Pyspark Install

Introduction to Massive Data Analysis

2021.09

Step 1 - Python Install and Setting

Download Python files

- Available download page: https://www.python.org/downloads/

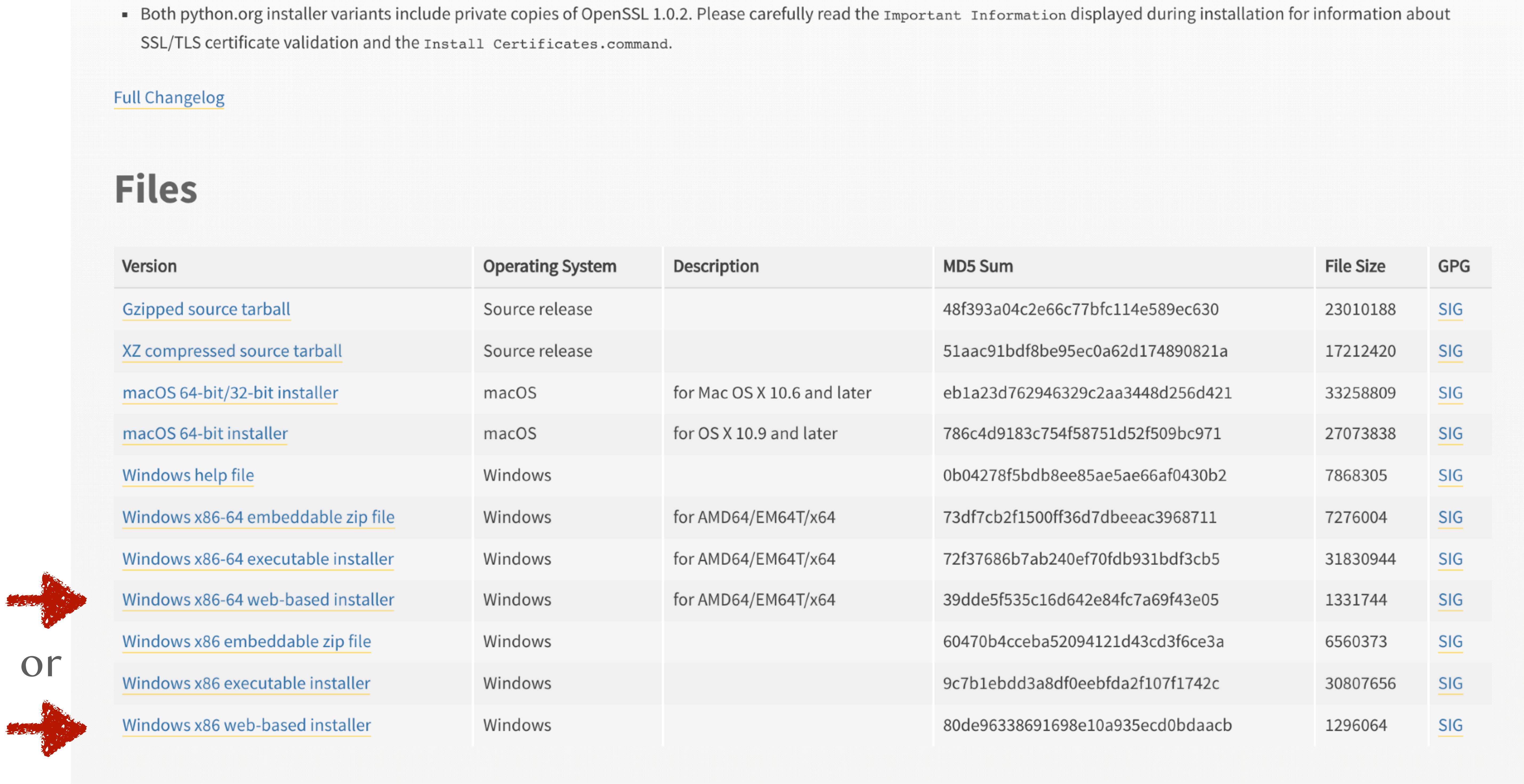
- Recommended version: python3.5.x, python 3.6.x

- It may cause some problems while downloading by Anaconda

Step 1 - Python Install

Download Python files

- Select **web-based installer**



- Both python.org installer variants include private copies of OpenSSL 1.0.2. Please carefully read the `Important Information` displayed during installation for information about SSL/TLS certificate validation and the `Install Certificates.command`.

