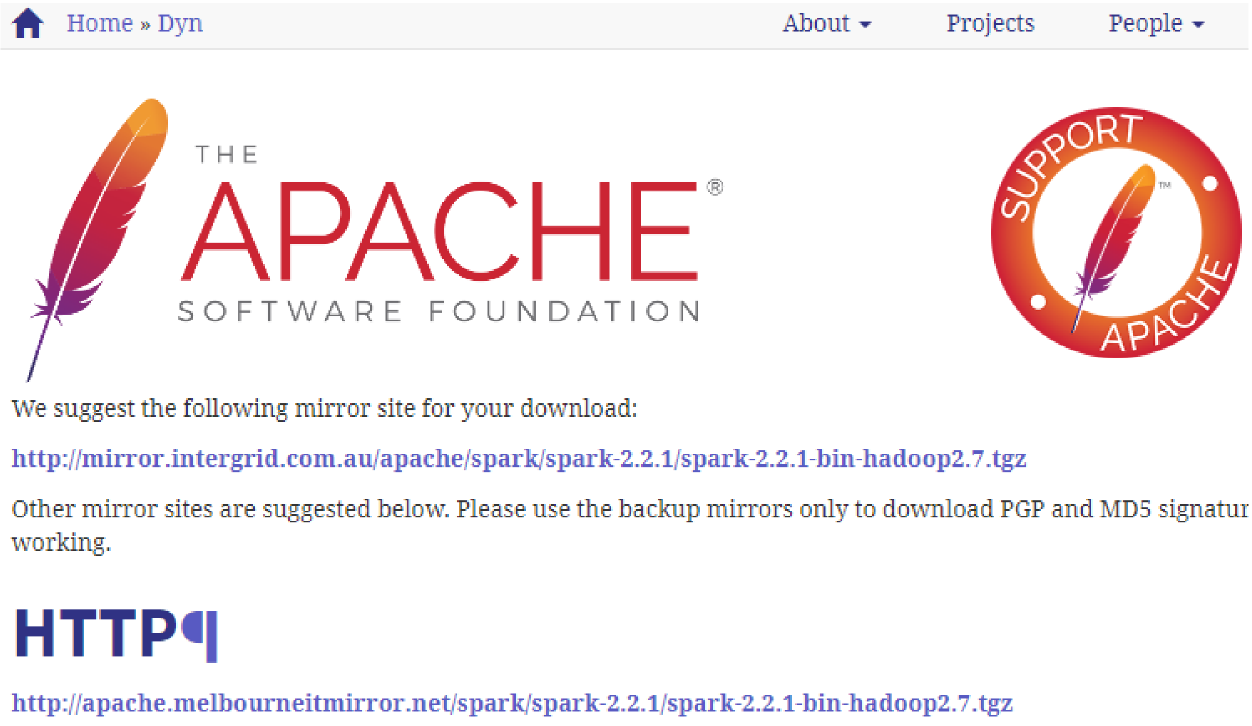
Archive



Download Apache Spark™

1. Choose a Spark release: 2.2.1 (Dec 01 2017) ▾
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later ▾
3. Download Spark: spark-2.2.1-bin-hadoop2.7.tgz
4. Verify this release using the 2.2.1 signatures and checksums and project release KEYS.

Click on spark-2.2.1-bin-hadoop2.7.tgz to download the tgz file.



We suggest the following mirror site for your download:

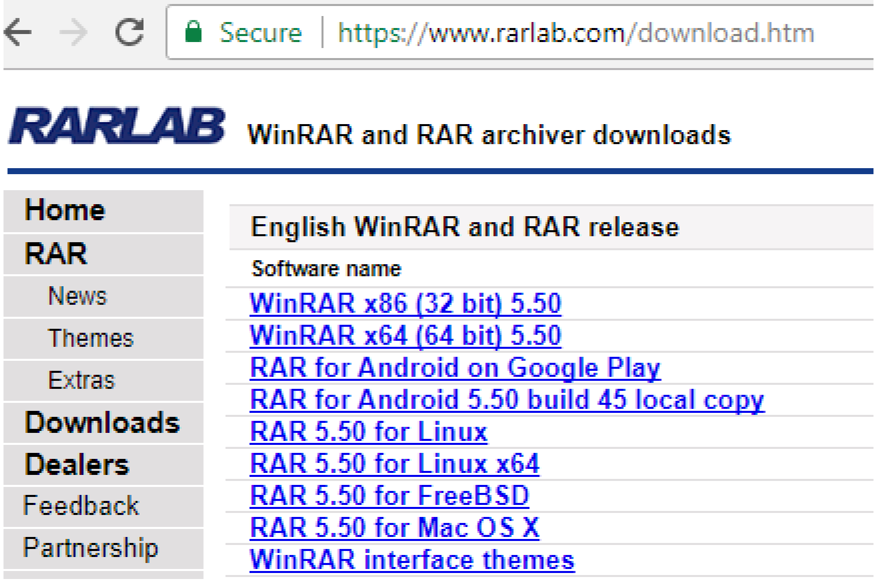
http://mirror.intergrid.com.au/apache/spark/spark-2.2.1/spark-2.2.1-bin-hadoop2.7.tgz

Other mirror sites are suggested below. Please use the backup mirrors only to download PGP and MD5 signatur working.

HTTP¶

http://apache.melbourneitmirror.net/spark/spark-2.2.1/spark-2.2.1-bin-hadoop2.7.tgz

Select the suggested mirror site and the download will start immediately. Keep in mind that the file is in tgz format and we need to use WinRAR to unzip this. So download and install WinRAR from the link in the picture shown below, I have used WinRAR x64 (64 bit) 5.50 for my Windows 10 machine (but can be used with other versions of windows).

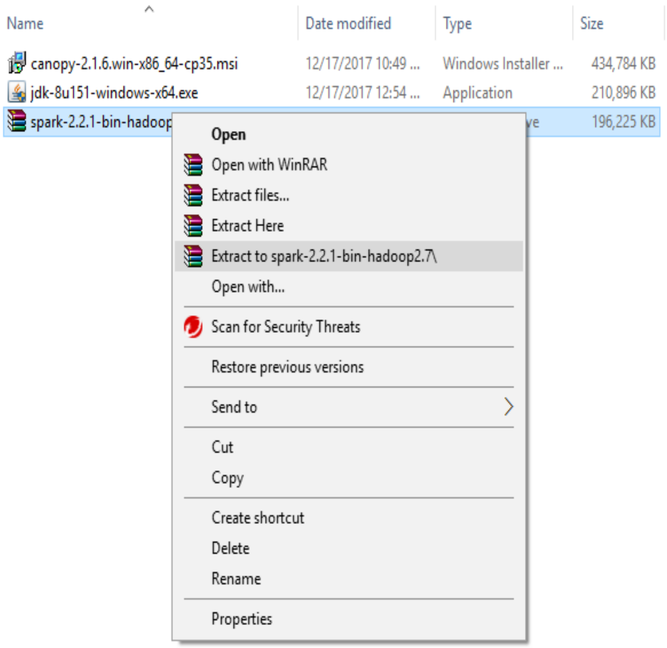


WinRAR and RAR archiver downloads

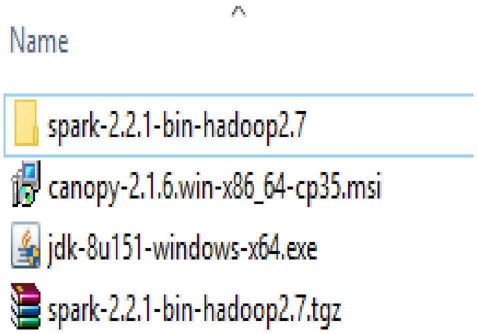
English WinRAR and RAR release

Software name
WinRAR x86 (32 bit) 5.50
WinRAR x64 (64 bit) 5.50
RAR for Android on Google Play
RAR for Android 5.50 build 45 local copy
RAR 5.50 for Linux
RAR 5.50 for Linux x64
RAR 5.50 for FreeBSD
RAR 5.50 for Mac OS X
WinRAR interface themes

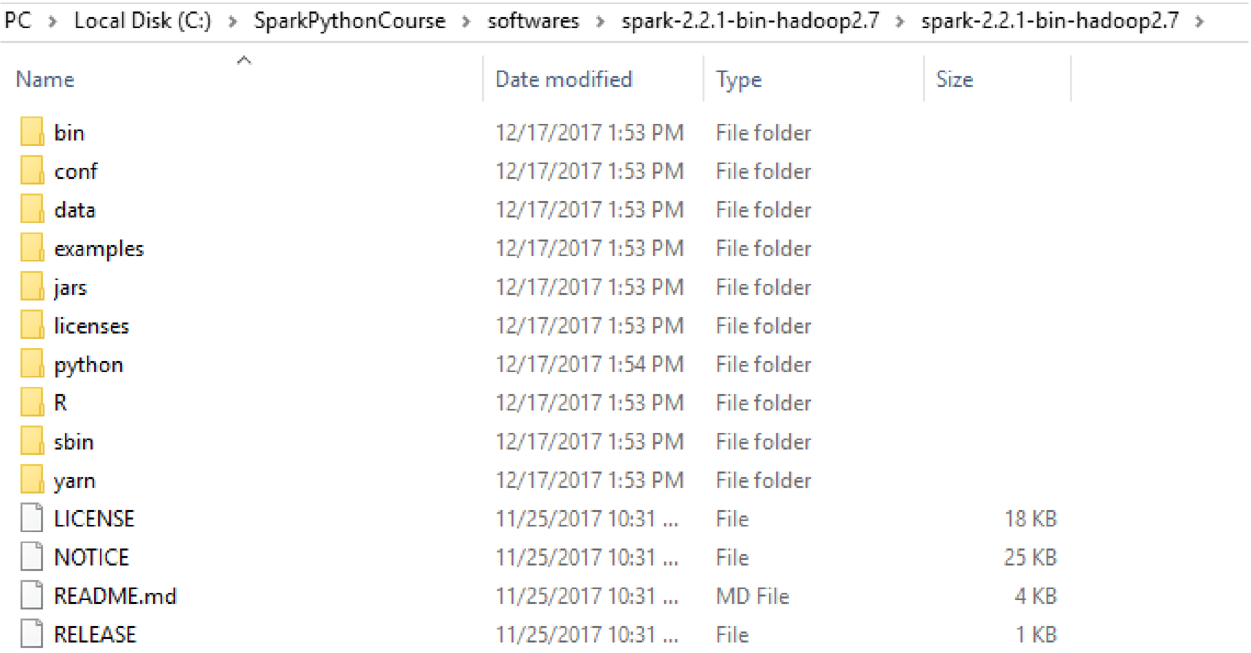
Go to the folder where the Spark tgz file is downloaded. Right click on the file and extract to spark-2.2.1-bin-hadoop2.7, this will create a folder and you will see many folders and files inside (as in the picture below).



Name	Date modified	Type	Size
canopy-2.1.6.win-x86_64-cp35.msi	12/17/2017 10:49 ...	Windows Installer ...	434,784 KB
jdk-8u151-windows-x64.exe	12/17/2017 12:54 ...	Application	210,896 KB
spark-2.2.1-bin-hadoop		ve	196,225 KB



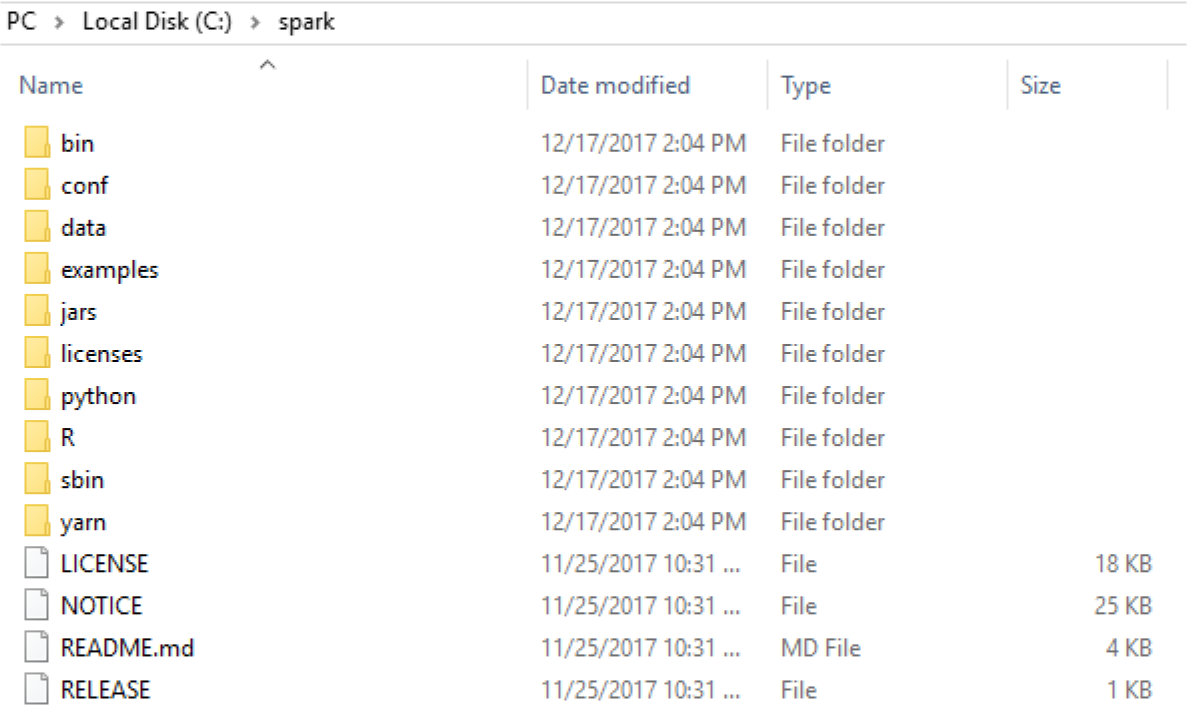
Name
spark-2.2.1-bin-hadoop2.7
canopy-2.1.6.win-x86_64-cp35.msi
jdk-8u151-windows-x64.exe
spark-2.2.1-bin-hadoop2.7.tgz



PC › Local Disk (C:) › SparkPythonCourse › softwares › spark-2.2.1-bin-hadoop2.7 › spark-2.2.1-bin-hadoop2.7 ›

