

# Simple Step-by-Step Install Guide for Stand-Alone Apache Spark on Windows 10

Jonathan Haynes

Last Updated September 14, 2018

## Overview

*Why use Spark?* Spark is an in-memory computing engine and a set of libraries for parallel data processing on computer clusters. Spark is among the best ways to work with data at scale. Spark supports multiple widely used programming and scripting languages, including Scala, R, Java, SQL, and Python. Spark is frequently used with Hadoop, but does not require Hadoop. (see Chambers and Zaharia (2018), “Spark: The Definitive Guide”, p 3.)

*Why use this install guide?* There are many existing tutorials and blog posts for getting up and running with Spark. I created this tutorial since I didn't find any step-by-step instructions for installing Spark stand-alone on Windows which contained every component I wanted in a single guide, and moreover, a screenshot of each step. If you have a team learning Spark, it's helpful to have them install Spark locally for learning purposes. The purpose of this document is to make it easy for someone with limited programming knowledge to get started quickly.<sup>1</sup>

## Installing Spark on your Windows 10 personal computer

**(1) Download prerequisite installers: Java SE 8u181, Scala 2.12.6, Python 2.7.15, and IntelliJ 2018.2**

Java Installer

<https://www.oracle.com/technetwork/java/javase/downloads/index.html>

Download the full JDK (Java SE 8u181 Development Kit), which includes the Java Runtime Environment.

---

<sup>1</sup> Disclaimer: These steps worked on a range of Lenovo and HP Windows 10 laptops. I'm sharing this guide with the hope that it may be helpful to others; use at your own risk.

Click 'Download':

The screenshot displays the Oracle Java SE Downloads page. The browser's address bar shows 'Java SE - Downloads | Oracle Te X'. The page header includes the Oracle logo, a menu icon, a search bar, and user options like 'Sign In' and 'Country/Region'. The main navigation bar has tabs for 'Overview', 'Downloads', 'Documentation', 'Community', 'Technologies', and 'Training'. The 'Downloads' tab is active, showing 'Java SE Downloads'. A sidebar on the left lists various Java products. The main content area features a 'Java Platform (JDK) 10' download button. Below this, there are sections for 'Java Platform, Standard Edition' and 'Java SE 10.0.2', each with a list of links for installation, release notes, license, and manuals. A red arrow points to the 'JDK DOWNLOAD +' button in the 'Java SE 8u181' section.

Read the license and if you agree, click 'Agree' to the licensing terms, and then select the Windows executable *jdk-8u181-windows-x64.exe* for download:

The screenshot shows the Oracle Java SE Development Kit 8 Downloads page. The page is titled "Java SE Development Kit 8 Downloads" and includes a "Downloads" tab. Below the tab, there is a section for "Java SE Development Kit 8u181" with a license agreement section and a table of download links. Two red arrows point to the "Accept License Agreement" radio button and the "jdk-8u181-windows-x64.exe" download link.

Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.95 MB	<a href="#">jdk-8u181-linux-arm32-vfp-hflt.tar.gz</a>
Linux ARM 64 Hard Float ABI	69.89 MB	<a href="#">jdk-8u181-linux-arm64-vfp-hflt.tar.gz</a>
Linux x86	165.06 MB	<a href="#">jdk-8u181-linux-i586.rpm</a>
Linux x86	179.87 MB	<a href="#">jdk-8u181-linux-i586.tar.gz</a>
Linux x64	162.15 MB	<a href="#">jdk-8u181-linux-x64.rpm</a>
Linux x64	177.05 MB	<a href="#">jdk-8u181-linux-x64.tar.gz</a>
Mac OS X, x64	242.83 MB	<a href="#">jdk-8u181-macosx-x64.dmg</a>
Solaris SPARC 64-bit (SVR4 package)	133.17 MB	<a href="#">jdk-8u181-solaris-sparcv9.tar.Z</a>
Solaris SPARC 64-bit	94.34 MB	<a href="#">jdk-8u181-solaris-sparcv9.tar.gz</a>
Solaris x64 (SVR4 package)	133.83 MB	<a href="#">jdk-8u181-solaris-x64.tar.Z</a>
Solaris x64	92.11 MB	<a href="#">jdk-8u181-solaris-x64.tar.gz</a>
Windows x86	194.41 MB	<a href="#">jdk-8u181-windows-i586.exe</a>
Windows x64	202.73 MB	<a href="#">jdk-8u181-windows-x64.exe</a>

Note, the JRE (Java Runtime Environment) alone is not sufficient since also need the JDK. Also, for some reason, the 10.0.2 version had issues with Spark on some Windows 10 Lenovo laptop configurations that I tested, so that is why version 8 is selected here.

## Scala Installer

Download Scala binaries for Windows; as of 7/30/18, the most recent is scala-2.12.6.msi. Note, IntelliJ will be installed separately later on.

<https://www.scala-lang.org/download/>

The screenshot shows the Scala download page in a browser. At the top, there is a code snippet: `java -version` with a note "(Make sure you have version 1.8.)" and a link to download Java. Below this is a section titled "2 Then, install Scala:" with the text "...either by installing an IDE such as IntelliJ, or sbt, Scala's build tool." There are two red buttons: "DOWNLOAD INTELLIJ" and "DOWNLOAD SBT". Under "DOWNLOAD INTELLIJ" are links for "Getting Started with Scala in IntelliJ", "Building a Scala Project with IntelliJ and sbt", and "Testing Scala in IntelliJ with ScalaTest". Under "DOWNLOAD SBT" are links for "Getting Started with Scala and sbt on the Command Line" and "Testing Scala with sbt and ScalaTest on the Command Line". Red arrows point from the buttons to their respective sections. Below these are two red annotations: "Best if you prefer a full-featured IDE (recommended for beginners)" and "Best if you are familiar with the command line". Further down is a section "Other ways to install Scala" with a red arrow pointing to a menu icon. The list includes "Download the Scala binaries for windows" (with a link "Need help running the binaries?") and "Use Scastie to run single-file Scala programs in your browser using multiple Scala compilers; the production Scala 2.x compilers, Scala.js, Dotty, and Typelevel Scala. Save and share executable Scala code snippets."

`java -version` (Make sure you have version 1.8.)  
(If you don't have it installed, [download Java here.](#))

**2** Then, install Scala:  
...either by installing an IDE such as IntelliJ, or sbt, Scala's build tool.

**DOWNLOAD INTELLIJ** or **DOWNLOAD SBT**

Getting Started with Scala in IntelliJ  
Building a Scala Project with IntelliJ and sbt  
Testing Scala in IntelliJ with ScalaTest

Getting Started with Scala and sbt on the Command Line  
Testing Scala with sbt and ScalaTest on the Command Line

*Best if you prefer a full-featured IDE (recommended for beginners)*

*Best if you are familiar with the command line*

Compared to other programming languages, installing Scala is a bit unusual. Scala is unusual because it is usually installed for each of your Scala projects rather than being installed system-wide. Both of the above options manage (via sbt) a specific Scala version per Scala project you create.

But it's also possible to "install" Scala in numerous other ways; e.g., grab Scala binaries and use Scala from the command line or use Scala in your browser!