Full Changelog

Files

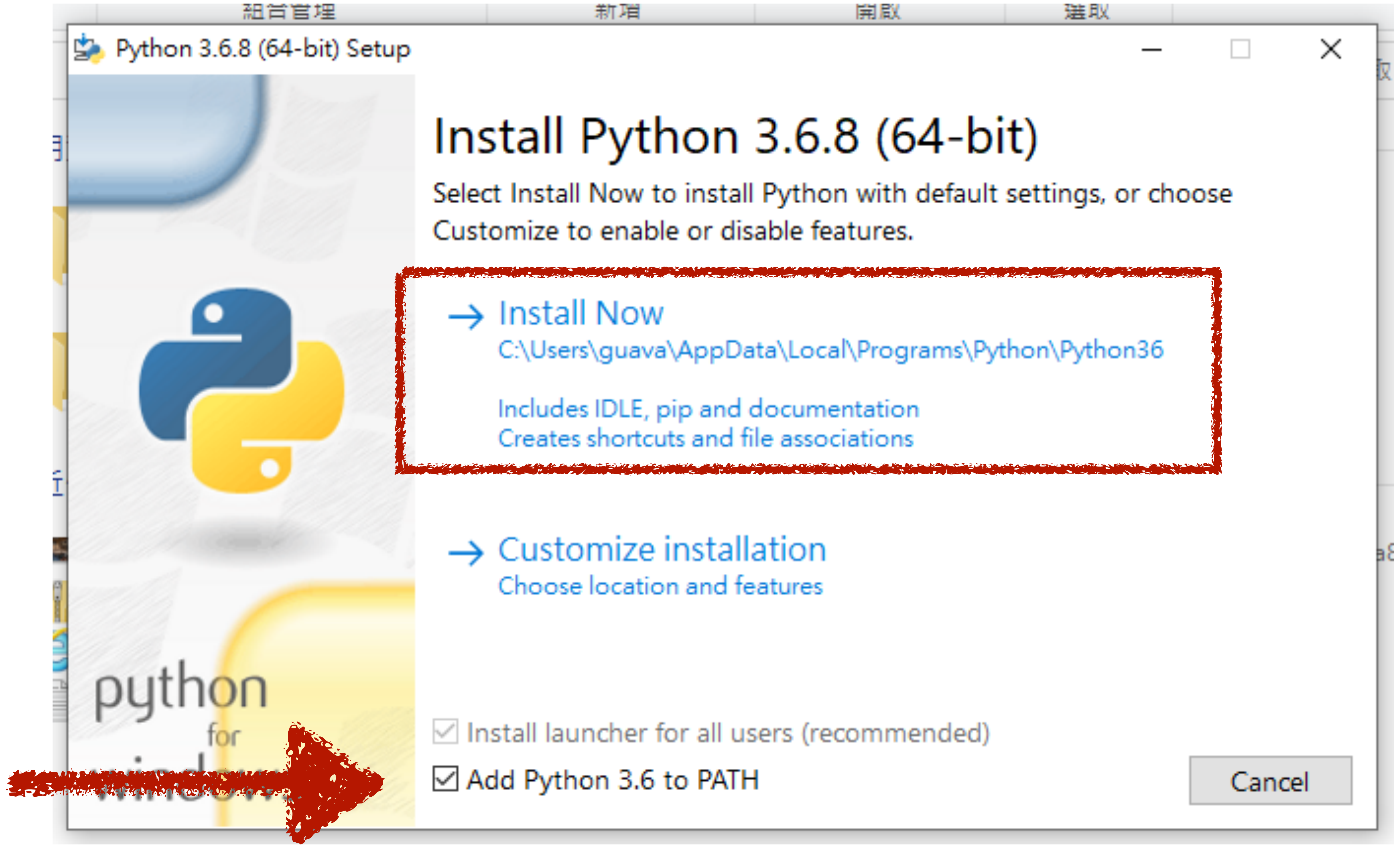
Version	Operating System	Description	MD5 Sum	File Size	GPG
Gzipped source tarball	Source release		48f393a04c2e66c77bfc114e589ec630	23010188	SIG
XZ compressed source tarball	Source release		51aac91bdf8be95ec0a62d174890821a	17212420	SIG
macOS 64-bit/32-bit installer	macOS	for Mac OS X 10.6 and later	eb1a23d762946329c2aa3448d256d421	33258809	SIG
macOS 64-bit installer	macOS	for OS X 10.9 and later	786c4d9183c754f58751d52f509bc971	27073838	SIG
Windows help file	Windows		0b04278f5bdb8ee85ae5ae66af0430b2	7868305	SIG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	73df7cb2f1500ff36d7dbeeac3968711	7276004	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	72f37686b7ab240ef70fdb931bdf3cb5	31830944	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	39dde5f535c16d642e84fc7a69f43e05	1331744	SIG
Windows x86 embeddable zip file	Windows		60470b4cceba52094121d43cd3f6ce3a	6560373	SIG
Windows x86 executable installer	Windows		9c7b1ebdd3a8df0eebfda2f107f1742c	30807656	SIG
Windows x86 web-based installer	Windows		80de96338691698e10a935ecd0bdaacb	1296064	SIG