Name	Date modified	Type	Size
bin	12/17/2017 1:53 PM	File folder	
conf	12/17/2017 1:53 PM	File folder	
data	12/17/2017 1:53 PM	File folder	
examples	12/17/2017 1:53 PM	File folder	
jars	12/17/2017 1:53 PM	File folder	
licenses	12/17/2017 1:53 PM	File folder	
python	12/17/2017 1:54 PM	File folder	
R	12/17/2017 1:53 PM	File folder	
sbin	12/17/2017 1:53 PM	File folder	
yarn	12/17/2017 1:53 PM	File folder	
LICENSE	11/25/2017 10:31 ...	File	18 KB
NOTICE	11/25/2017 10:31 ...	File	25 KB
README.md	11/25/2017 10:31 ...	MD File	4 KB
RELEASE	11/25/2017 10:31 ...	File	1 KB

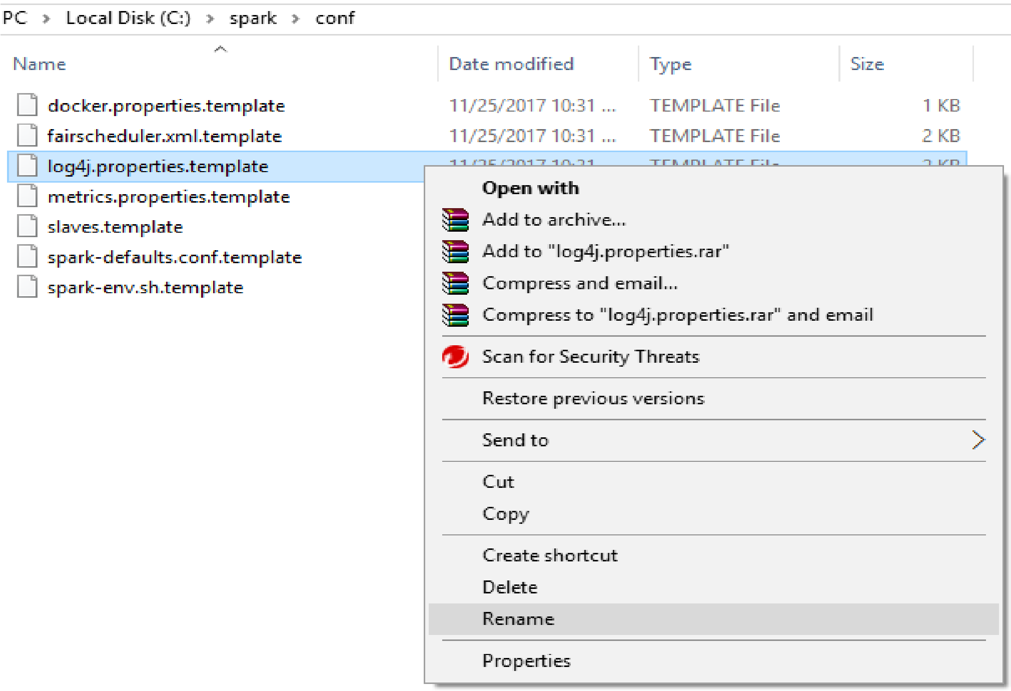
We are now ready to install Apache Spark. Create a new folder in C drive names spark and copy whole content (shown in the above picture) into that folder (as in picture below).



PC › Local Disk (C:) › spark

Name	Date modified	Type	Size
bin	12/17/2017 2:04 PM	File folder	
conf	12/17/2017 2:04 PM	File folder	
data	12/17/2017 2:04 PM	File folder	
examples	12/17/2017 2:04 PM	File folder	
jars	12/17/2017 2:04 PM	File folder	
licenses	12/17/2017 2:04 PM	File folder	
python	12/17/2017 2:04 PM	File folder	
R	12/17/2017 2:04 PM	File folder	
sbin	12/17/2017 2:04 PM	File folder	
yarn	12/17/2017 2:04 PM	File folder	
LICENSE	11/25/2017 10:31 ...	File	18 KB
NOTICE	11/25/2017 10:31 ...	File	25 KB
README.md	11/25/2017 10:31 ...	MD File	4 KB
RELEASE	11/25/2017 10:31 ...	File	1 KB

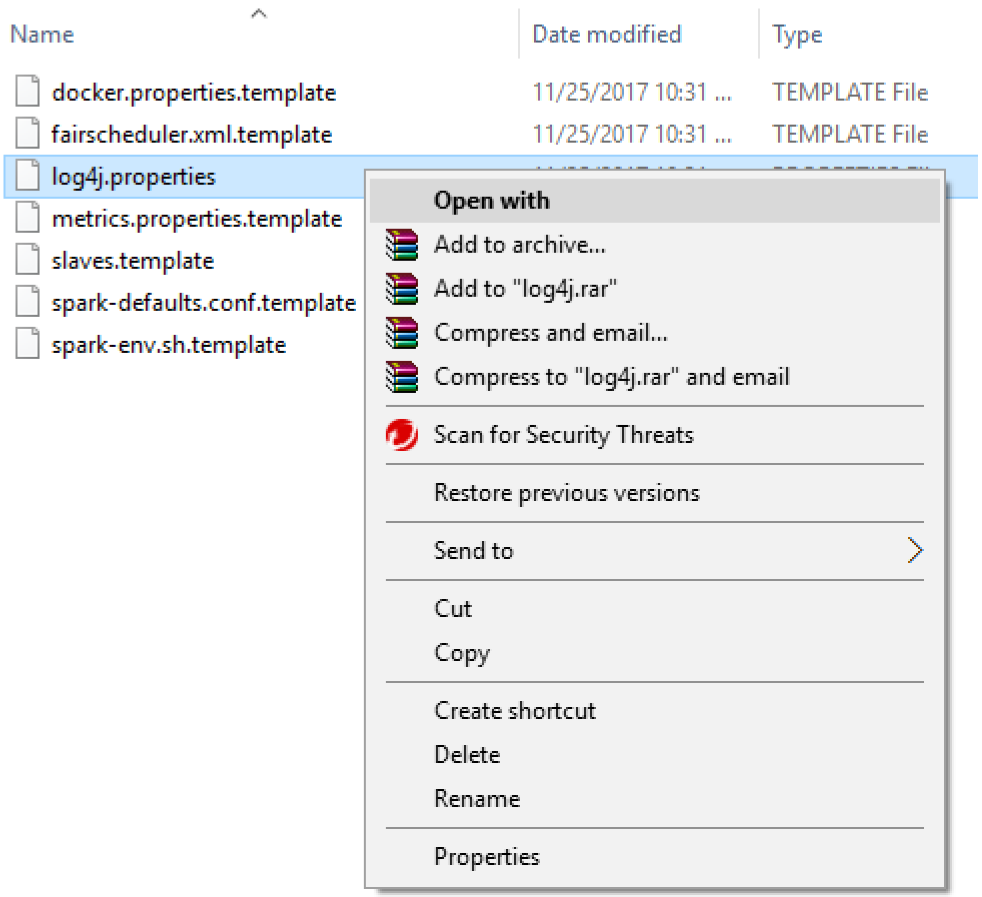
Now go to the conf folder and rename the file log4j.properties.template to log4j.properties by excluding the last part of the name.



PC › Local Disk (C:) › spark › conf

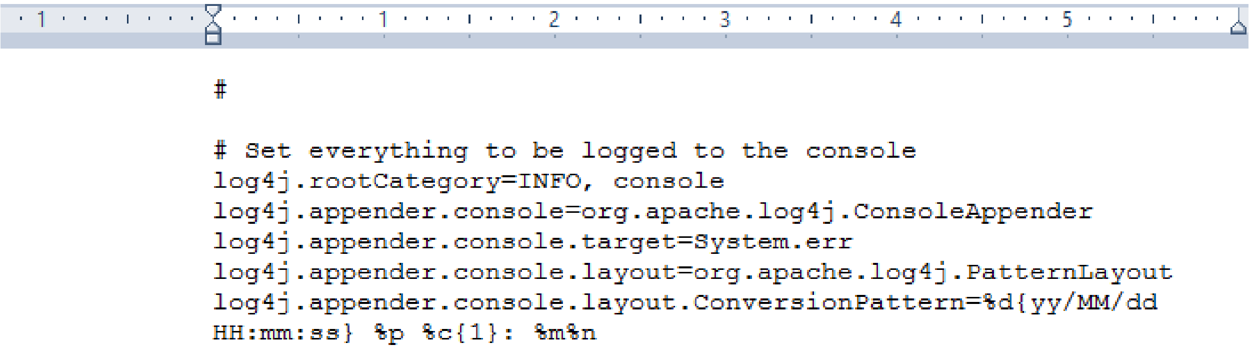
Name	Date modified	Type	Size
docker.properties.template	11/25/2017 10:31 ...	TEMPLATE File	1 KB
fairscheduler.xml.template	11/25/2017 10:31 ...	TEMPLATE File	2 KB
log4j.properties.template	11/25/2017 10:31	TEMPLATE File	2 KB
metrics.properties.template			
slaves.template			
spark-defaults.conf.template			
spark-env.sh.template			

Open with
Add to archive...
Add to "log4j.properties.rar"
Compress and email...
Compress to "log4j.properties.rar" and email
Scan for Security Threats
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties



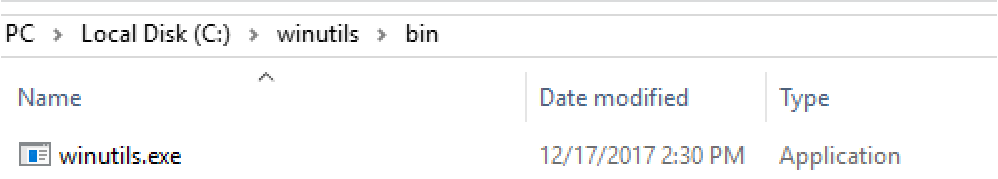
Name	Date modified	Type
docker.properties.template	11/25/2017 10:31 ...	TEMPLATE File
fairscheduler.xml.template	11/25/2017 10:31 ...	TEMPLATE File
log4j.properties		
metrics.properties.template		
slaves.template		
spark-defaults.conf.template		
spark-env.sh.template		

Next, open the renamed file by using *Open With* and open the file with WordPad. Here replace INFO by ERROR in the line log4j.rootCategory=INFO, console (shown below) then save and close the file in usual way.



```
#

# Set everything to be logged to the console
log4j.rootCategory=INFO, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd
HH:mm:ss} %p %c{1}: %m%n
```

Then you have to download winutils.exe from the website (skip this if you are not a windows user) http://sundog-spark.s3.amazonaws.com/winutils.exe. Create a new folder in C drive named winutils, inside the winutils folder create another folder named bin and copy this winutils.exe file to that bin folder (shown below).