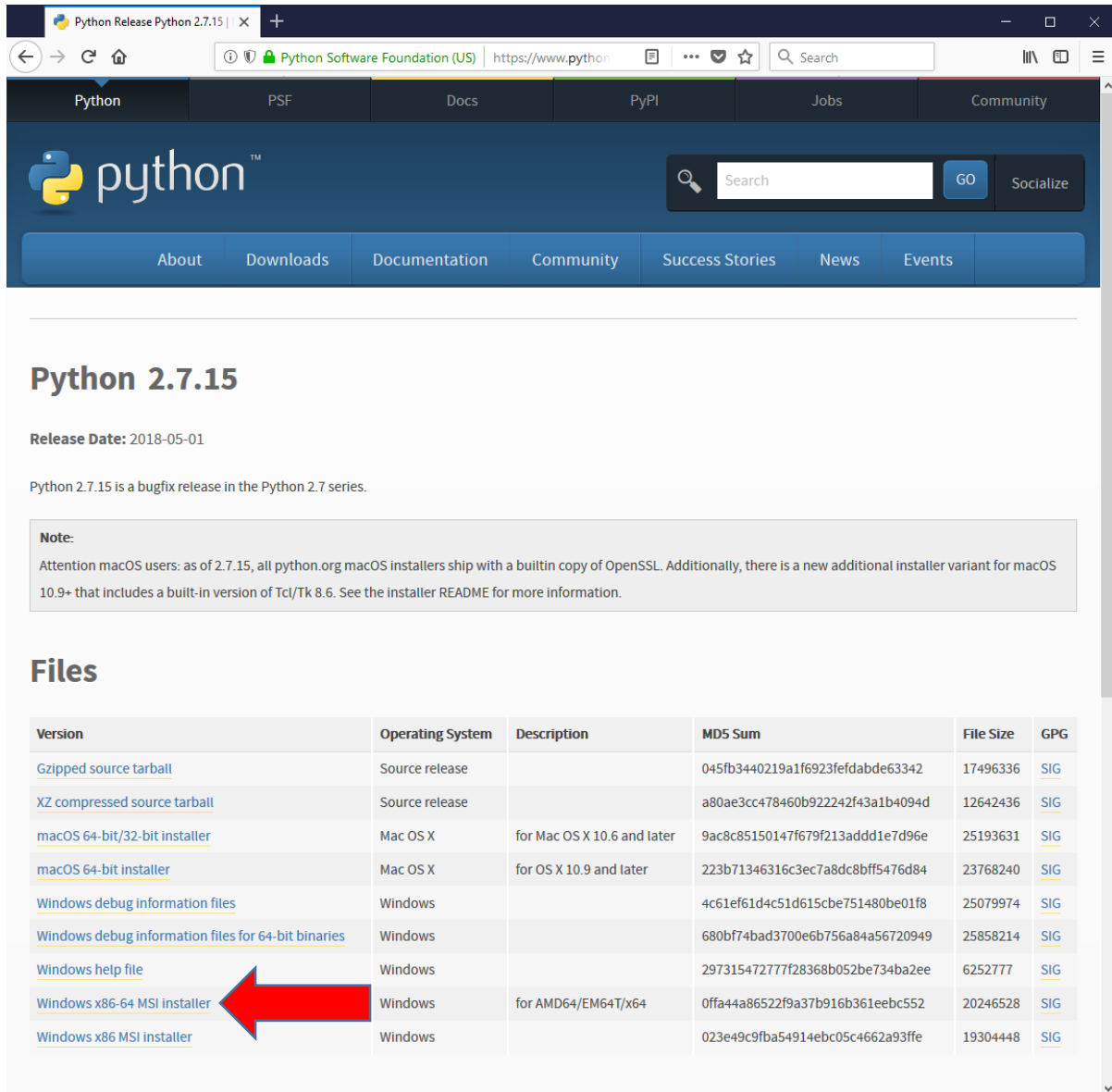
**Other ways to install Scala**

- [Download the Scala binaries for windows](#)  
*Need help running the binaries?*
- Use [Scastie](#) to run single-file Scala programs in your browser using multiple Scala compilers; the production Scala 2.x compilers, Scala.js, Dotty, and Typelevel Scala. *Save and share executable Scala code snippets.*

## Python Installer

Download the latest Python 2.7 release; as of 7/30/18, the most recent is python-2.7.15. Be sure to download the x64 bit version, *python-2.7.15.amd64.msi*.

<https://www.python.org/downloads/windows/>



**Python 2.7.15**

**Release Date:** 2018-05-01

Python 2.7.15 is a bugfix release in the Python 2.7 series.

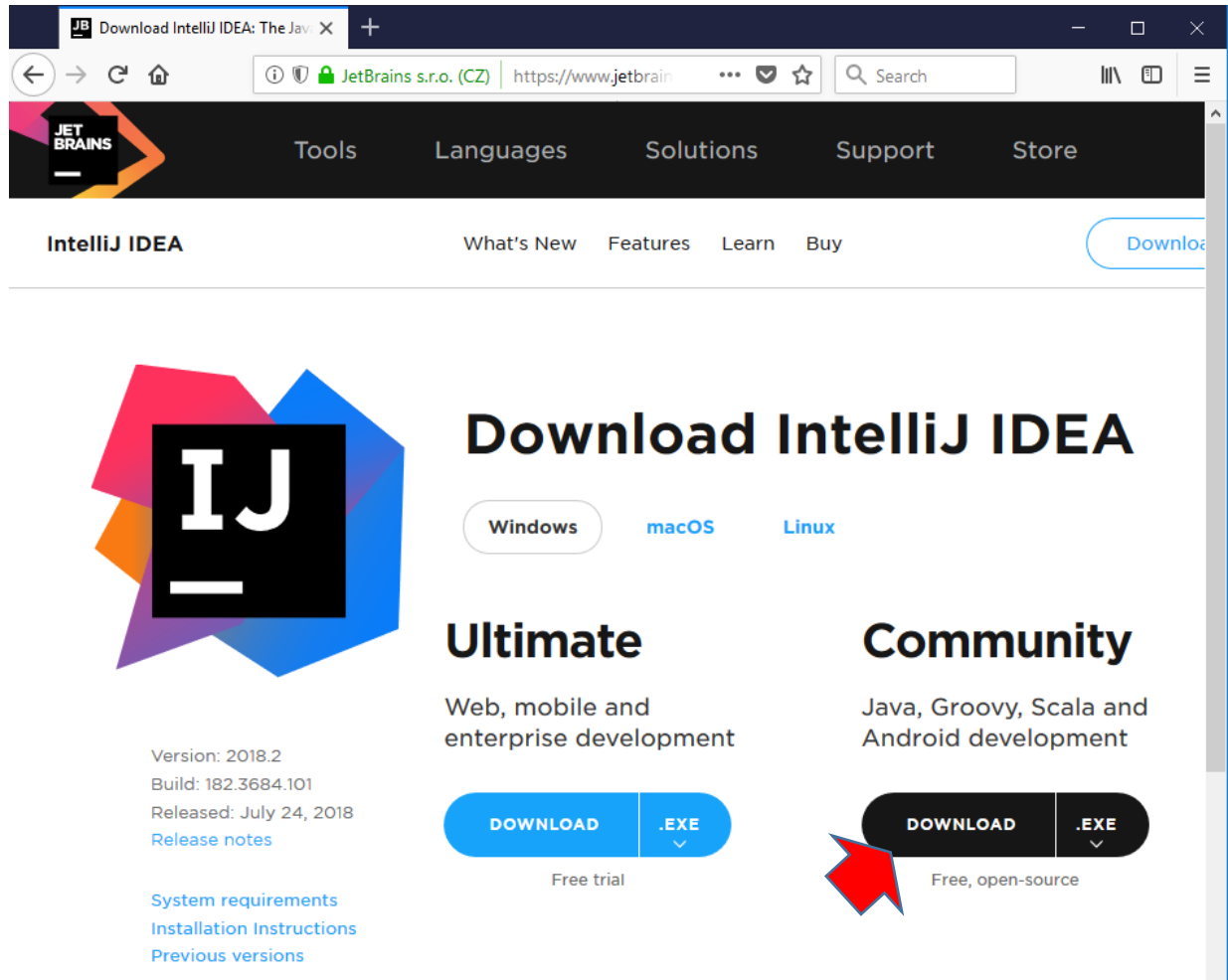
**Note:**  
Attention macOS users: as of 2.7.15, all python.org macOS installers ship with a builtin copy of OpenSSL. Additionally, there is a new additional installer variant for macOS 10.9+ that includes a built-in version of Tcl/Tk 8.6. See the installer README for more information.