or

Step 1 - Python Install

Installing and setting environment variables

- Unzip and run the downloaded file
- **Select "Add to PATH"**
- Install now



Python 3.6.8 (64-bit) Setup

Install Python 3.6.8 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

→ Install Now
C:\Users\guava\AppData\Local\Programs\Python\Python36

Includes IDLE, pip and documentation
Creates shortcuts and file associations

→ Customize installation
Choose location and features

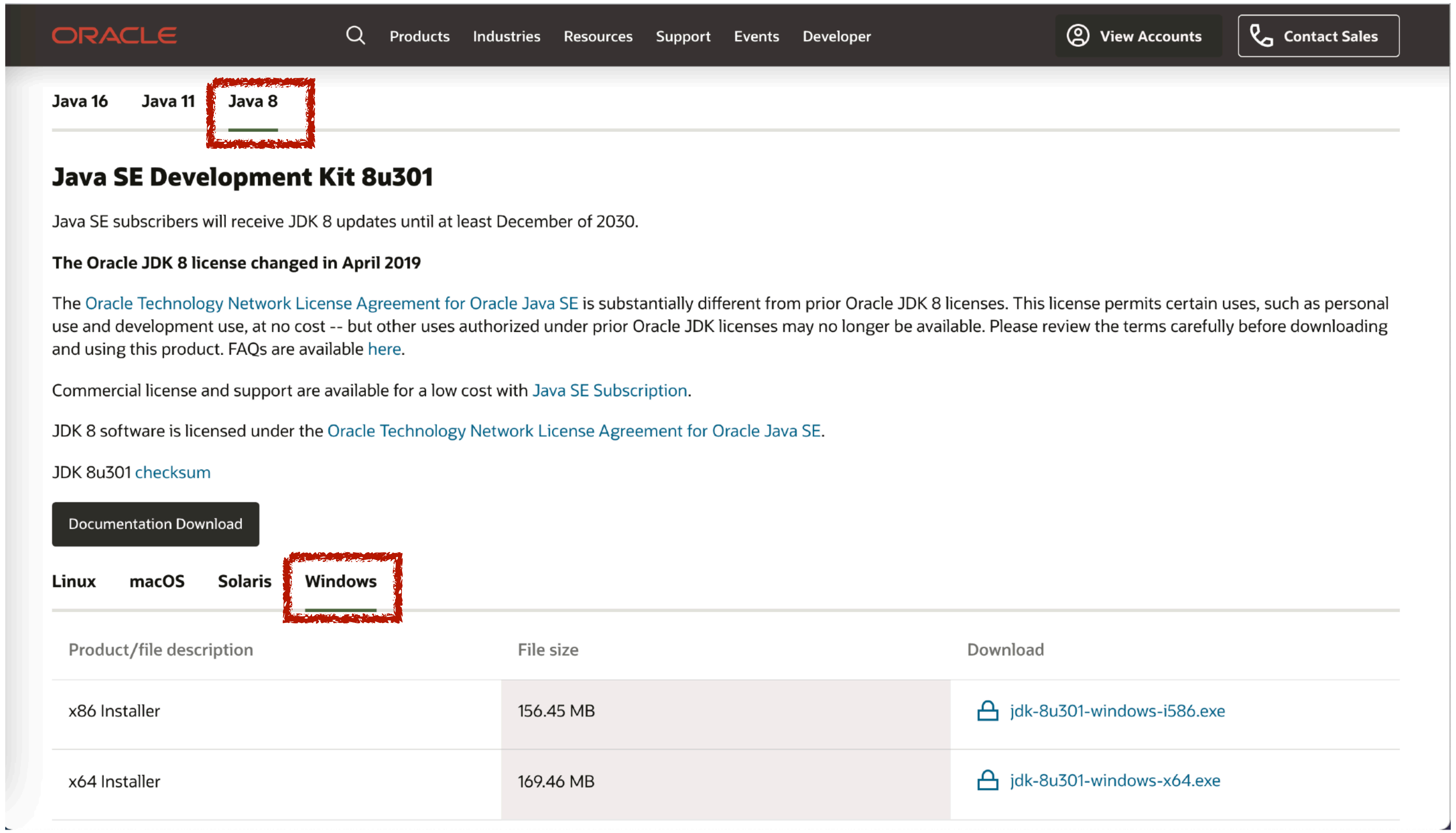
☑ Install launcher for all users (recommended)
☑ Add Python 3.6 to PATH