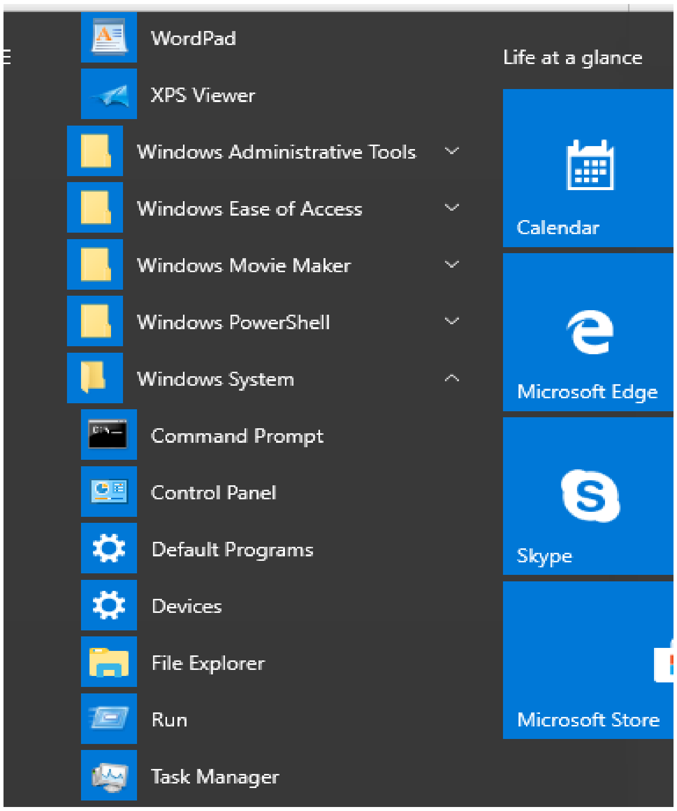


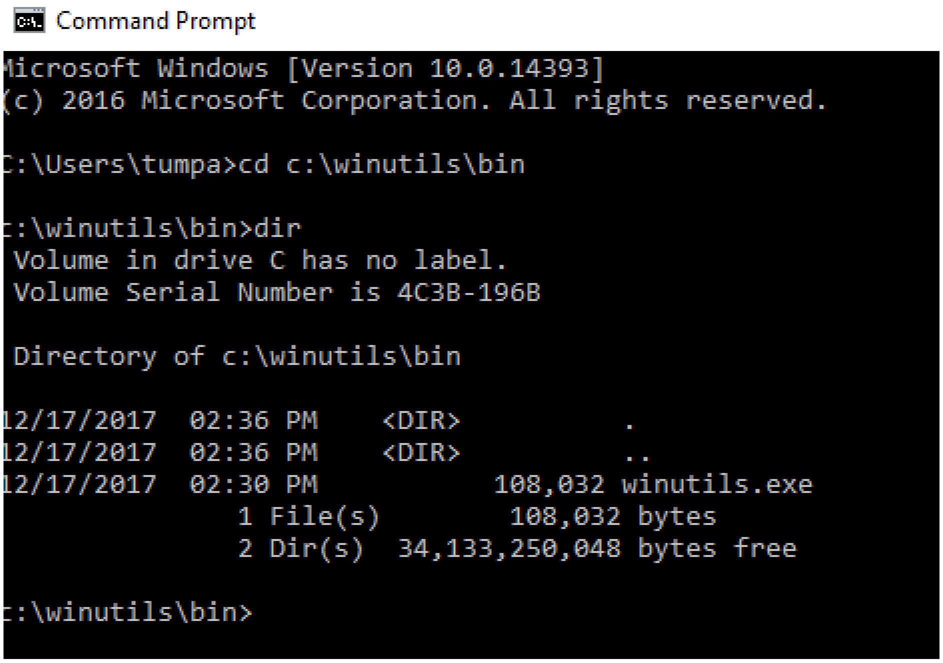
PC › Local Disk (C:) › winutils › bin

Name	Date modified	Type
winutils.exe	12/17/2017 2:30 PM	Application

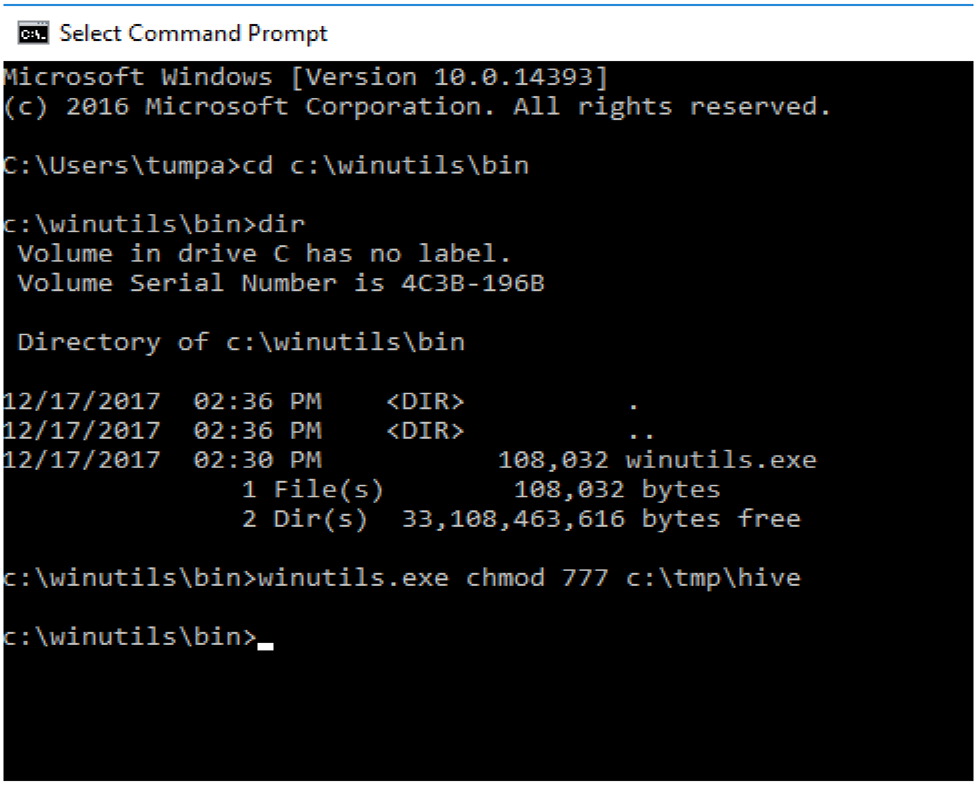
In the next step we have to integrate everything to make the system workable. Let us create a folder named tmp in the C drive and within that create another folder named hive, that is, c:\tmp\hive would be the directory for that folder.

We are nearly done. Go to the *Command Prompt* window under Windows System in Windows 10 (other versions may have different view)





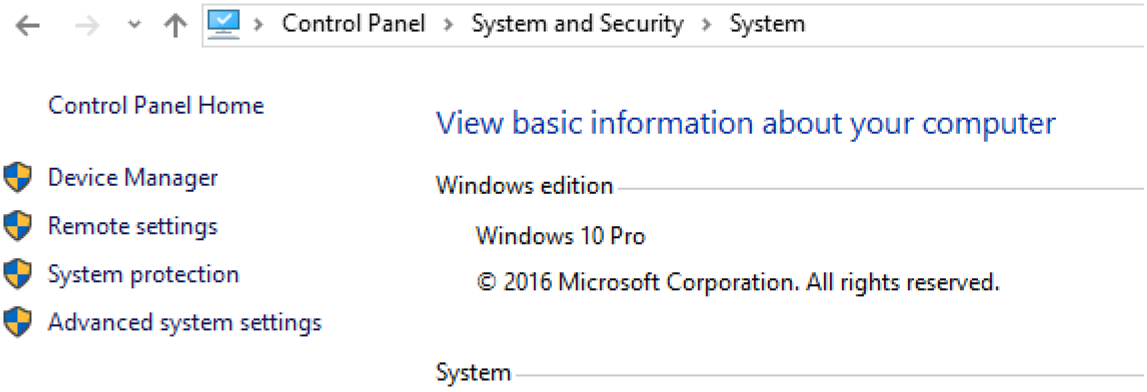
```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\tumpa>cd c:\winutils\bin

c:\winutils\bin>dir
 Volume in drive C has no label.
 Volume Serial Number is 4C3B-196B

 Directory of c:\winutils\bin

12/17/2017  02:36 PM    <DIR>          .
12/17/2017  02:36 PM    <DIR>          ..
12/17/2017  02:30 PM           108,032 winutils.exe
               1 File(s)        108,032 bytes
               2 Dir(s)  34,133,250,048 bytes free

c:\winutils\bin>
```



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\tumpa>cd c:\winutils\bin

c:\winutils\bin>dir
 Volume in drive C has no label.
 Volume Serial Number is 4C3B-196B

 Directory of c:\winutils\bin

12/17/2017  02:36 PM    <DIR>          .
12/17/2017  02:36 PM    <DIR>          ..
12/17/2017  02:30 PM           108,032 winutils.exe
               1 File(s)        108,032 bytes
               2 Dir(s)  33,108,463,616 bytes free