**Files**

Version	Operating System	Description	MD5 Sum	File Size	GPG
<a href="#">Gzipped source tarball</a>	Source release		045fb3440219a1f6923fefdbabde63342	17496336	<a href="#">SIG</a>
<a href="#">XZ compressed source tarball</a>	Source release		a80ae3cc478460b922242f43a1b4094d	12642436	<a href="#">SIG</a>
<a href="#">macOS 64-bit/32-bit installer</a>	Mac OS X	for Mac OS X 10.6 and later	9ac8c85150147f679f213addd1e7d96e	25193631	<a href="#">SIG</a>
<a href="#">macOS 64-bit installer</a>	Mac OS X	for OS X 10.9 and later	223b71346316c3ec7a8dc8bff5476d84	23768240	<a href="#">SIG</a>
<a href="#">Windows debug information files</a>	Windows		4c61ef61d4c51d615cbe751480be01f8	25079974	<a href="#">SIG</a>
<a href="#">Windows debug information files for 64-bit binaries</a>	Windows		680bf74bad3700e6b756a84a56720949	25858214	<a href="#">SIG</a>
<a href="#">Windows help file</a>	Windows		297315472777f28368b052be734ba2ee	6252777	<a href="#">SIG</a>
<a href="#">Windows x86-64 MSI installer</a>	Windows	for AMD64/EM64T/x64	0ffa44a86522f9a37b916b361eebc552	20246528	<a href="#">SIG</a>
<a href="#">Windows x86 MSI installer</a>	Windows		023e49c9fba54914ebc05c4662a93ffe	19304448	<a href="#">SIG</a>

## IntelliJ Installer

Download the free community edition: <https://www.jetbrains.com/idea/download/#section=windows>



The screenshot shows the JetBrains website's download page for IntelliJ IDEA. The browser address bar shows the URL <https://www.jetbrains.com/idea/download/#section=windows>. The page features the IntelliJ IDEA logo on the left, with version details: Version: 2018.2, Build: 182.3684.101, Released: July 24, 2018, and a link to Release notes. Below the logo are links for System requirements, Installation Instructions, and Previous versions. The main content area is titled 'Download IntelliJ IDEA' and offers three operating system options: Windows (selected), macOS, and Linux. Two product editions are presented: 'Ultimate' (Web, mobile and enterprise development) with a 'Free trial' offer, and 'Community' (Java, Groovy, Scala and Android development) with a 'Free, open-source' offer. Each edition has a 'DOWNLOAD .EXE' button. A red arrow points to the 'DOWNLOAD .EXE' button for the Community edition.

## (2) Install software

The four installers you'll need to have ready are:

*jdk-8u181-windows-x64.exe*

*ideaIC-2018.2.exe*

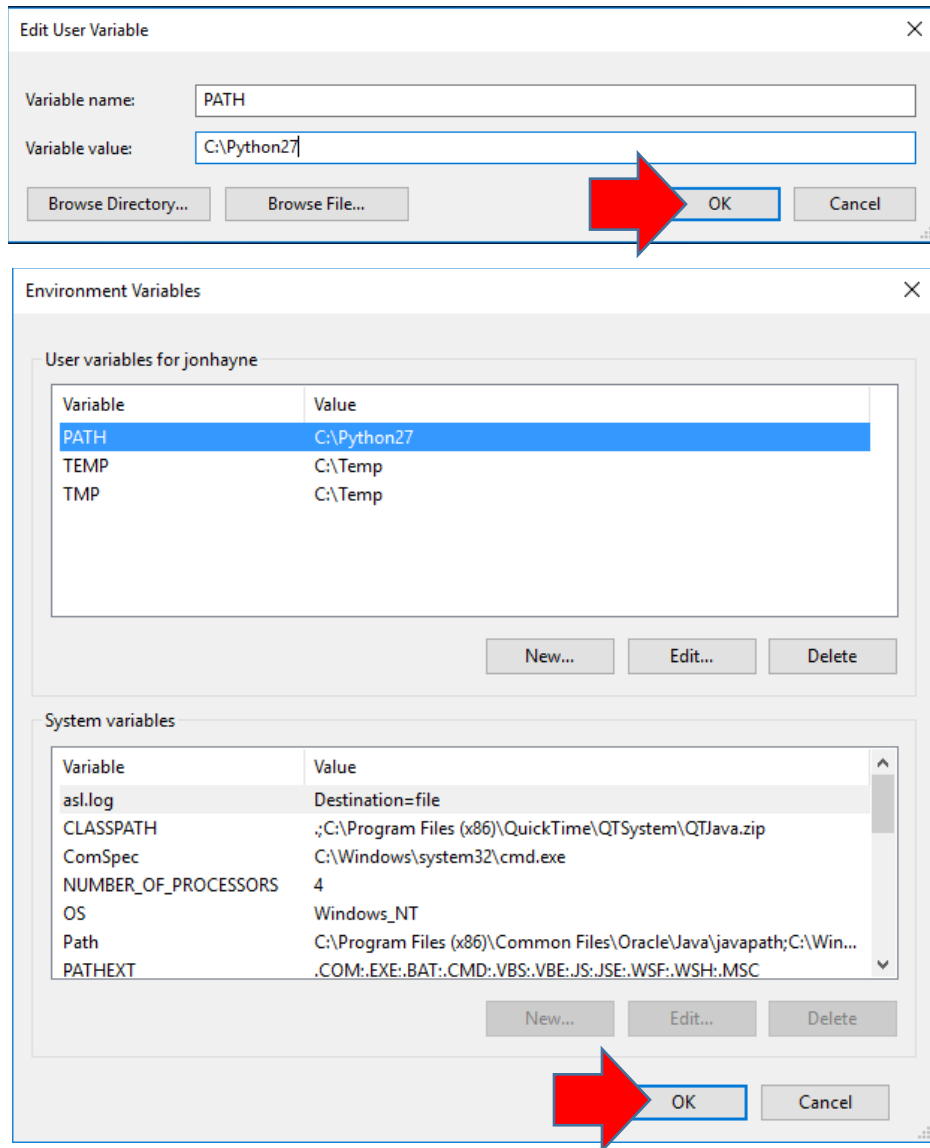
*scala-2.12.6.msi*

*python-2.7.15.amd64.msi*

Install all with default options selected.

### (3) Update environment variables

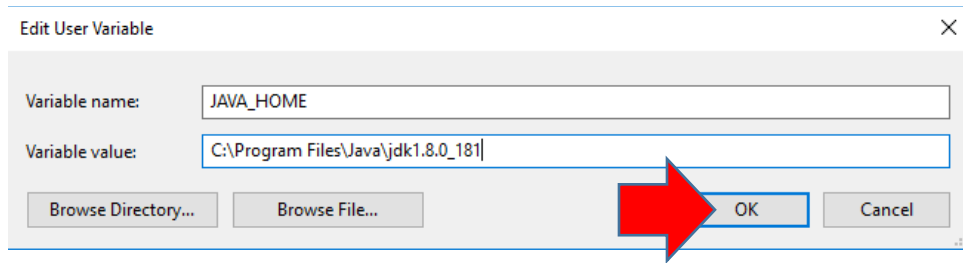
Add Python to the environmental PATH variable. Enter 'path' in the windows search bar and select 'Edit environment variables for your account'. If no 'PATH' variable already exists, then create a new user variable called 'PATH' with the value 'C:\Python27':



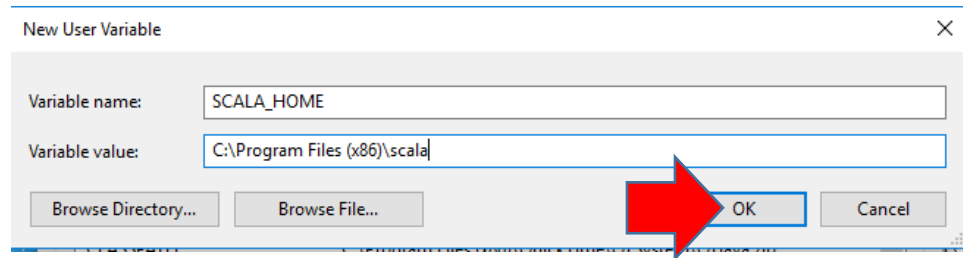
If a user variable named 'PATH' already exists, then do not delete it, but instead select it and click 'Edit'. Then add ';C:\Python27' to the end of the string and click 'OK' to modify the variable.