Step 2 - Java Install and Setting

Download Java jdk installer

- Available download page: https://www.oracle.com/java/technologies/downloads/#java8-windows

- Should select **Java8**



Java SE Development Kit 8u301

Java SE subscribers will receive JDK 8 updates until at least December of 2030.

The Oracle JDK 8 license changed in April 2019

The Oracle Technology Network License Agreement for Oracle Java SE is substantially different from prior Oracle JDK 8 licenses. This 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. FAQs are available here.

Commercial license and support are available for a low cost with Java SE Subscription.

JDK 8 software is licensed under the Oracle Technology Network License Agreement for Oracle Java SE.

JDK 8u301 checksum

Product/file description	File size	Download
x86 Installer	156.45 MB	jdk-8u301-windows-i586.exe
x64 Installer	169.46 MB	jdk-8u301-windows-x64.exe

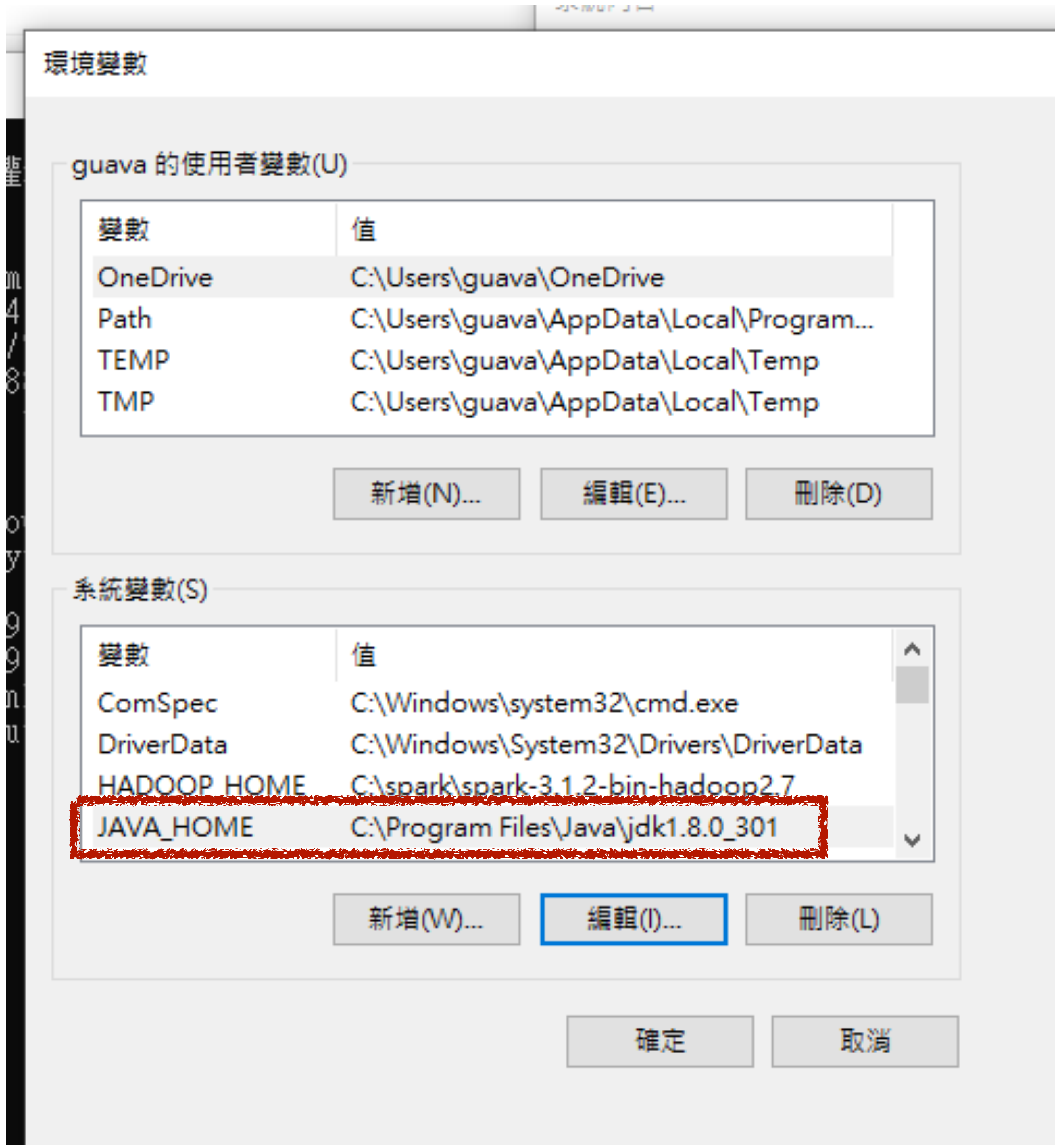
Step 2 - Java Install and Setting

Setting environment variables

- At Environment Variables > System Variables

- Add JAVA_HOME

- Fill in the **actual address** of your jdk folder (not the one in picture)



環境變數

guava 的使用者變數(U)

變數	值
OneDrive	C:\Users\guava\OneDrive
Path	C:\Users\guava\AppData\Local\Program...
TEMP	C:\Users\guava\AppData\Local\Temp
TMP	C:\Users\guava\AppData\Local\Temp

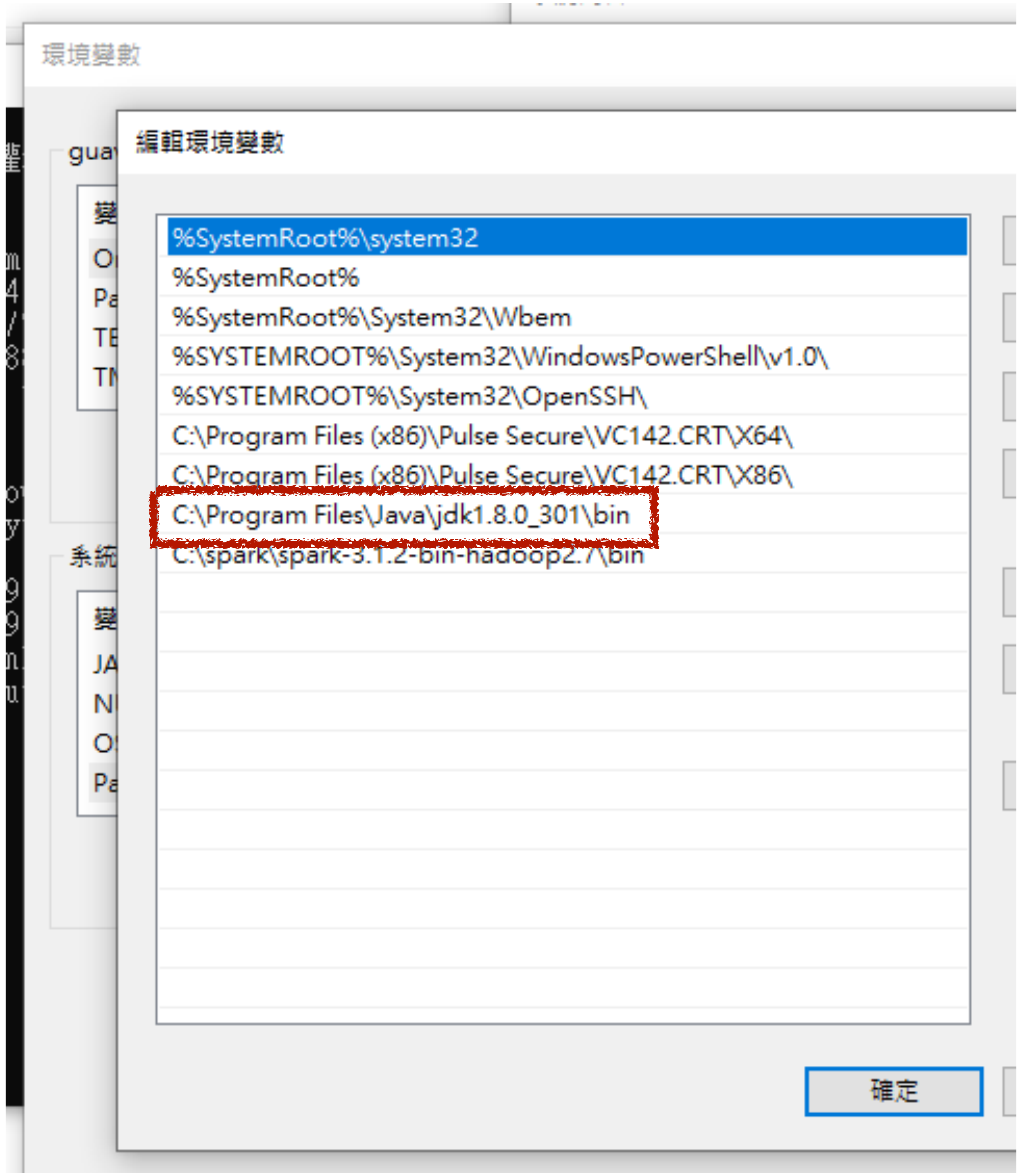
系統變數(S)