c:\winutils\bin>winutils.exe chmod 777 c:\tmp\hive

c:\winutils\bin>_
```

Okay, it is passed outright! We are done! This completes all file formation to run Spark successfully in the system. (Note: If not, you may see some error message related to access issues or path recognition. In such situations, you may need to provide access permission to the folder c:\tmp\hive, hope this will solve the problem.)

Environmental Variable: Enter from *Control Panel* to *System and Security* and then to *System* (shown below).



View basic information about your computer

Windows edition

Windows 10 Pro

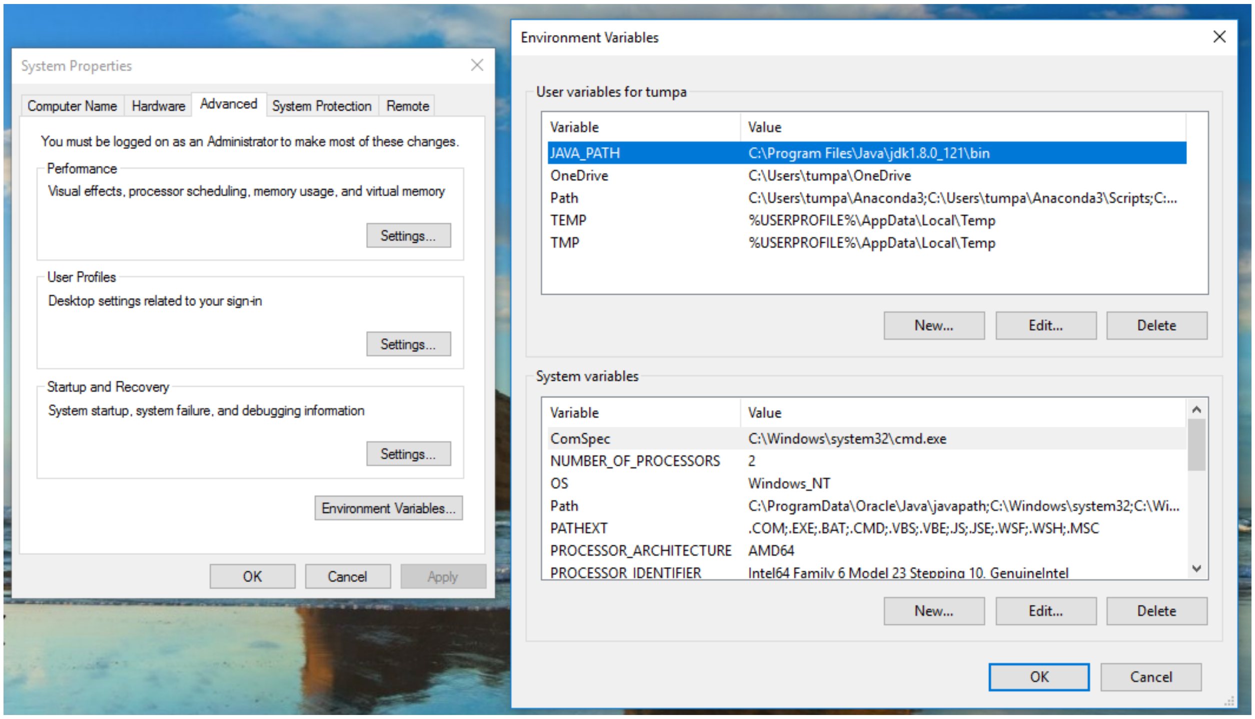
© 2016 Microsoft Corporation. All rights reserved.

System

Click on *Advanced system settings* and then *Environment Variables* (shown below, left window). This will open the right window in the picture below. Then we select the new *User variable for user (user name)* by clicking on *New,* write new user variable name

Variable name: SPARK_HOME

Variable value: C:\spark



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

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

Desktop settings related to your sign-in

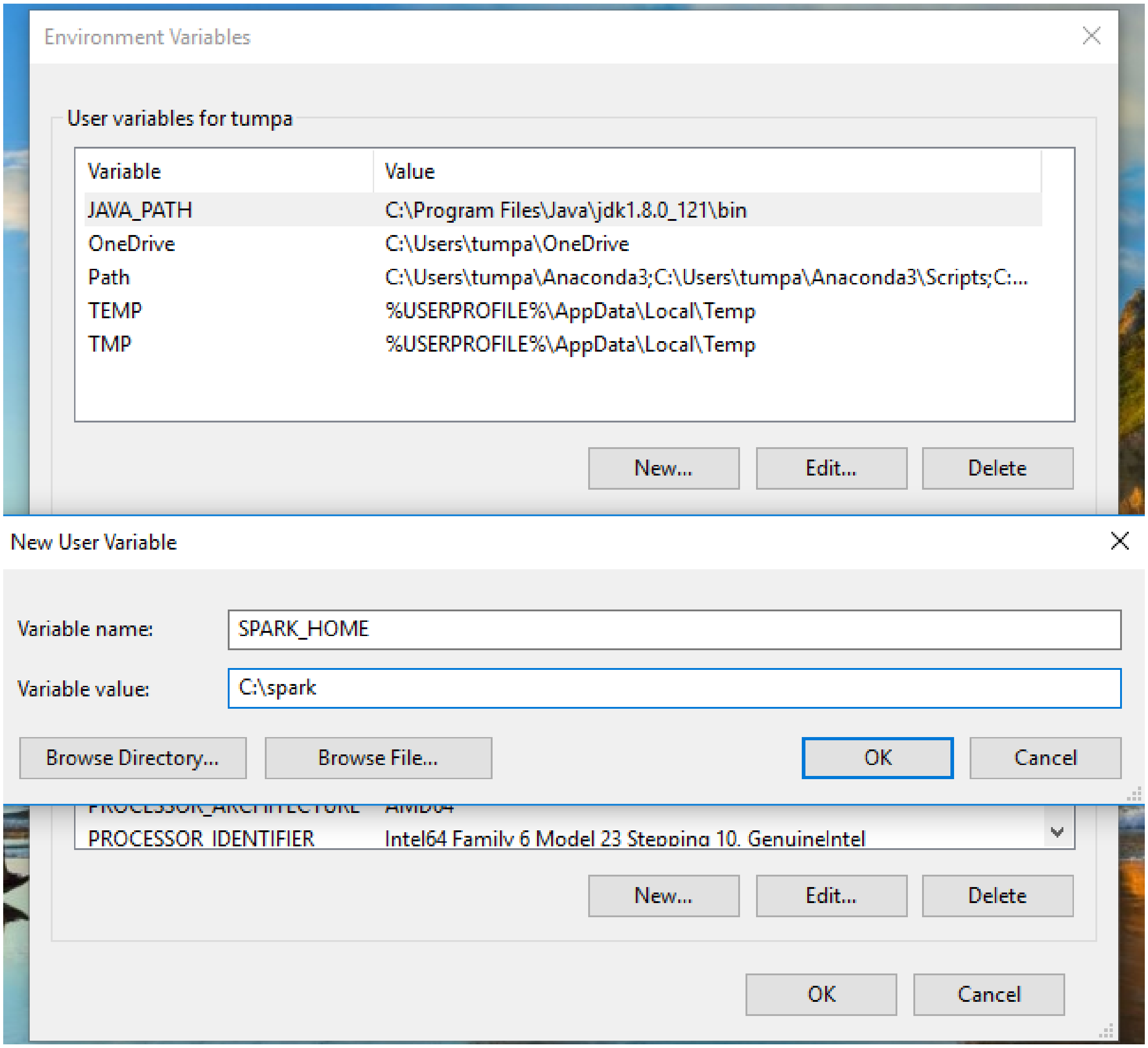
System startup, system failure, and debugging information

User variables for tumpa

Variable	Value
JAVA_PATH	C:\Program Files\Java\jdk1.8.0_121\bin
OneDrive	C:\Users\tumpa\OneDrive
Path	C:\Users\tumpa\Anaconda3;C:\Users\tumpa\Anaconda3\Scripts;C:...
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
NUMBER_OF_PROCESSORS	2
OS	Windows_NT
Path	C:\ProgramData\Oracle\Java\javapath;C:\Windows\system32;C:\Wi...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 23 Stepping 10, GenuineIntel



Variable	Value
JAVA_PATH	C:\Program Files\Java\jdk1.8.0_121\bin
OneDrive	C:\Users\tumpa\OneDrive
Path	C:\Users\tumpa\Anaconda3;C:\Users\tumpa\Anaconda3\Scripts;C:...
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

Variable name: SPARK_HOME

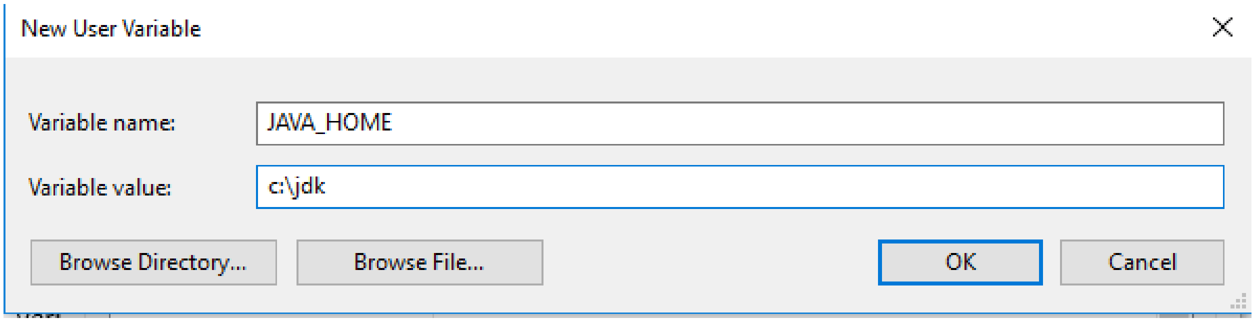
Variable value: C:\spark

PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 23 Stepping 10, GenuineIntel

Similarly, enter another user variable name for Java with value.

Variable name: JAVA_HOME

Variable value: c:\jdk



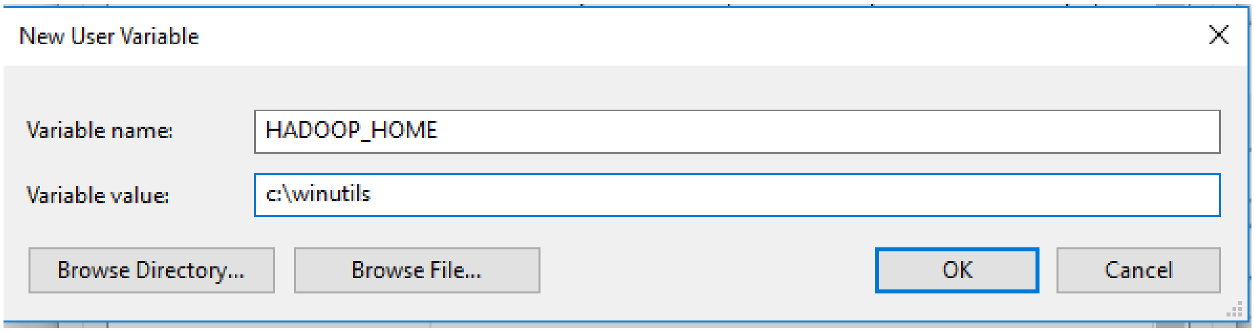
Variable name: JAVA_HOME

Variable value: c:\jdk

Finally, we have to assign the path for Hadoop home by creating a new variable in the above mentioned process. This will complete the environmental variable set up.

Variable name: HADOOP_HOME

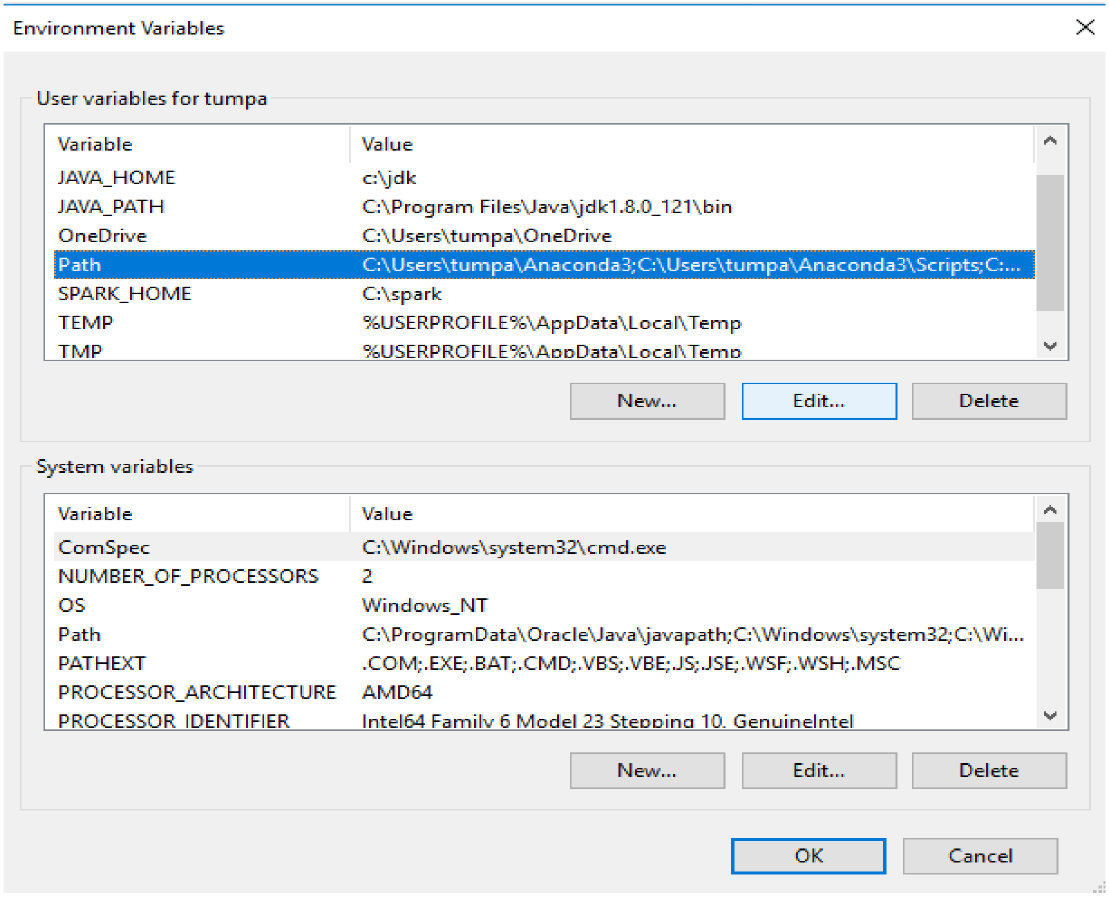
Variable value: c:\winutils



Variable name: HADOOP_HOME

Variable value: c:\winutils

Further, we have to edit the path for these variables. Click on *Path* (as highlighted in the screen below) and then pressing *Edit* will provide another window given below.

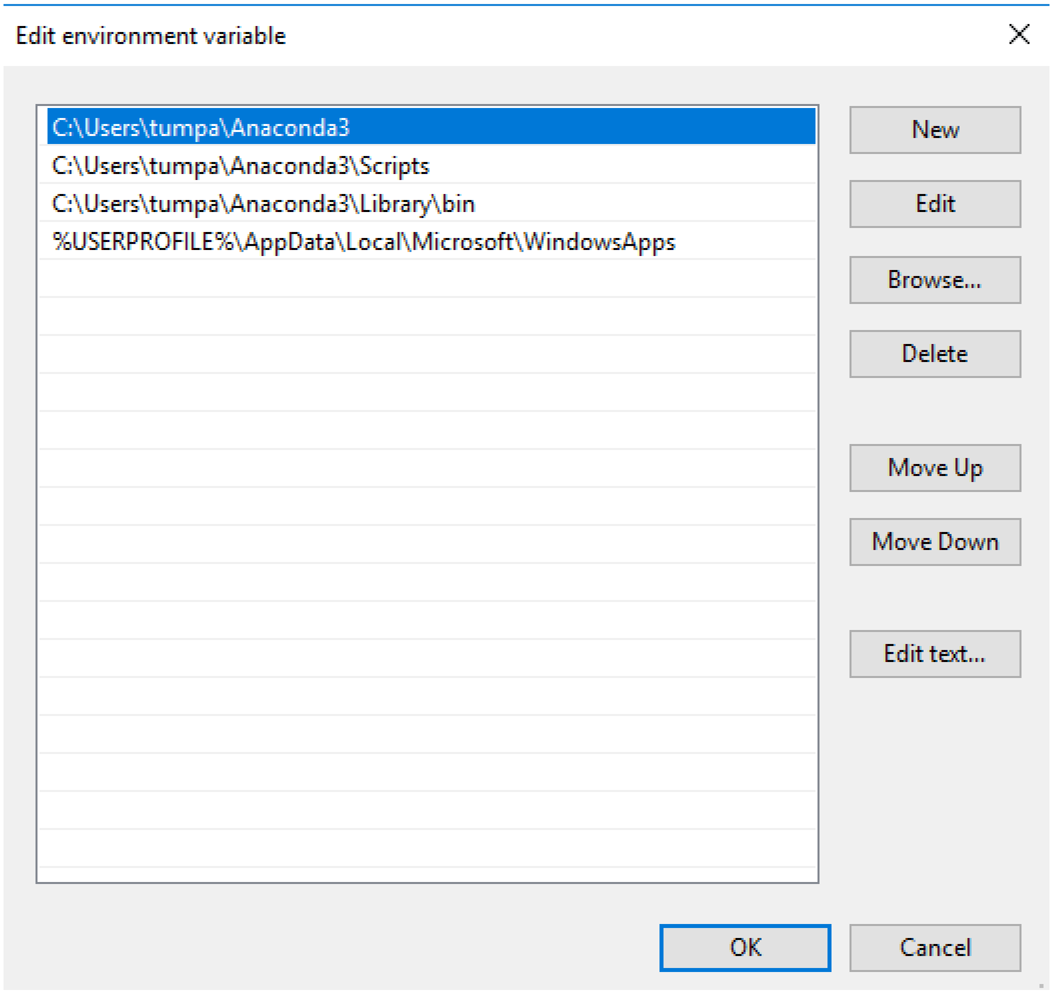


User variables for tumpa

Variable	Value
JAVA_HOME	c:\jdk
JAVA_PATH	C:\Program Files\Java\jdk1.8.0_121\bin
OneDrive	C:\Users\tumpa\OneDrive
Path	C:\Users\tumpa\Anaconda3;C:\Users\tumpa\Anaconda3\Scripts;C:...
SPARK_HOME	C:\spark
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

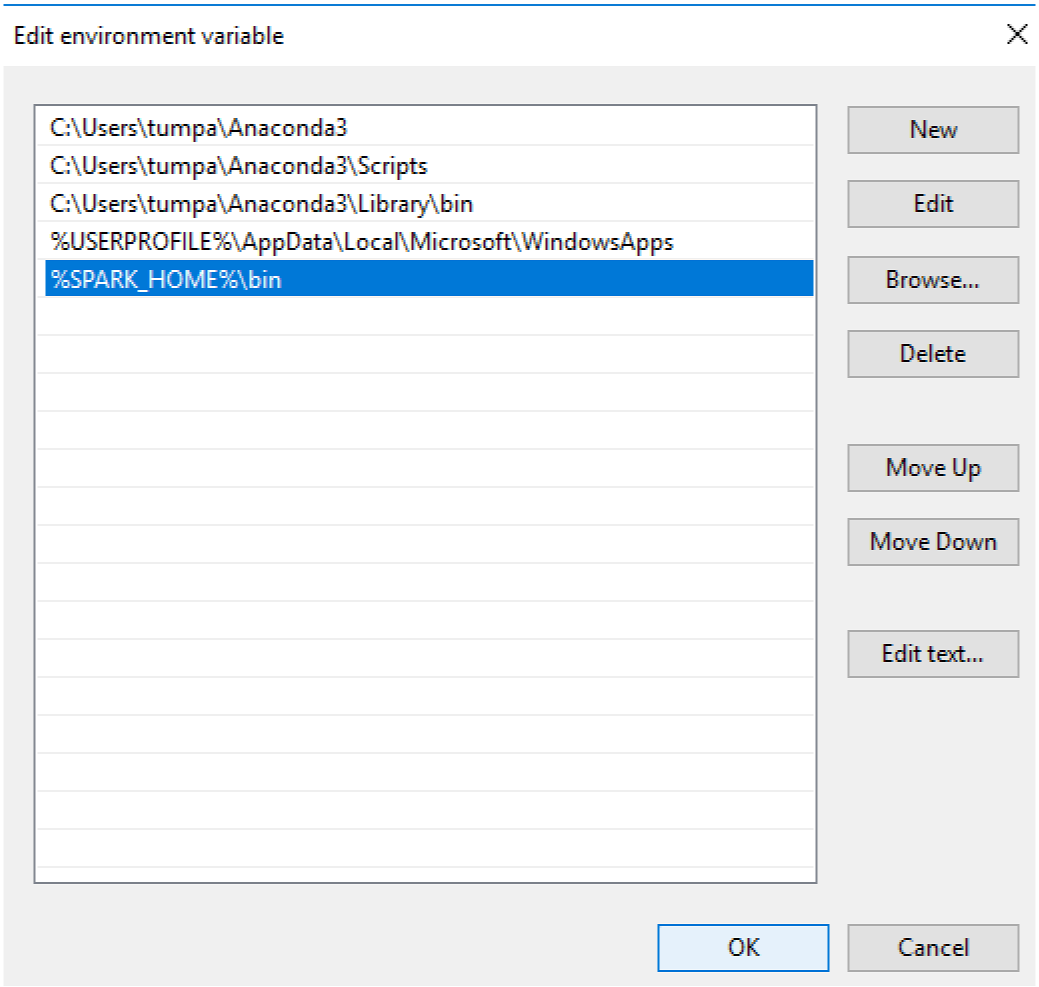
Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
NUMBER_OF_PROCESSORS	2
OS	Windows_NT
Path	C:\ProgramData\Oracle\Java\javapath;C:\Windows\system32;C:\Wi...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	Intel64 Family 6 Model 23 Stepping 10, GenuineIntel



Edit environment variable

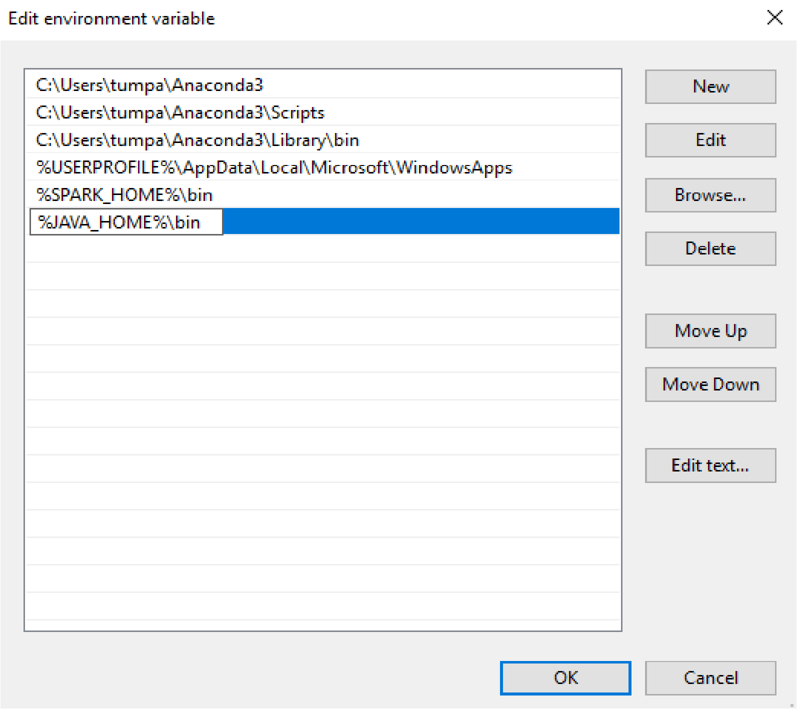
C:\Users\tumpa\Anaconda3
C:\Users\tumpa\Anaconda3\Scripts
C:\Users\tumpa\Anaconda3\Library\bin
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps

In the above window, click on *New* and enter %SPARK_HOME%\bin



C:\Users\tumpa\Anaconda3
C:\Users\tumpa\Anaconda3\Scripts
C:\Users\tumpa\Anaconda3\Library\bin
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
%SPARK_HOME%\bin

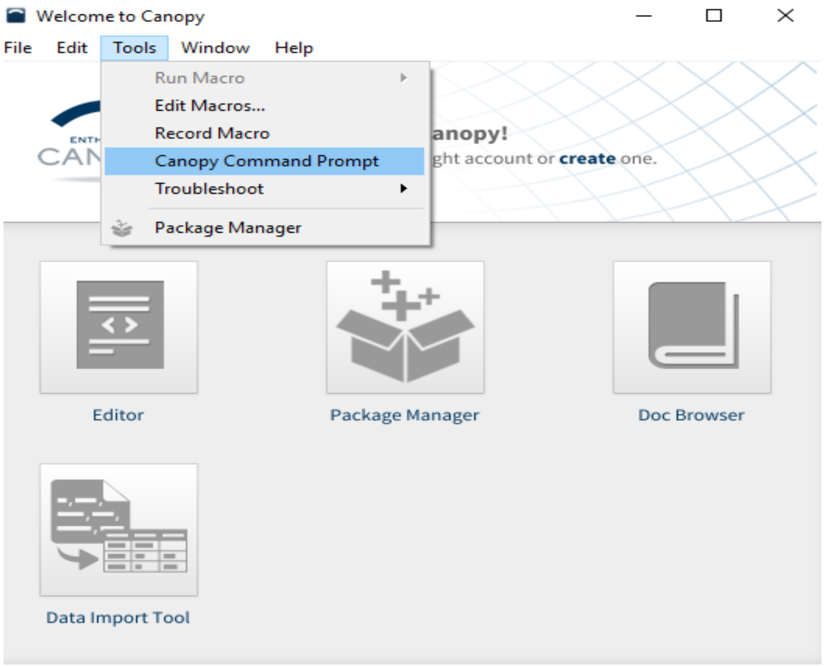
Add another one, %JAVA_HOME%\bin.

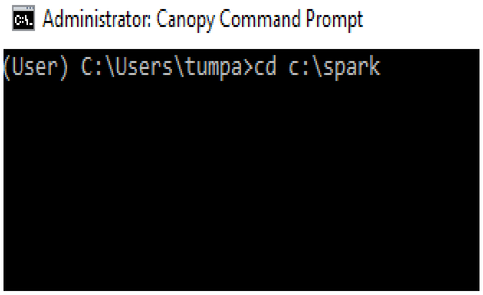

Edit environment variable

C:\Users\tumpa\Anaconda3
C:\Users\tumpa\Anaconda3\Scripts
C:\Users\tumpa\Anaconda3\Library\bin
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
%SPARK_HOME%\bin
%JAVA_HOME%\bin

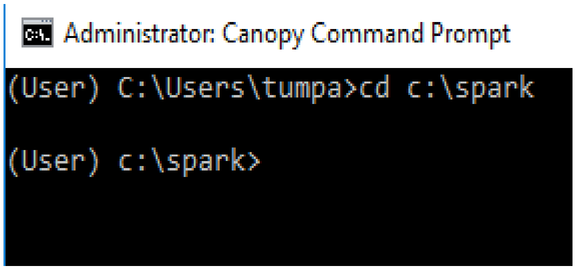
All done now. Keep clicking on *OK* until all is OK for this set up and close the window.

It is now time to start checking whether the whole lot of settings work or not. Let us start by checking the Canopy. If you double click on the icon to start, you will see a screen like shown below. Now, click on *Tools* menu and start with *Canopy Command Prompt.*





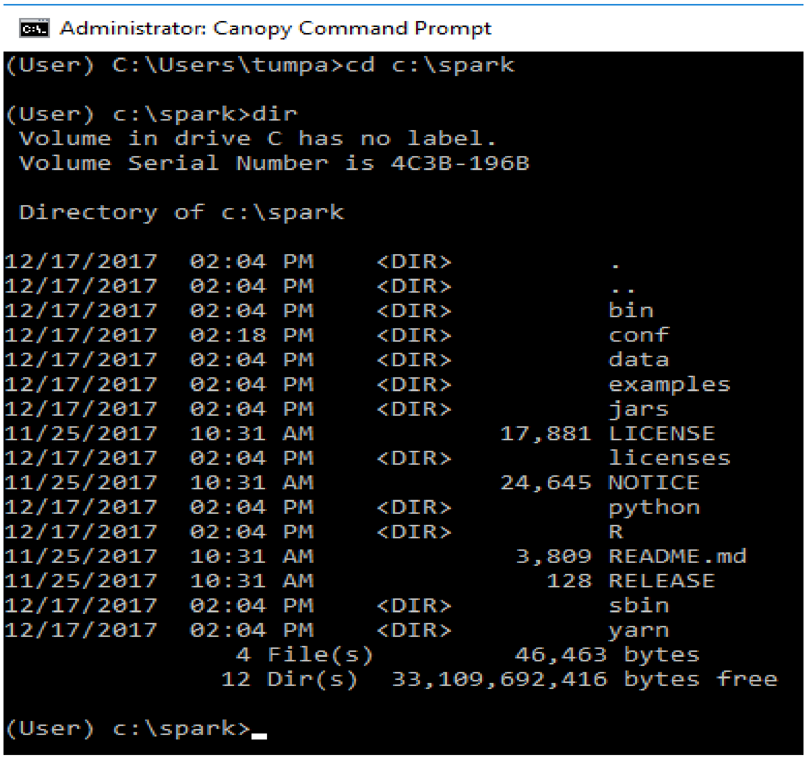
```
(User) C:\Users\tumpa>cd c:\spark
```