Now open a command prompt and type 'python'. Python should launch and show a >>> prompt.

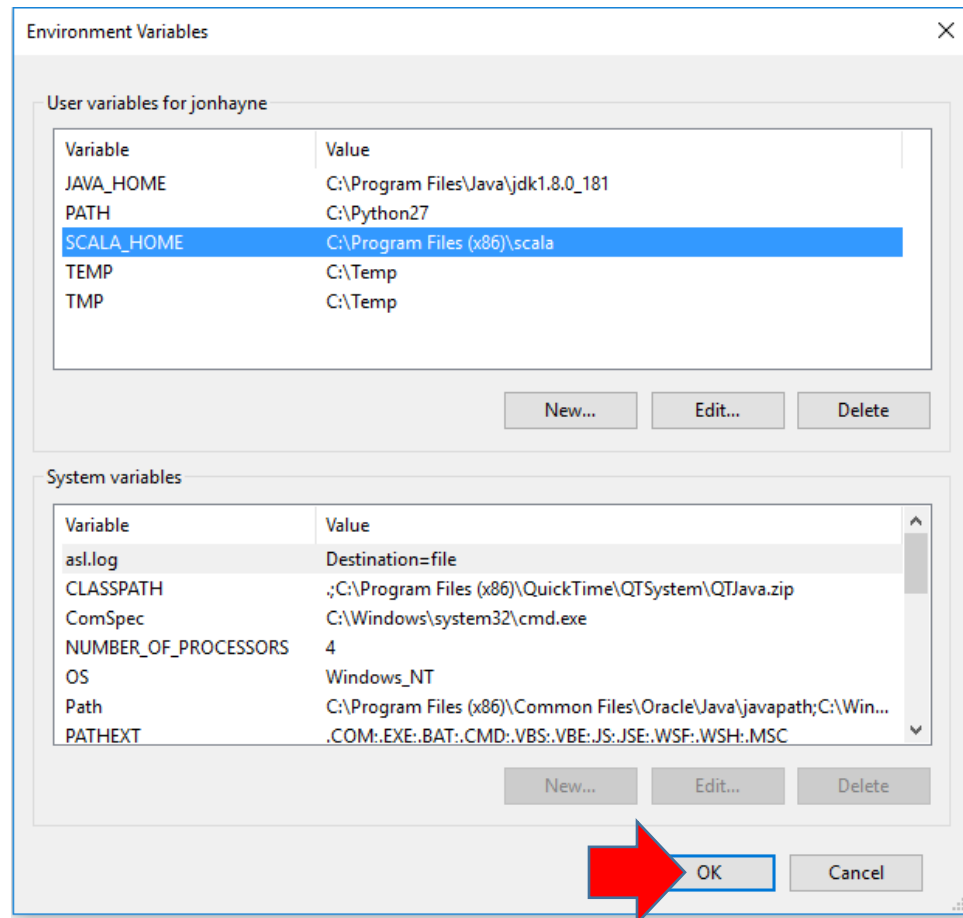
Create a new user variable called 'JAVA\_HOME' with the value 'C:\Program Files\Java\jdk1.8.0\_181':



Finally, do the same for Scala:



Double-check that everything is spelled correctly, then click 'OK':





#### (4) Install Spark

Open a browser and navigate to <https://spark.apache.org/downloads.html>

Download Spark by clicking the link:

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.3.1-bin-hadoop2.7.tgz](#)
4. Verify this release using the [2.3.1 signatures and checksums](#) and [project release KEYS](#).

Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with [Scala 2.10 support](#).

Link with Spark

Spark artifacts are hosted in [Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.11
version: 2.3.1
```

Installing with PyPi

PySpark is now available in pypi. To install just run `pip install pyspark`.

Latest News

- Spark+AI Summit (October 2-4th, 2018, London) agenda posted (Jul 24, 2018)
- Spark 2.2.2 released (Jul 02, 2018)
- Spark 2.1.3 released (Jun 29, 2018)
- Spark 2.3.1 released (Jun 08, 2018)

APACHECON North America September 24-27, 2018 Montréal, Canada

Download Spark

Built-in Libraries:

After download, verify the integrity of the .tgz file using the Windows certUtil application. This will validate the SHA512 checksum.

For example, if the file `spark-2.3.1-bin-hadoop2.7.tgz` is on your Desktop, go to the Windows search bar and type 'cmd' and open a command prompt. Then enter 'cd Desktop', then 'certutil -hashfile spark-2.3.1-bin-hadoop2.7.tgz SHA512'.

This should return the following checksum:

*SHA512 hash of file spark-2.3.1-bin-hadoop2.7.tgz:*

*dc 3a 97 f3 d9 97 91 d3 63 e4 f7 0a 62 2b 84 d6 e3 13 bd 85 2f 6f db c7 77 d3 1e ab 44 cb c1 12 ce ea a2  
Of 7b f8 35 49 2f b6 54 f4 8a e5 7e 99 69 f9 3d 3b 0e 6e c9 20 76 d1 c5 e1 b4 0b 46 96*

```

C:\Users\jonhayne>cd Desktop

C:\Users\jonhayne\Desktop>dir
Volume in drive C is Windows
Volume Serial Number is

Directory of C:\Users\jonhayne\Desktop

07/30/2018  01:38 PM    <DIR>          .
07/30/2018  01:38 PM    <DIR>          ..
07/30/2018  12:48 PM           20,246,528  python-2.7.15.amd64.msi
06/01/2018  03:49 PM    <DIR>          spark-2.3.1-bin-hadoop2.7
07/30/2018  09:35 AM           225,883,783  spark-2.3.1-bin-hadoop2.7.tgz
07/30/2018  01:38 PM              268  spark-2.3.1-bin-hadoop2.7.tgz.sha512
07/30/2018  01:22 PM    <DIR>          spark_class
                4 File(s)      246,173,441 bytes
                7 Dir(s)   119,660,699,648 bytes free

C:\Users\jonhayne\Desktop>certutil -hashfile spark-2.3.1-bin-hadoop2.7.tgz SHA512
SHA512 hash of file spark-2.3.1-bin-hadoop2.7.tgz:
dc 3a 97 f3 d9 97 91 d3 63 e4 f7 0a 62 2b 84 d6 e3 13 bd 85 2f 6f db c7 77 d3 1e ab 44
cb c1 12 ce ea a2 0f 7b f8 35 49 2f b6 54 f4 8a e5 7e 99 69 f9 3d 3b 0e 6e c9 20 76 d1
c5 e1 b4 0b 46 96
CertUtil: -hashfile command completed successfully.