變數	值
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
HADOOP_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_301

Step 2 - Java Install and Setting

Setting environment variables

- At Environment Variables > System Variables
- Modify and add **PATH**
- Fill in the **actual address** of the **bin** file in your jdk folder

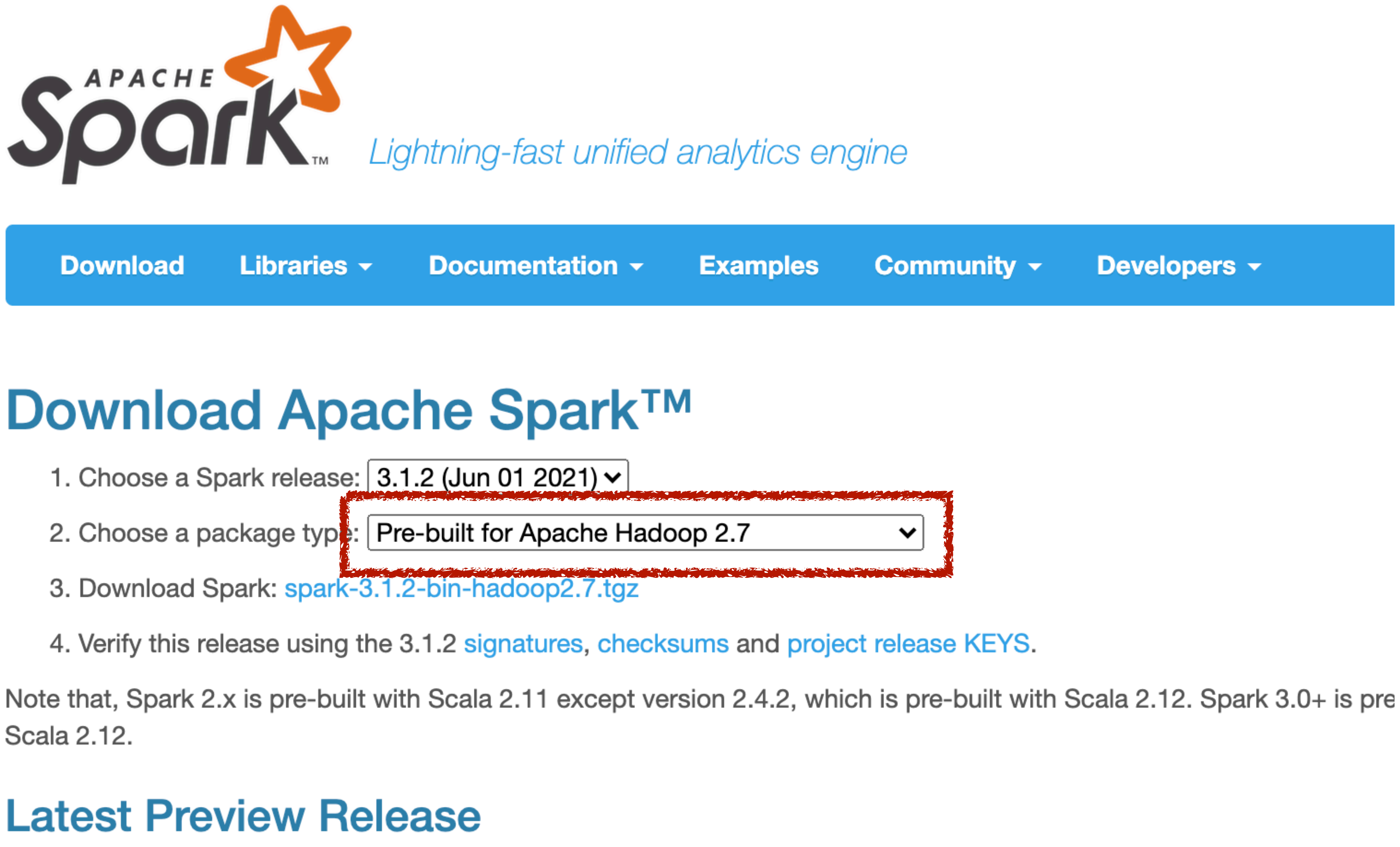


```
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files (x86)\Pulse Secure\VC142.CRT\X64\
C:\Program Files (x86)\Pulse Secure\VC142.CRT\X86\
C:\Program Files\Java\jdk1.8.0_301\bin
C:\spark\spark-3.1.2-bin-hadoop2.7\bin
```