```
(User) C:\Users\tumpa>cd c:\spark

(User) c:\spark>
```

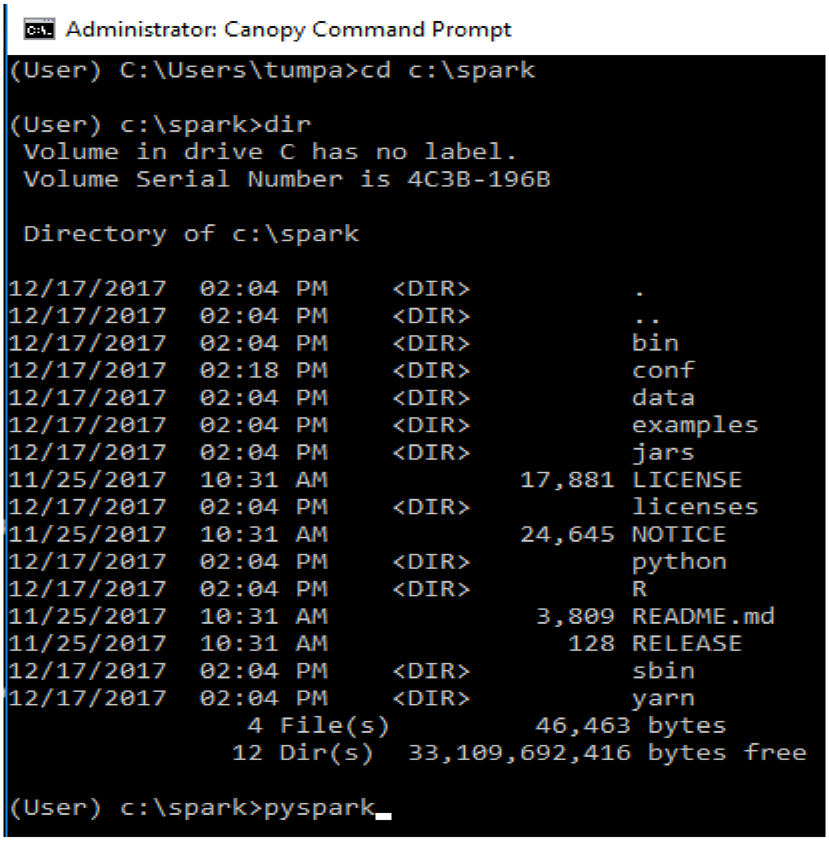
The above windows confirm the installed Spark. You can see the directory by using the command

C:\spark>dir



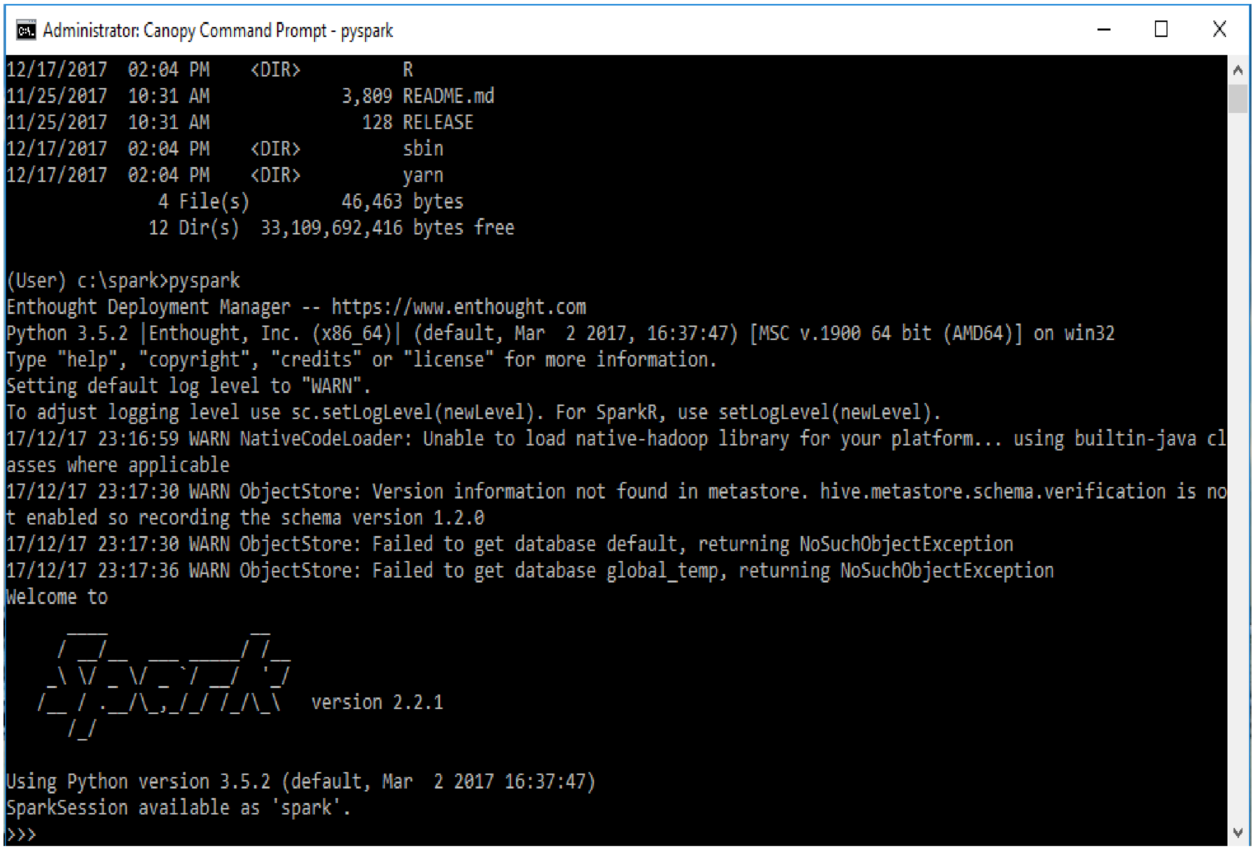
```
(User) C:\Users\tumpa>cd c:\spark

(User) c:\spark>dir
 Volume in drive C has no label.
 Volume Serial Number is 4C3B-196B

 Directory of c:\spark

12/17/2017  02:04 PM    <DIR>          .
12/17/2017  02:04 PM    <DIR>          ..
12/17/2017  02:04 PM    <DIR>          bin
12/17/2017  02:18 PM    <DIR>          conf
12/17/2017  02:04 PM    <DIR>          data
12/17/2017  02:04 PM    <DIR>          examples
12/17/2017  02:04 PM    <DIR>          jars
11/25/2017  10:31 AM            17,881 LICENSE
12/17/2017  02:04 PM    <DIR>          licenses
11/25/2017  10:31 AM            24,645 NOTICE
12/17/2017  02:04 PM    <DIR>          python
12/17/2017  02:04 PM    <DIR>          R
11/25/2017  10:31 AM             3,809 README.md
11/25/2017  10:31 AM               128 RELEASE
12/17/2017  02:04 PM    <DIR>          sbin
12/17/2017  02:04 PM    <DIR>          yarn
               4 File(s)         46,463 bytes
              12 Dir(s)  33,109,692,416 bytes free