C:\Users\jonhayne\Desktop>

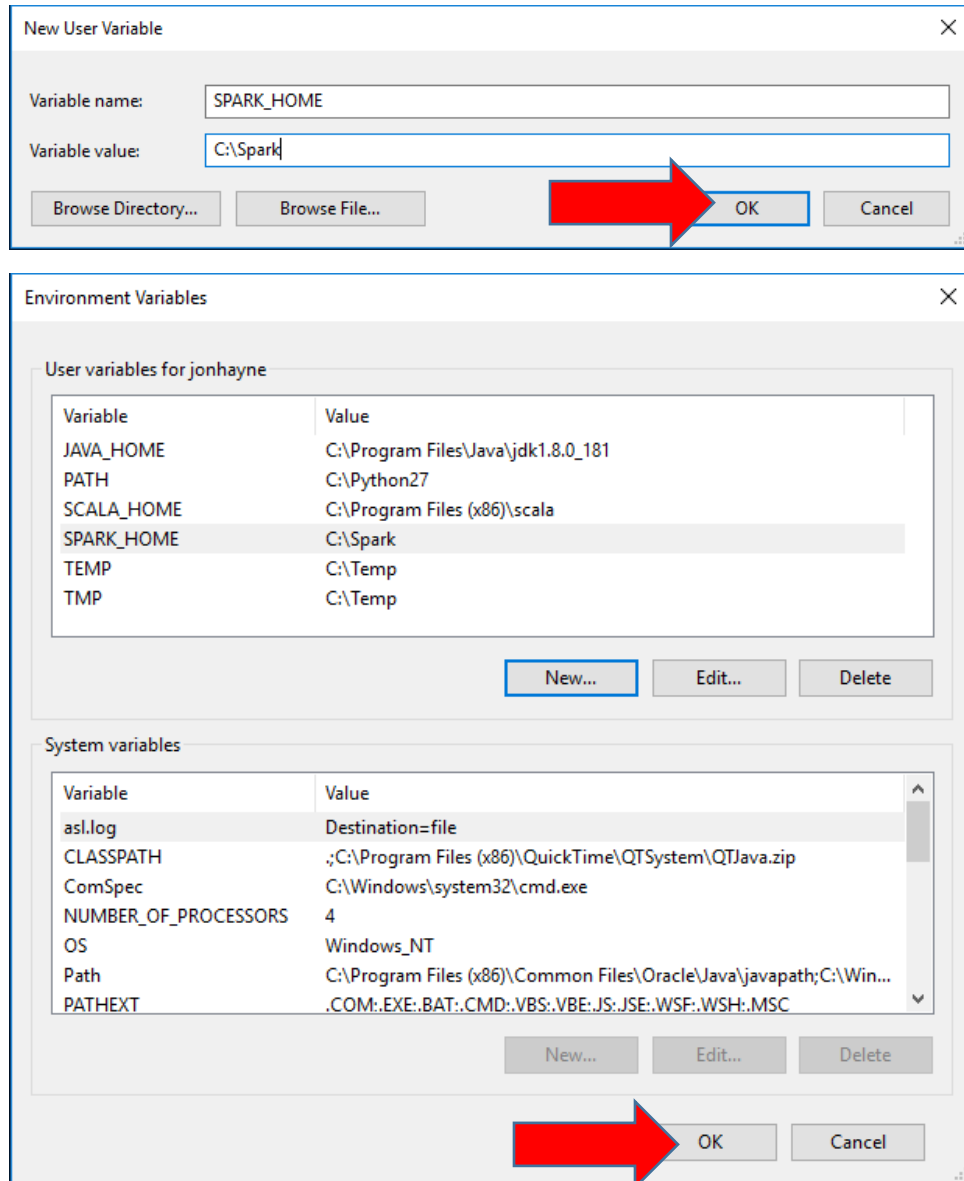
```

If you download a different version of this file the SHA512 checksum will be different, in which case, you'll need to check the signatures and checksums download page for the matching .SHA512 file:

Name	Last modified	Size	Description
<a href="#">Parent Directory</a>	-	-	-
<a href="#">SparkR 2.3.1.tar.gz</a>	2018-06-01 20:59	303K	
<a href="#">SparkR 2.3.1.tar.gz.asc</a>	2018-06-01 20:59	819	
<a href="#">SparkR 2.3.1.tar.gz.sha512</a>	2018-06-01 20:59	207	
<a href="#">pyspark-2.3.1.tar.gz</a>	2018-06-01 20:59	202M	
<a href="#">pyspark-2.3.1.tar.gz.asc</a>	2018-06-01 20:59	819	
<a href="#">pyspark-2.3.1.tar.gz.sha512</a>	2018-06-01 20:59	210	
<a href="#">spark-2.3.1-bin-hadoop2.6.tgz</a>	2018-06-01 20:59	214M	
<a href="#">spark-2.3.1-bin-hadoop2.6.tgz.asc</a>	2018-06-01 20:59	819	
<a href="#">spark-2.3.1-bin-hadoop2.6.tgz.sha512</a>	2018-06-01 20:59	268	
<a href="#">spark-2.3.1-bin-hadoop2.7.tgz</a>	2018-06-01 20:59	215M	
<a href="#">spark-2.3.1-bin-hadoop2.7.tgz.asc</a>	2018-06-01 20:59	819	
<a href="#">spark-2.3.1-bin-hadoop2.7.tgz.sha512</a>	2018-06-01 20:59	268	
<a href="#">spark-2.3.1-bin-without-hadoop.tgz</a>	2018-06-01 20:59	147M	
<a href="#">spark-2.3.1-bin-without-hadoop.tgz.asc</a>	2018-06-01 20:59	819	
<a href="#">spark-2.3.1-bin-without-hadoop.tgz.sha512</a>	2018-06-01 20:59	288	
<a href="#">spark-2.3.1.tgz</a>	2018-06-01 20:59	15M	
<a href="#">spark-2.3.1.tgz.asc</a>	2018-06-01 20:59	819	
<a href="#">spark-2.3.1.tgz.sha512</a>	2018-06-01 20:59	195	

Unzip and unpack the contents of the tgz file to the directory C:\Spark

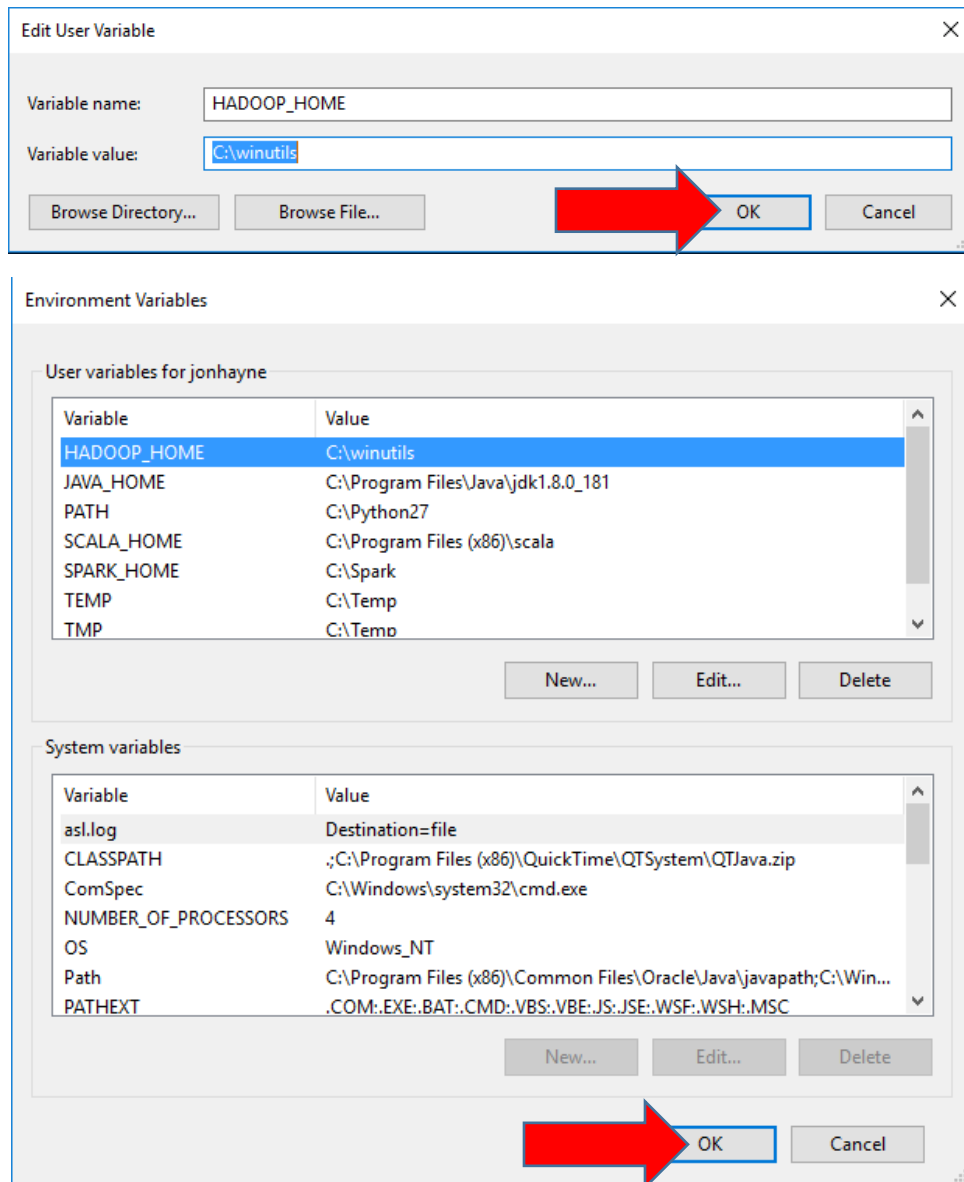
Update the environment variables. Enter 'path' in the windows search bar and select 'Edit environment variables for your account'. Then create a new user variable called 'SPARK\_HOME' with the value 'C:\Spark':



### (5) Download winutils.exe

We'll be installing Spark without connecting to a Hadoop cluster, but are using a compiled version of Spark which expects Hadoop to be present. The solution to this is to download an executable called winutils.exe, available here: <https://github.com/stevloughran/winutils/tree/master/hadoop-2.6.0/bin>

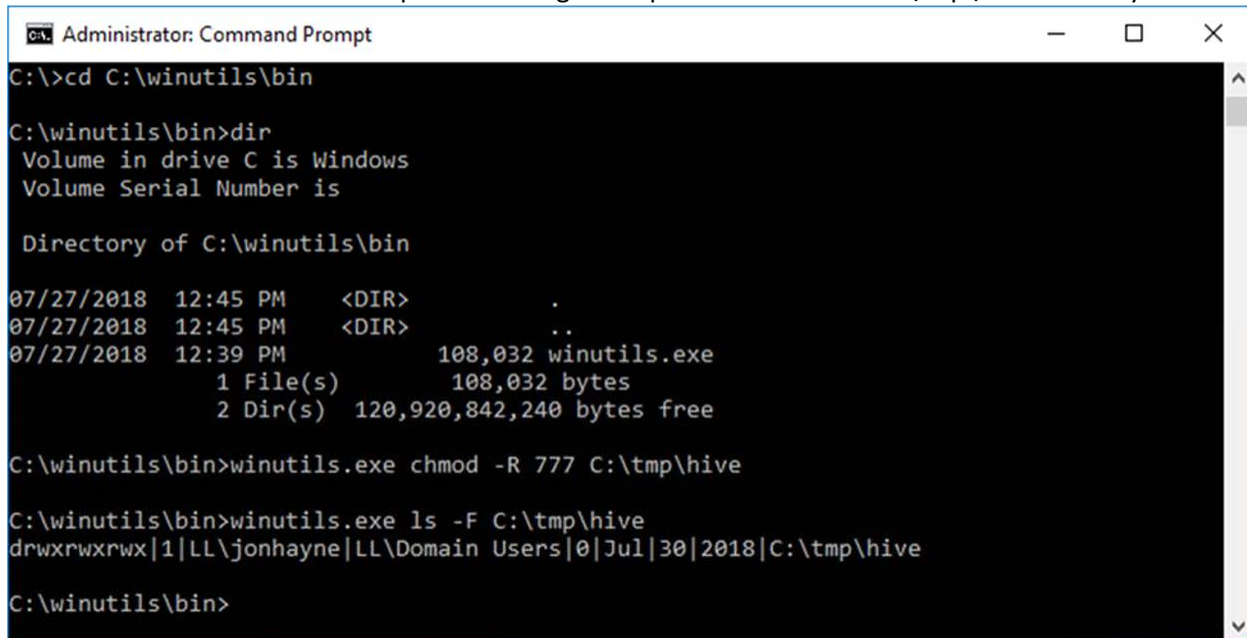
Create a directory C:\winutils\bin\ and put winutils.exe there, then add a HADOOP\_HOME path to the environment variables where the value is C:\winutils:



Finally, you need to make the temporary directories writable by Spark. Confirm that the directories C:\tmp\hive and C:\tmp\mydir exist. If not, create them. Then open a command prompt as an administrator and type the following:

```
C:\>cd C:\winutils\bin
C:\>winutils.exe chmod -R 777 C:\tmp\hive
C:\>winutils.exe chmod -R 777 C:\tmp\mydir
```

See below for a screenshot example that changes the permissions on the C:\tmp\hive directory:



```
Administrator: Command Prompt
C:\>cd C:\winutils\bin
C:\winutils\bin>dir
Volume in drive C is Windows
Volume Serial Number is

Directory of C:\winutils\bin
07/27/2018 12:45 PM <DIR>      .
07/27/2018 12:45 PM <DIR>      ..
07/27/2018 12:39 PM          108,032 winutils.exe
                1 File(s)          108,032 bytes
                2 Dir(s)      120,920,842,240 bytes free

C:\winutils\bin>winutils.exe chmod -R 777 C:\tmp\hive

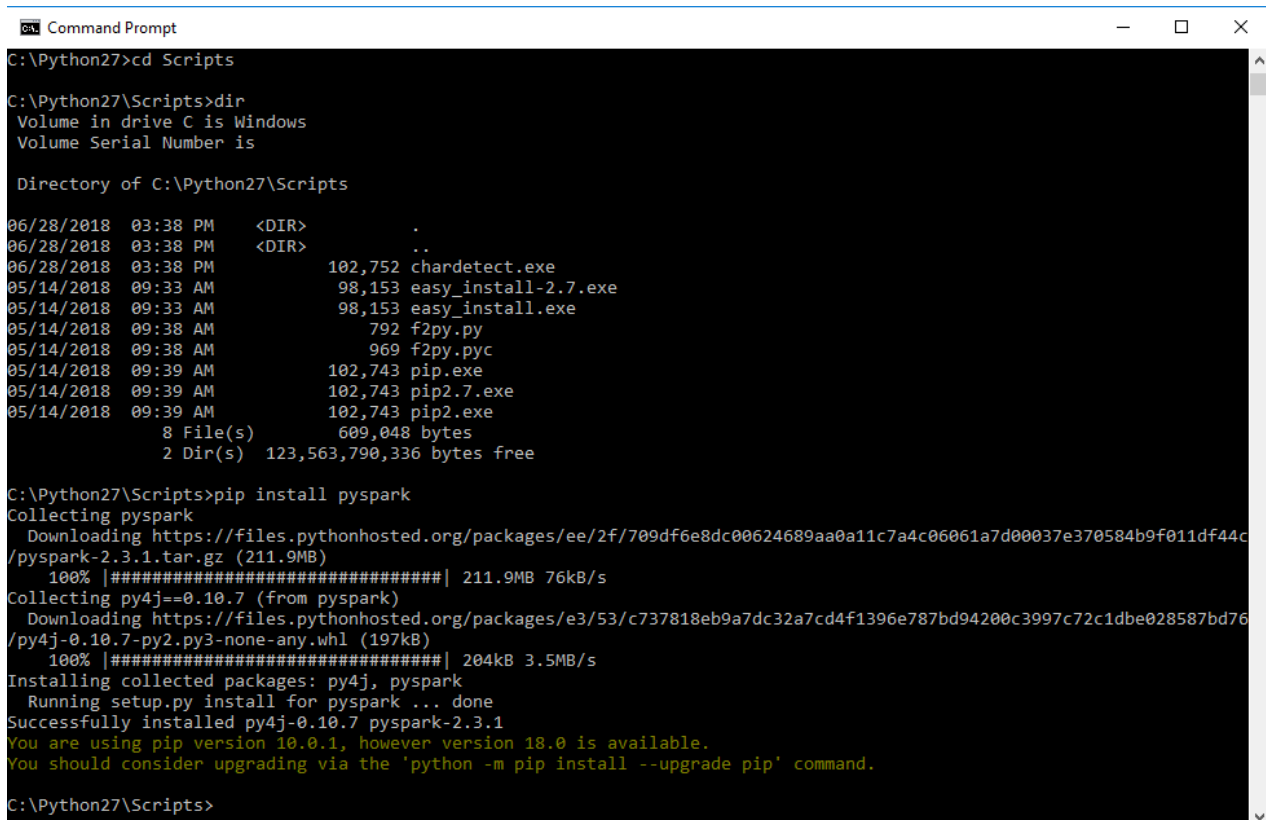
C:\winutils\bin>winutils.exe ls -F C:\tmp\hive
drwxrwxrwx|1|LL\jonhayne|LL\Domain Users|0|Jul|30|2018|C:\tmp\hive