Step 3 - Spark Install and Setting

Download Spark files

- Available download page: http://spark.apache.org/downloads.html
- Should select **Hadoop 2.7**



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
4. Verify this release using the 3.1.2 signatures, checksums and project release KEYS.

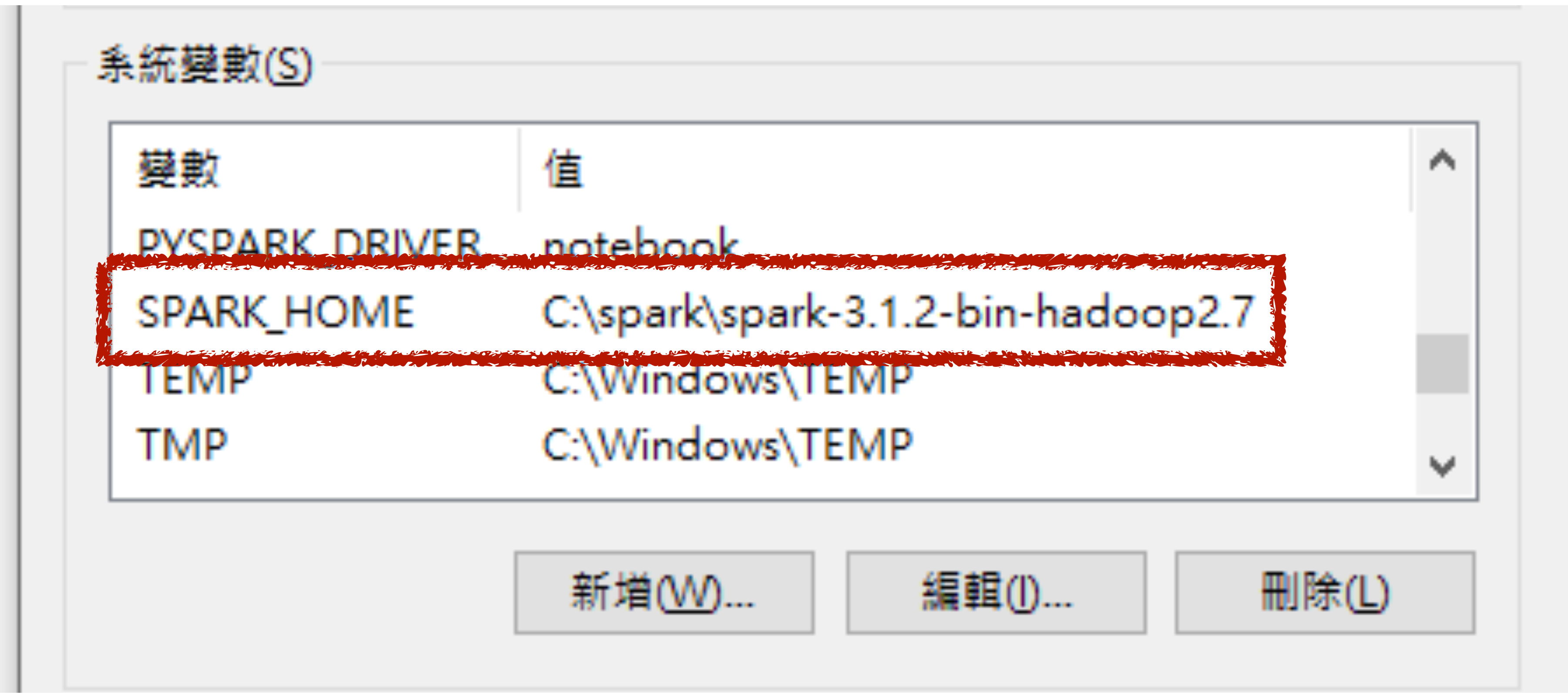
Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre
Scala 2.12.