(User) c:\spark>
```

Finally, start pyspark from the command line as in the screen below. If it is successful, you will see a welcome screen with Spark as shown in picture



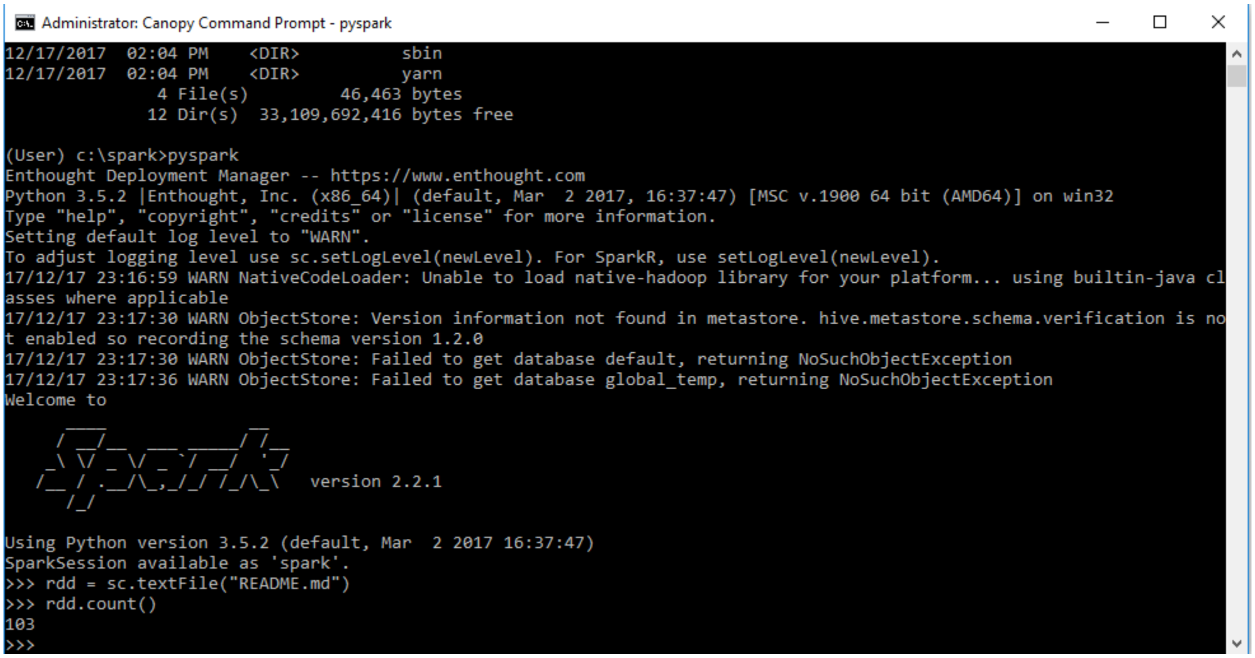
Administrator: Canopy Command Prompt

```
(User) C:\Users\tumpa>cd c:\spark

(User) c:\spark>dir
 Volume in drive C has no label.
 Volume Serial Number is 4C3B-196B

 Directory of c:\spark

12/17/2017  02:04 PM    <DIR>          .
12/17/2017  02:04 PM    <DIR>          ..
12/17/2017  02:04 PM    <DIR>          bin
12/17/2017  02:18 PM    <DIR>          conf
12/17/2017  02:04 PM    <DIR>          data
12/17/2017  02:04 PM    <DIR>          examples
12/17/2017  02:04 PM    <DIR>          jars
11/25/2017  10:31 AM            17,881 LICENSE
12/17/2017  02:04 PM    <DIR>          licenses
11/25/2017  10:31 AM            24,645 NOTICE
12/17/2017  02:04 PM    <DIR>          python
12/17/2017  02:04 PM    <DIR>          R
11/25/2017  10:31 AM             3,809 README.md
11/25/2017  10:31 AM               128 RELEASE
12/17/2017  02:04 PM    <DIR>          sbin
12/17/2017  02:04 PM    <DIR>          yarn
               4 File(s)         46,463 bytes
              12 Dir(s)  33,109,692,416 bytes free

(User) c:\spark>pyspark
```

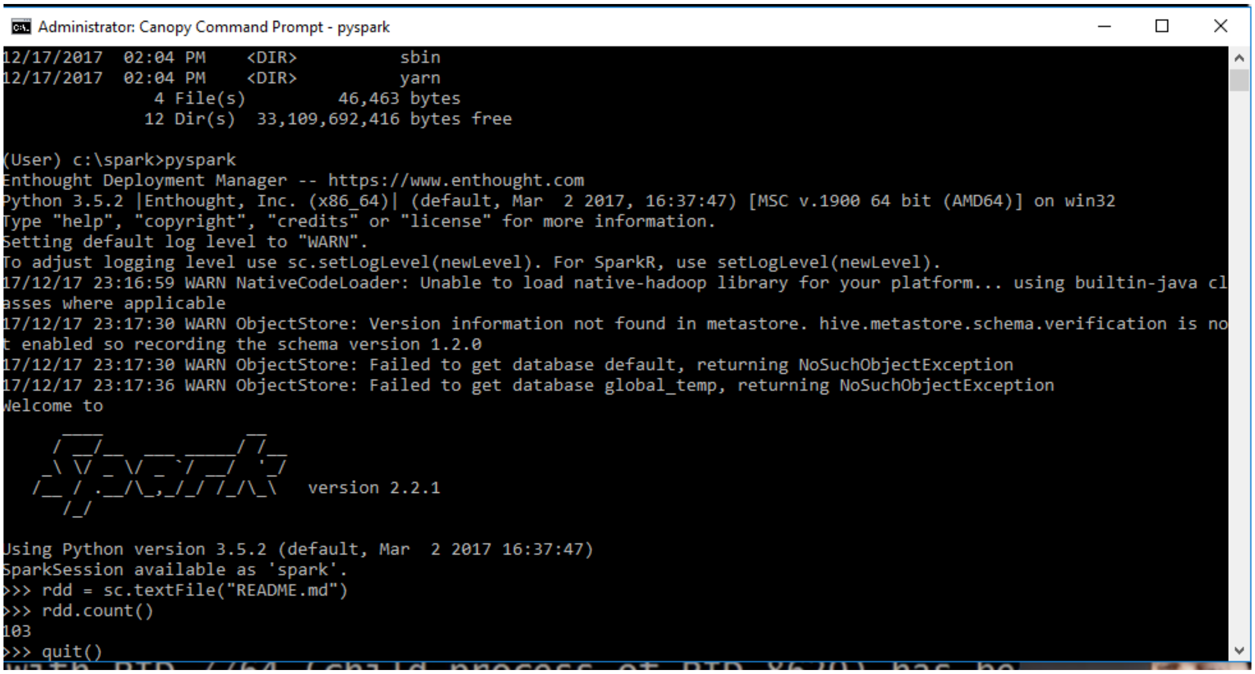


```
12/17/2017  02:04 PM    <DIR>          R
11/25/2017  10:31 AM             3,809 README.md
11/25/2017  10:31 AM               128 RELEASE
12/17/2017  02:04 PM    <DIR>          sbin
12/17/2017  02:04 PM    <DIR>          yarn
               4 File(s)         46,463 bytes
              12 Dir(s)  33,109,692,416 bytes free

(User) c:\spark>pyspark
Enthought Deployment Manager -- https://www.enthought.com
Python 3.5.2 |Enthought, Inc. (x86_64)| (default, Mar  2 2017, 16:37:47) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/12/17 23:16:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/12/17 23:17:30 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is no
t enabled so recording the schema version 1.2.0
17/12/17 23:17:30 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/12/17 23:17:36 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Python version 3.5.2 (default, Mar  2 2017 16:37:47)
SparkSession available as 'spark'.
>>>
```

We saw that in the directory there is a README.md file (in some other version this can be a text file with extension txt). We read the file and count the number of line. Cool, there are 103 lines: so Spark is working with Python.

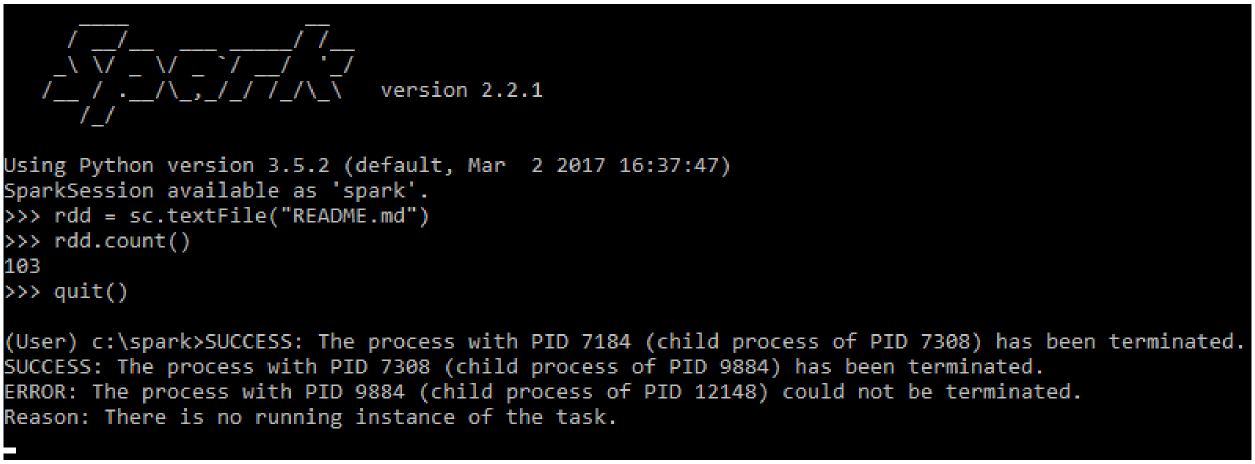


```
12/17/2017  02:04 PM    <DIR>          sbin
12/17/2017  02:04 PM    <DIR>          yarn
               4 File(s)         46,463 bytes
              12 Dir(s)  33,109,692,416 bytes free

(User) c:\spark>pyspark
Enthought Deployment Manager -- https://www.enthought.com
Python 3.5.2 |Enthought, Inc. (x86_64)| (default, Mar  2 2017, 16:37:47) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/12/17 23:16:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/12/17 23:17:30 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is no
t enabled so recording the schema version 1.2.0
17/12/17 23:17:30 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/12/17 23:17:36 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Python version 3.5.2 (default, Mar  2 2017 16:37:47)
SparkSession available as 'spark'.
>>> rdd = sc.textFile("README.md")
>>> rdd.count()
103
>>>
```



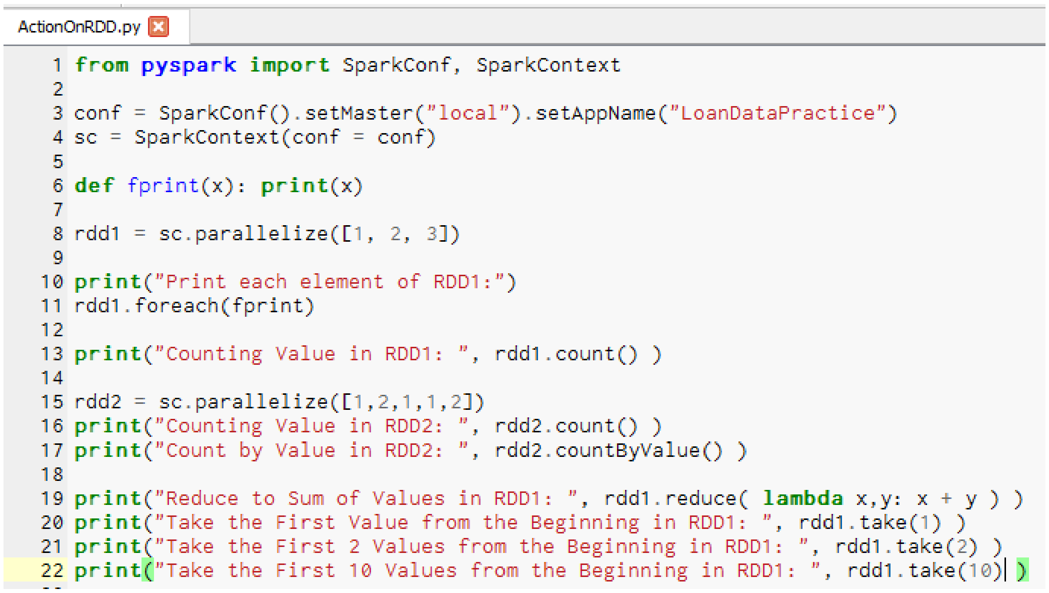
```
12/17/2017  02:04 PM    <DIR>          sbin
12/17/2017  02:04 PM    <DIR>          yarn
               4 File(s)         46,463 bytes
              12 Dir(s)  33,109,692,416 bytes free

(User) c:\spark>pyspark
Enthought Deployment Manager -- https://www.enthought.com
Python 3.5.2 |Enthought, Inc. (x86_64)| (default, Mar  2 2017, 16:37:47) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/12/17 23:16:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
17/12/17 23:17:30 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is no
t enabled so recording the schema version 1.2.0
17/12/17 23:17:30 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/12/17 23:17:36 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Python version 3.5.2 (default, Mar  2 2017 16:37:47)
SparkSession available as 'spark'.
>>> rdd = sc.textFile("README.md")
>>> rdd.count()
103
>>> quit()
```

Type quit() if you want to get exit from the screen. Once you see the following scenario, you are ready to close the window from the top rightmost corner.

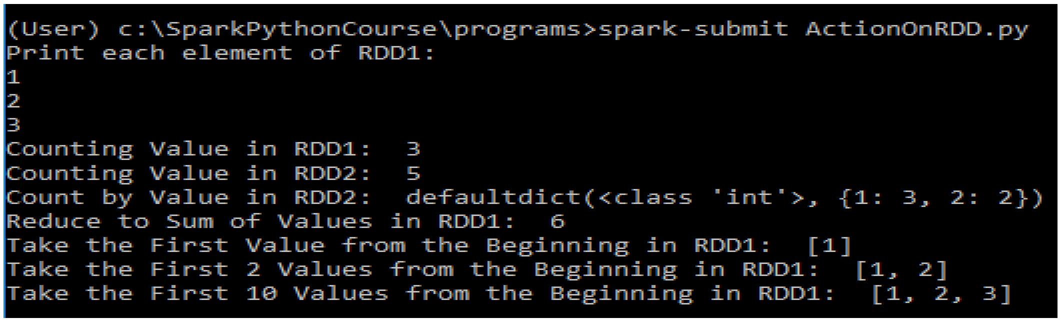


```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Python version 3.5.2 (default, Mar  2 2017 16:37:47)
SparkSession available as 'spark'.
>>> rdd = sc.textFile("README.md")
>>> rdd.count()
103
>>> quit()

(User) c:\spark>SUCCESS: The process with PID 7184 (child process of PID 7308) has been terminated.
SUCCESS: The process with PID 7308 (child process of PID 9884) has been terminated.
ERROR: The process with PID 9884 (child process of PID 12148) could not be terminated.
Reason: There is no running instance of the task.
```

Fire off PySpark: Write a python script (file name of our example script is ActionOnRDD.py) as shown in the snippet below. Then go to the directory of the written script from the command prompt and run the script by using the command spark-submit ActionOnRDD.py. If everything is alright, you will see outcomes as has been displayed in the screenshot.