C:\winutils\bin>
```

## (6) Install PySpark

Open a command prompt and navigate to the directory C:\Python27\Scripts.

Type: *pip install pyspark*



```
Command Prompt
C:\Python27>cd Scripts
C:\Python27\Scripts>dir
Volume in drive C is Windows
Volume Serial Number is

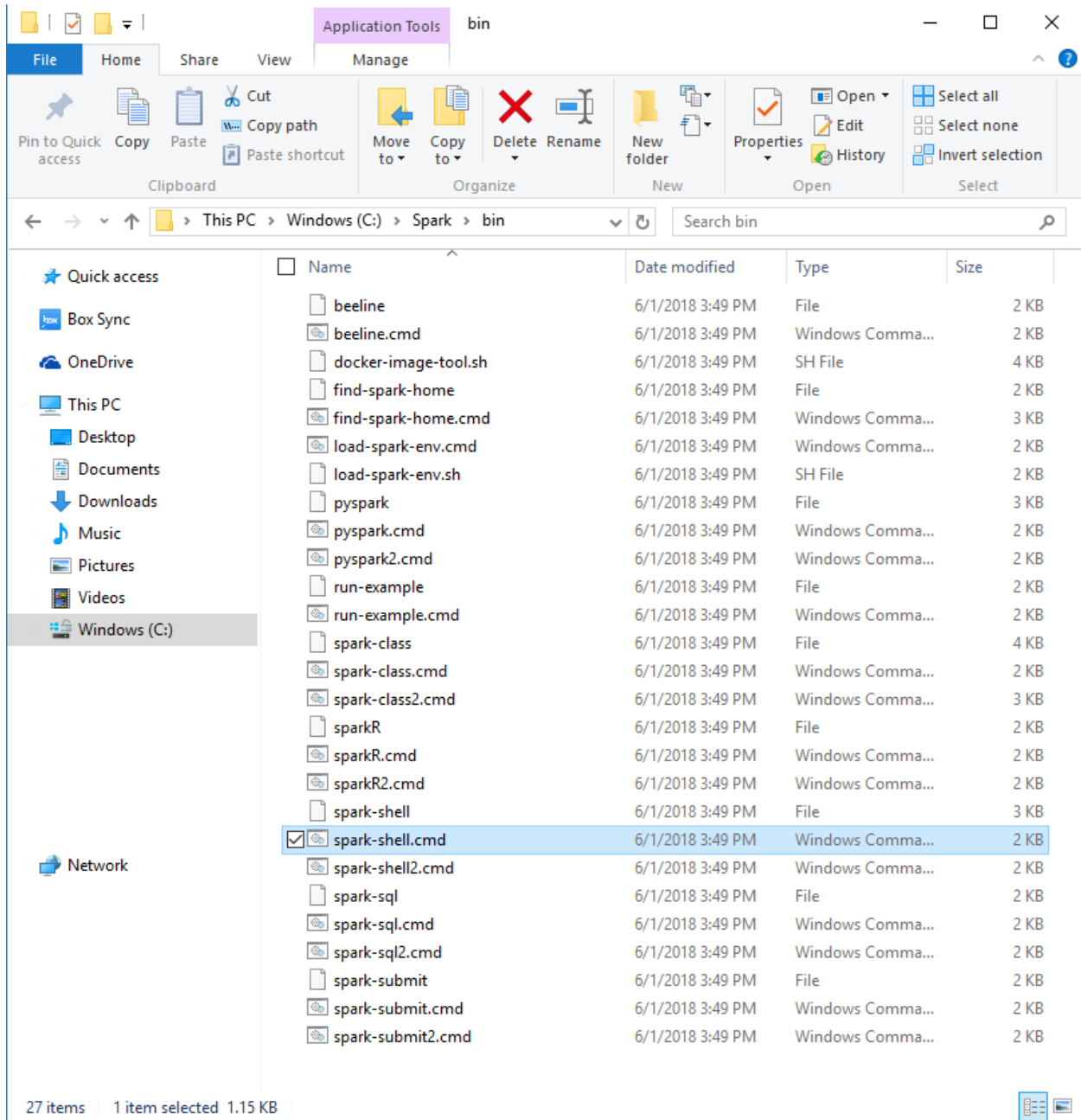
Directory of C:\Python27\Scripts
06/28/2018 03:38 PM <DIR>      .
06/28/2018 03:38 PM <DIR>      ..
06/28/2018 03:38 PM          102,752 chardetect.exe
05/14/2018 09:33 AM          98,153 easy_install-2.7.exe
05/14/2018 09:33 AM          98,153 easy_install.exe
05/14/2018 09:38 AM           792 f2py.py
05/14/2018 09:38 AM           969 f2py.pyc
05/14/2018 09:39 AM          102,743 pip.exe
05/14/2018 09:39 AM          102,743 pip2.7.exe
05/14/2018 09:39 AM          102,743 pip2.exe
                8 File(s)          609,048 bytes
                2 Dir(s)      123,563,790,336 bytes free

C:\Python27\Scripts>pip install pyspark
Collecting pyspark
  Downloading https://files.pythonhosted.org/packages/ee/2f/709df6e8dc00624689aa0a11c7a4c06061a7d00037e370584b9f011df44c/pyspark-2.3.1.tar.gz (211.9MB)
    100% |#####| 211.9MB 76kB/s
Collecting py4j==0.10.7 (from pyspark)
  Downloading https://files.pythonhosted.org/packages/e3/53/c737818eb9a7dc32a7cd4f1396e787bd94200c3997c72c1dbe028587bd76/py4j-0.10.7-py2.py3-none-any.whl (197kB)
    100% |#####| 204kB 3.5MB/s
Installing collected packages: py4j, pyspark
  Running setup.py install for pyspark ... done
Successfully installed py4j-0.10.7 pyspark-2.3.1
You are using pip version 10.0.1, however version 18.0 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.