```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("LoanDataPractice")
sc = SparkContext(conf = conf)

def fprint(x): print(x)

rdd1 = sc.parallelize([1, 2, 3])

print("Print each element of RDD1:")
rdd1.foreach(fprint)

print("Counting Value in RDD1: ", rdd1.count() )

rdd2 = sc.parallelize([1,2,1,1,2])
print("Counting Value in RDD2: ", rdd2.count() )
print("Count by Value in RDD2: ", rdd2.countByValue() )

print("Reduce to Sum of Values in RDD1: ", rdd1.reduce( lambda x,y: x + y ) )
print("Take the First Value from the Beginning in RDD1: ", rdd1.take(1) )
print("Take the First 2 Values from the Beginning in RDD1: ", rdd1.take(2) )
print("Take the First 10 Values from the Beginning in RDD1: ", rdd1.take(10) )
```

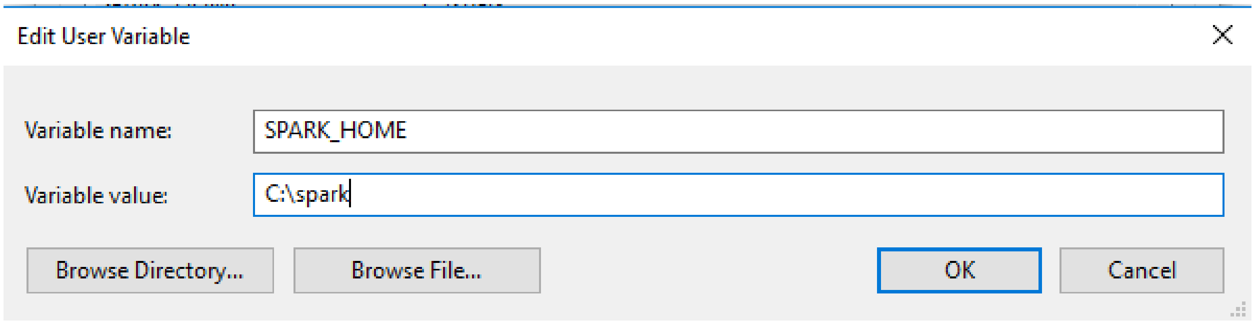


```
(User) c:\SparkPythonCourse\programs>spark-submit ActionOnRDD.py
Print each element of RDD1:
1
2
3
Counting Value in RDD1:  3
Counting Value in RDD2:  5
Count by Value in RDD2:  defaultdict(<class 'int'>, {1: 3, 2: 2})
Reduce to Sum of Values in RDD1:  6
Take the First Value from the Beginning in RDD1:  [1]
Take the First 2 Values from the Beginning in RDD1:  [1, 2]
Take the First 10 Values from the Beginning in RDD1:  [1, 2, 3]
```

1.2 Anaconda with Jupyter Notebook for Windows 10

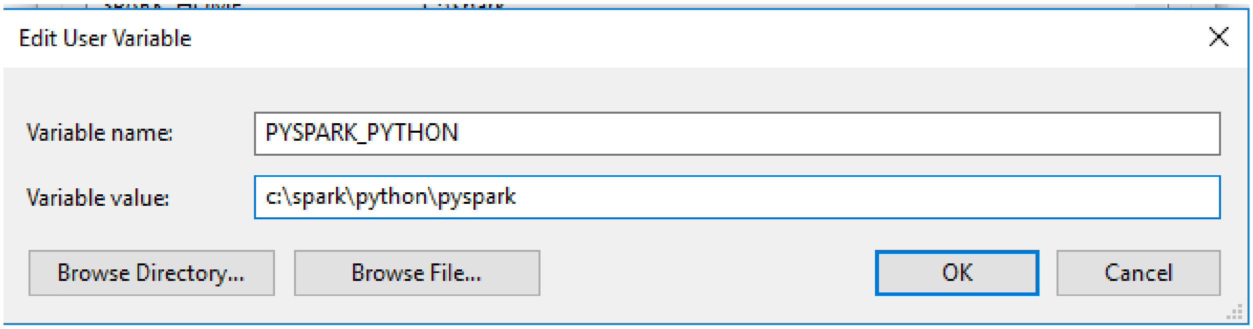
Install Python IDE: Download and install Anaconda in the C drive with the path c:\Anaconda3

Note: Install JAVA, SPARK and all relevant tools as has been done in the as has been done in the previous section 1.1. In addition to the path and environmental variable settings make sure the following settings for environmental variables and path.



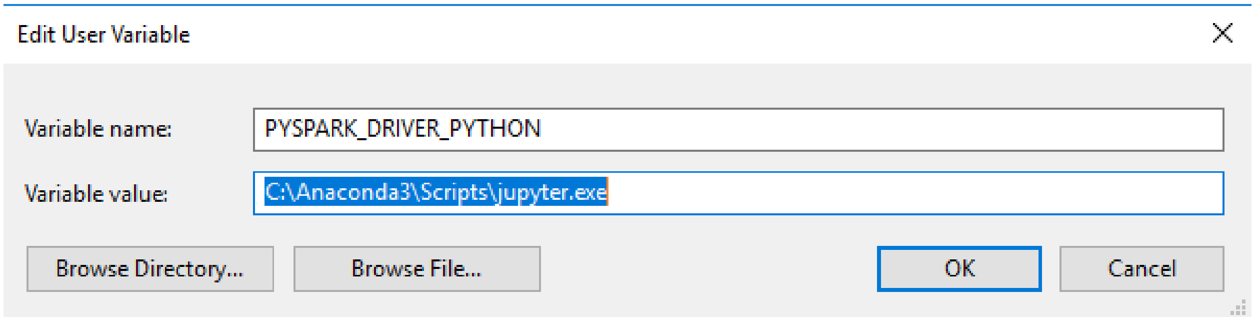
Variable name: SPARK_HOME

Variable value: C:\spark



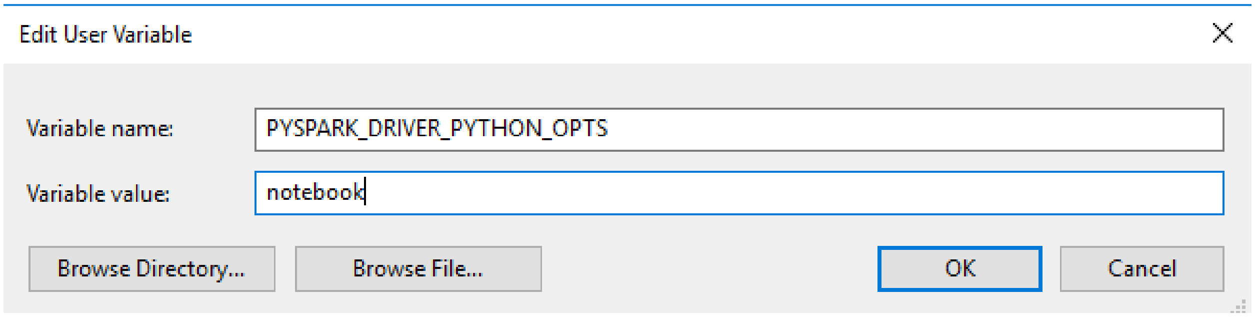
Variable name: PYSPARK_PYTHON

Variable value: c:\spark\python\pyspark



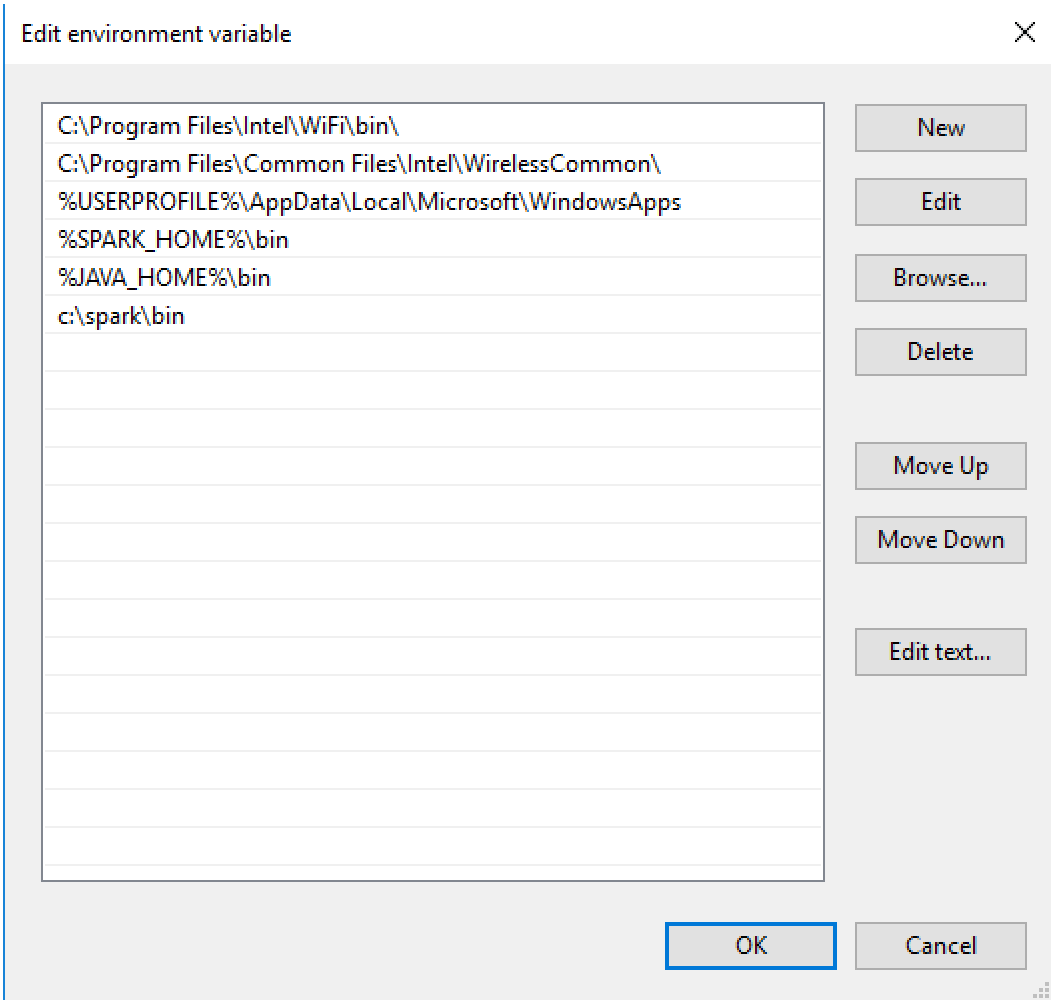
Variable name: PYSPARK_DRIVER_PYTHON

Variable value: C:\Anaconda3\Scripts\jupyter.exe



Variable name: PYSPARK_DRIVER_PYTHON_OPTS

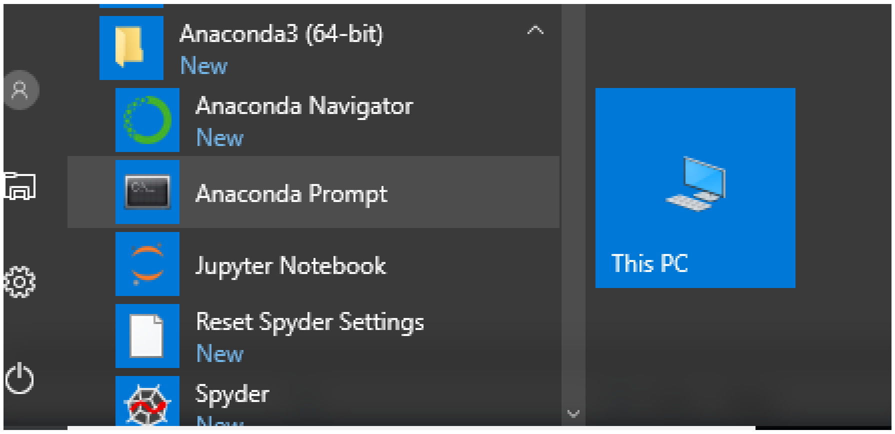
Variable value: notebook

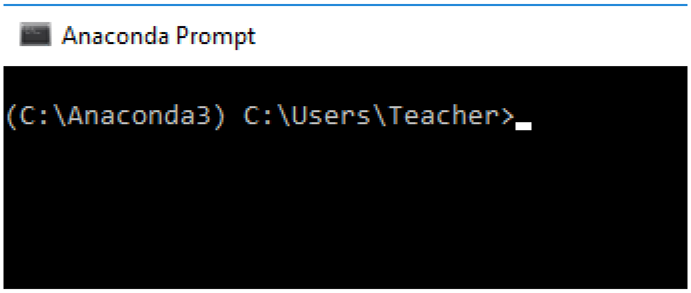


Edit environment variable

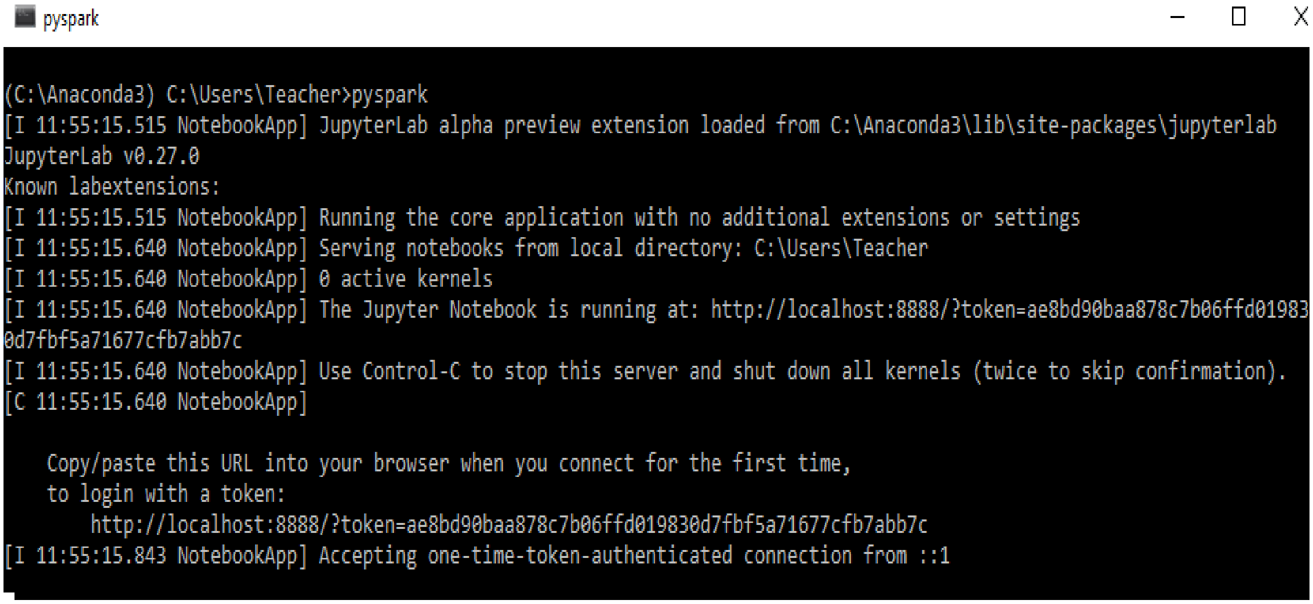
C:\Program Files\Intel\WiFi\bin\
C:\Program Files\Common Files\Intel\WirelessCommon\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
%SPARK_HOME%\bin
%JAVA_HOME%\bin
c:\spark\bin

Fire off PySpark: To fire off PySpark, find Anaconda from the program list via the start menu and select either Anaconda Prompt (or Jupyter Notebook) directly. After opening a untitled notebook, you will see a screen with Jupyter Ipython notebook kernel.

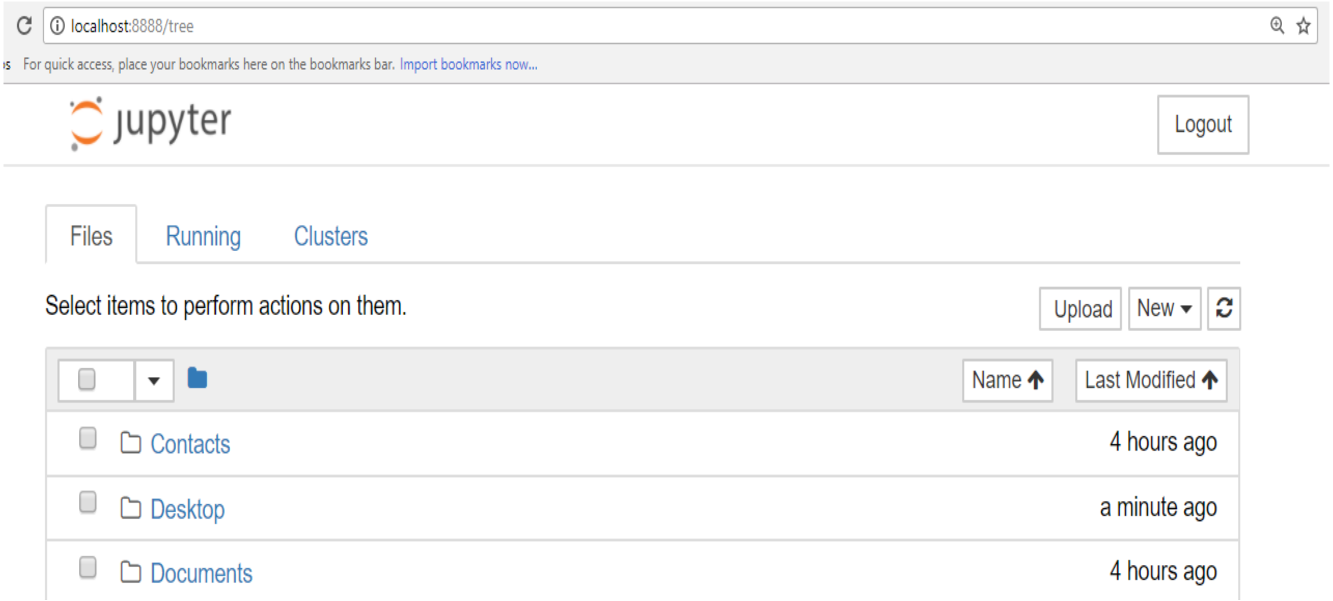




```
(C:\Anaconda3) C:\Users\Teacher>_
```



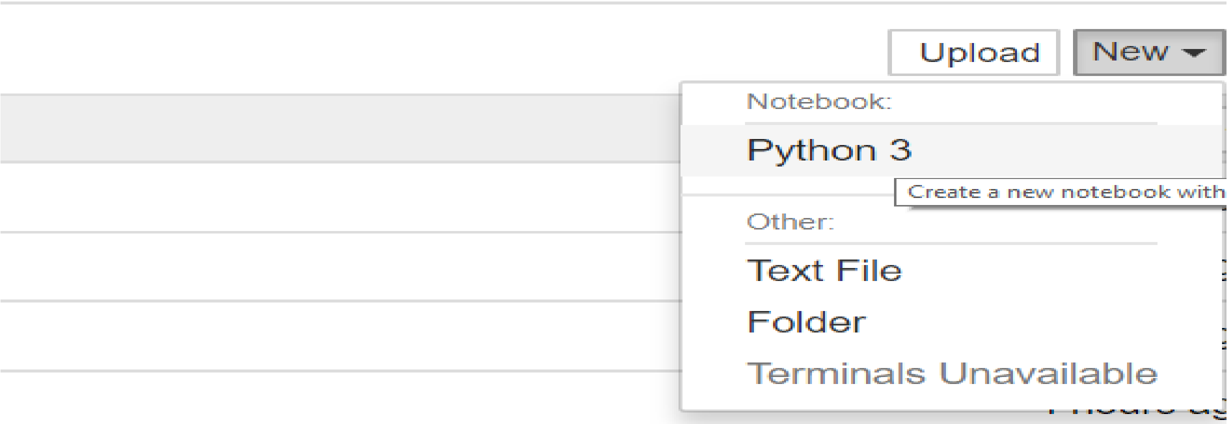
```
(C:\Anaconda3) C:\Users\Teacher>pyspark
[I 11:55:15.515 NotebookApp] JupyterLab alpha preview extension loaded from C:\Anaconda3\lib\site-packages\jupyterlab
JupyterLab v0.27.0
Known labextensions:
[I 11:55:15.515 NotebookApp] Running the core application with no additional extensions or settings
[I 11:55:15.640 NotebookApp] Serving notebooks from local directory: C:\Users\Teacher
[I 11:55:15.640 NotebookApp] 0 active kernels
[I 11:55:15.640 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/?token=ae8bd90baa878c7b06ffd01983
0d7fbf5a71677cfb7abb7c
[I 11:55:15.640 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 11:55:15.640 NotebookApp]

    Copy/paste this URL into your browser when you connect for the first time,
    to login with a token:
        http://localhost:8888/?token=ae8bd90baa878c7b06ffd019830d7fbf5a71677cfb7abb7c
[I 11:55:15.843 NotebookApp] Accepting one-time-token-authenticated connection from ::1
```


localhost:8888/tree
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
jupyter
Logout
Files
Running
Clusters
Upload
New
Name
Last Modified

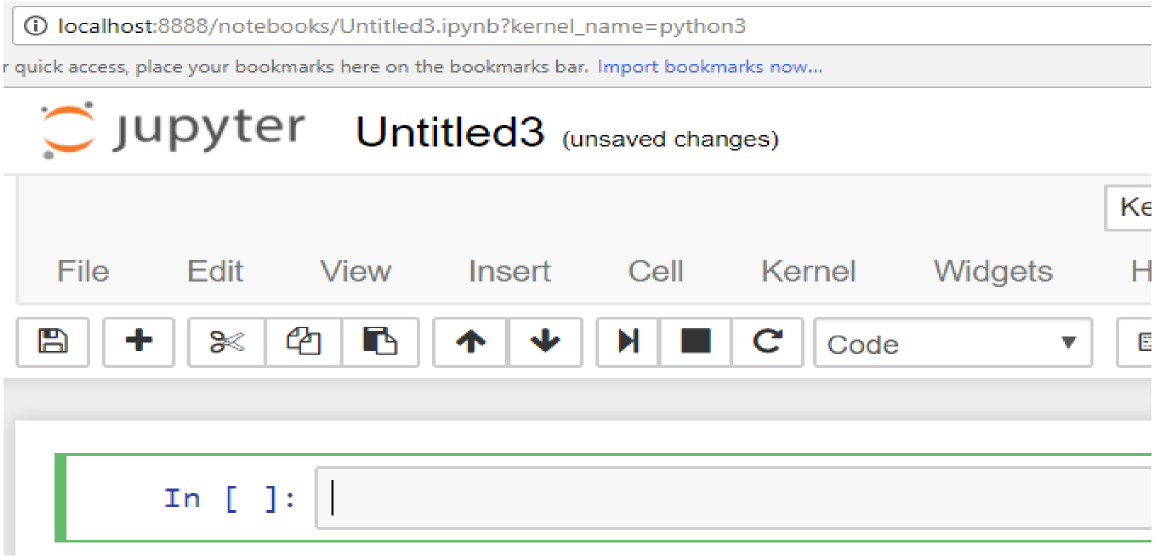
Select items to perform actions on them.

Contacts	4 hours ago
Desktop	a minute ago
Documents	4 hours ago

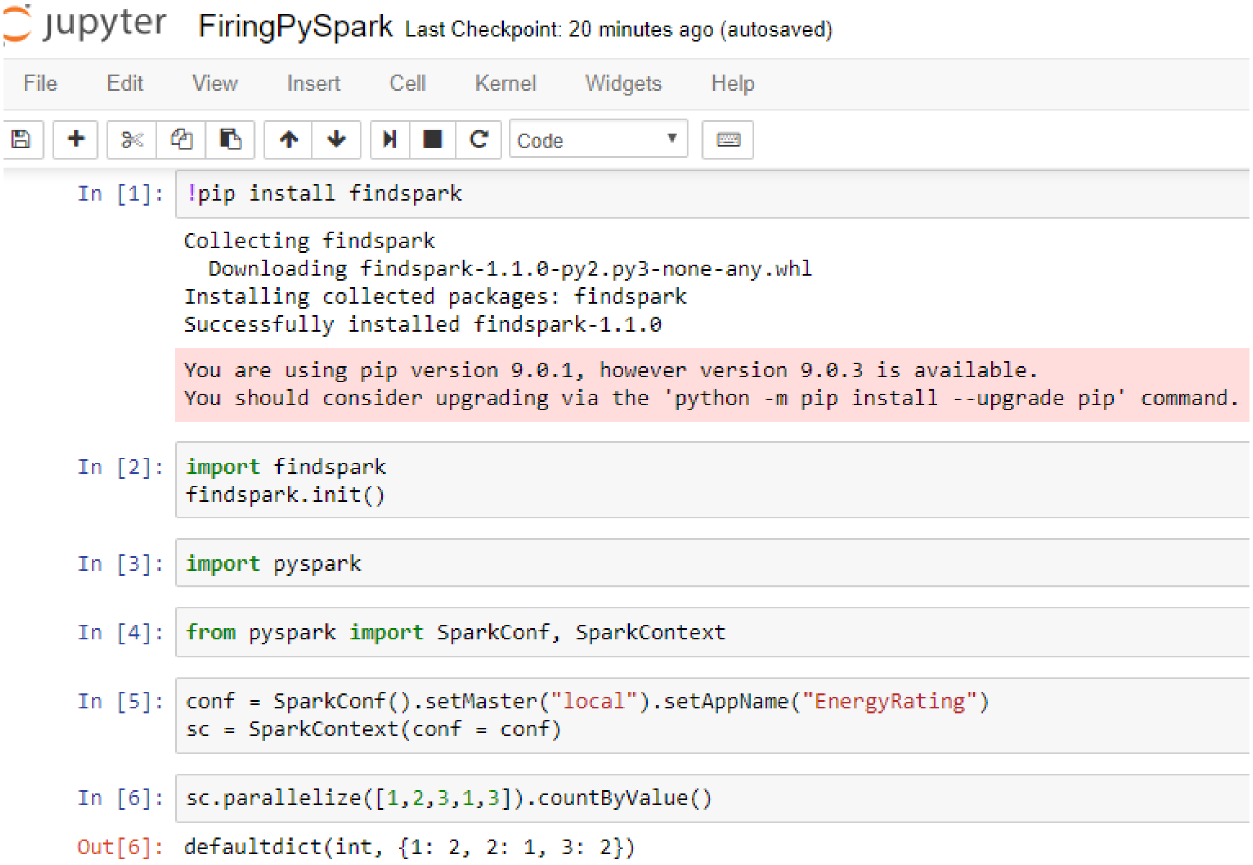
Look carefully the empty and untitled Jupyter Notebook. On the top-right corner of the kernel, click on **New** and open a notebook for **Python 3** (as can be seen below).



Create a new notebook with


localhost:8888/notebooks/Untitled3.ipynb?kernel_name=python3
r quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
jupyter
Untitled3 (unsaved changes)
Ke
File
Edit
View
Insert
Cell
Kernel
Widgets
H
Code
In [ ]:

Rename the file (if and as you wish) and use following commands (shown in the snippet) to call pyspark environment in jupyter notebook.



```
In [1]: !pip install findspark

        Collecting findspark
          Downloading findspark-1.1.0-py2.py3-none-any.whl
        Installing collected packages: findspark
        Successfully installed findspark-1.1.0

        You are using pip version 9.0.1, however version 9.0.3 is available.
        You should consider upgrading via the 'python -m pip install --upgrade pip' command.

In [2]: import findspark
        findspark.init()

In [3]: import pyspark

In [4]: from pyspark import SparkConf, SparkContext

In [5]: conf = SparkConf().setMaster("local").setAppName("EnergyRating")
        sc = SparkContext(conf = conf)

In [6]: sc.parallelize([1,2,3,1,3]).countByValue()

Out[6]: defaultdict(int, {1: 2, 2: 1, 3: 2})
```

1.3 Jupyter Notebook via Ubuntu on VirtualBox for Windows 10

This section has been supplemented in an appendix, please see Appendix A.