C:\Python27\Scripts>
```

## (7) Test Spark

Launch the Scala console in the Spark directory to confirm it works, by clicking *spark-shell.cmd*



If installation was successful, you will see a prompt as follows:

```
C:\Windows\system32\cmd.exe
2018-07-27 12:47:23 WARN NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-
java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://
Spark context available as 'sc' (master = local[*], app id = local-1532713653131).
Spark session available as 'spark'.
Welcome to

  ____
 /  __ \
/   /  \
/_____/  version 2.3.1

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_172)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Test the first two examples from the book, *Spark: The Definitive Guide*, creating a Spark DataFrame and performing a calculation, see pages 17 & 19:

```
scala>val myRange = spark.range(1000).toDF("number")
scala>myRange.show(10)
scala>val divisBy2 = myRange.where("number % 2 = 0")
scala>divisBy2.show(10)
```

This should return the following:

```
scala> val myRange = spark.range(1000).toDF("number")
2018-08-02 10:32:50 WARN ObjectStore:568 - Failed to get database global_temp, returning NoSuchObjectException
myRange: org.apache.spark.sql.DataFrame = [number: bigint]

scala> myRange.show(10)
+-----+
|number|
+-----+
|  0|
|  1|
|  2|
|  3|
|  4|
|  5|
|  6|
|  7|
|  8|
|  9|
+-----+
only showing top 10 rows
```

```
scala> val divisBy2 = myRange.where("number % 2 = 0")
divisBy2: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [number: bigint]

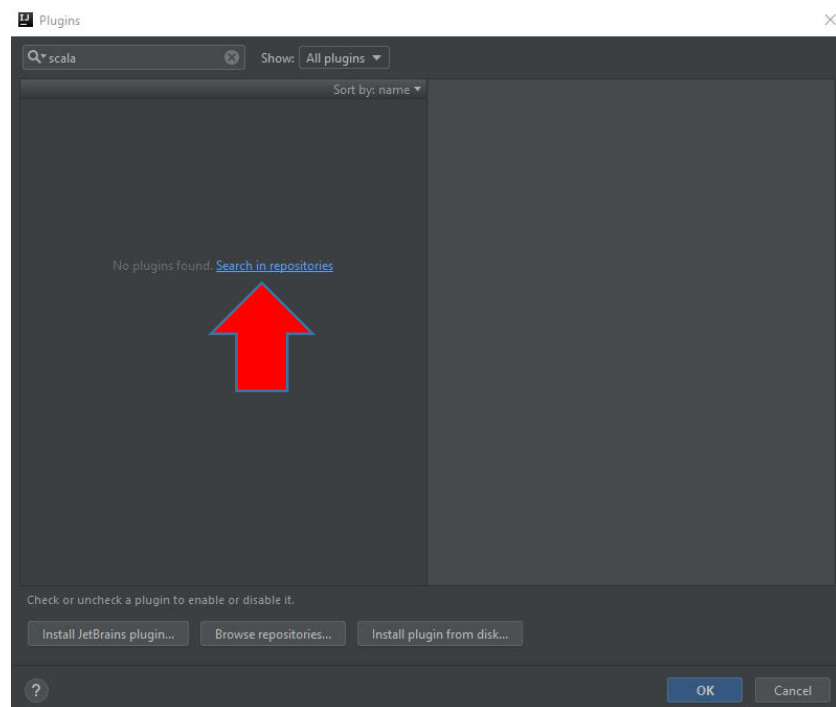
scala> divisBy2.show(10)
+-----+
|number|
+-----+
|  0|
|  2|
|  4|
|  6|
|  8|
| 10|
| 12|
| 14|
| 16|
| 18|
+-----+
only showing top 10 rows
```

## (8) IntelliJ, community edition

Find and launch the new application IntelliJ in the Program menu, "All Apps". If it installed successfully, you should see the following:

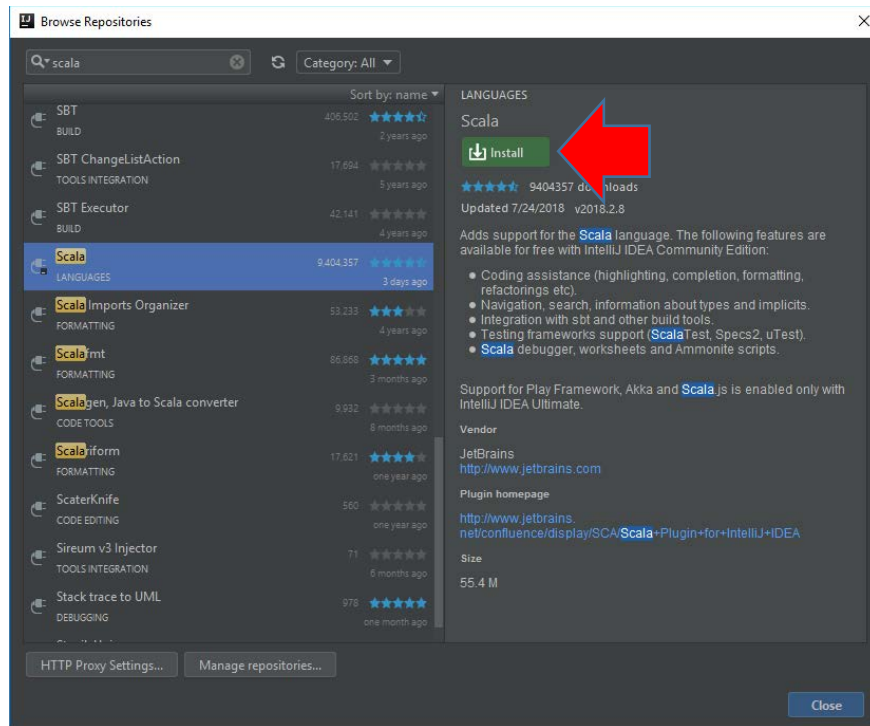


Now you need to install the Scala plugin. Select 'Configure' and then 'Plugin', and then enter 'Scala' and click 'Search in repositories'.

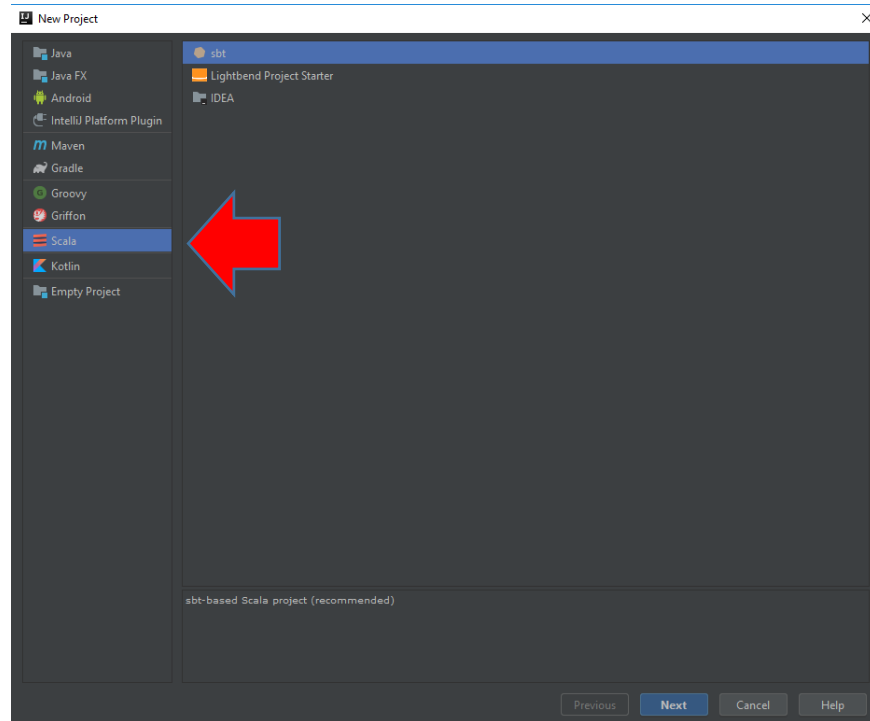




Scroll down to 'Scala', click 'Install':



When you relaunch IntelliJ, it should now be possible to create a Scala project:



# Congratulations! Spark is now ready to use on your computer!

## Optional

(9) Apache Maven

<http://maven.apache.org/>

---

*What if I don't want to install all these applications, but still want to work through the examples in the "Spark: The Definitive Guide" book?*

## **Running Spark in the Cloud**

Databricks Community Edition (free)

<https://databricks.com/try-databricks>

## **Example Code from "Spark: The Definitive Guide"**

<https://github.com/databricks/Spark-The-Definitive-Guide>