Latest Preview Release

Step 3 - Spark Install and Setting

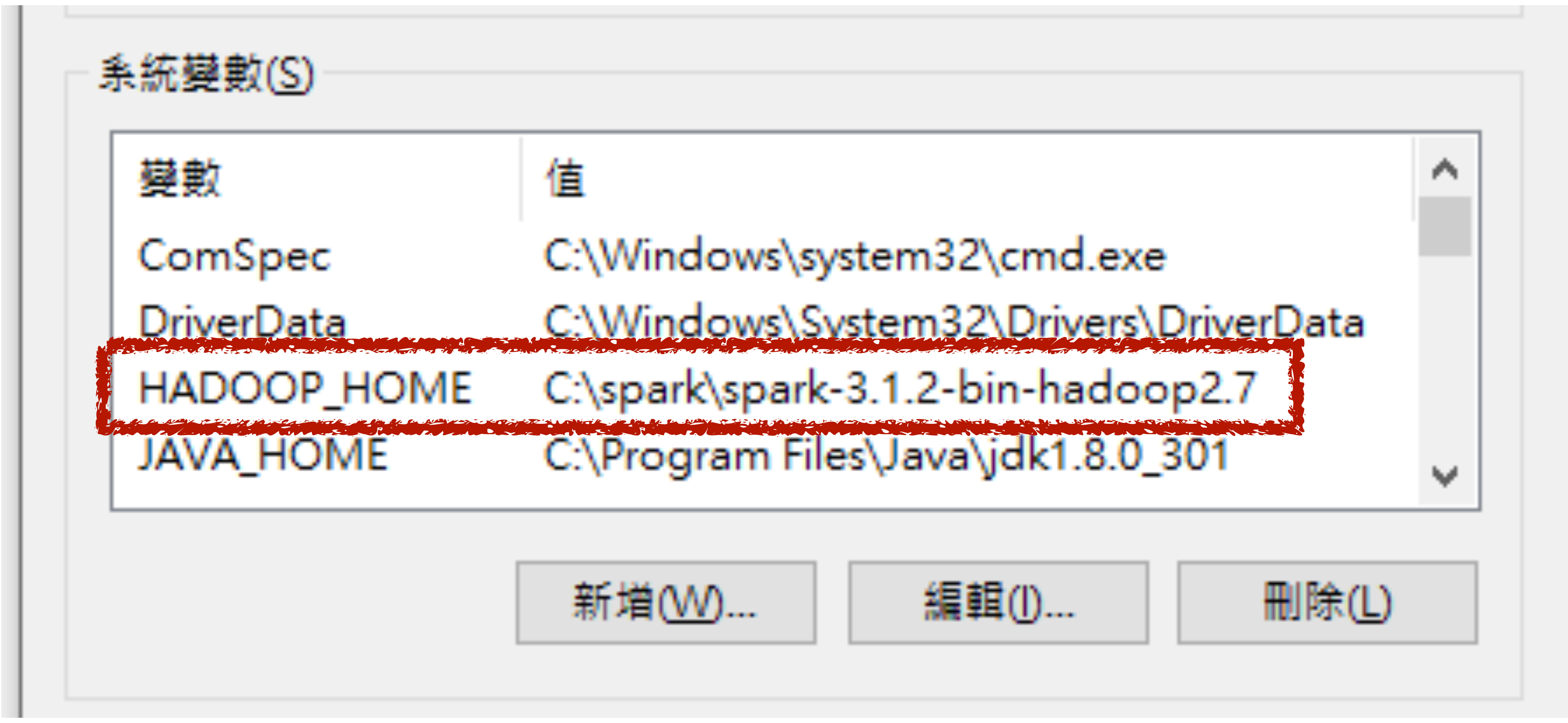
Setting environment variables

- At Environment Variables > System Variables
- Add **SPARK_HOME** and **HADOOP_HOME**
- Fill in the **actual address** of your spark folder (not the one in picture)



系統變數(S)

變數	值
PYSPARK_DRIVER	notebook
SPARK_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
TEMP	C:\Windows\TEMP
TMP	C:\Windows\TEMP



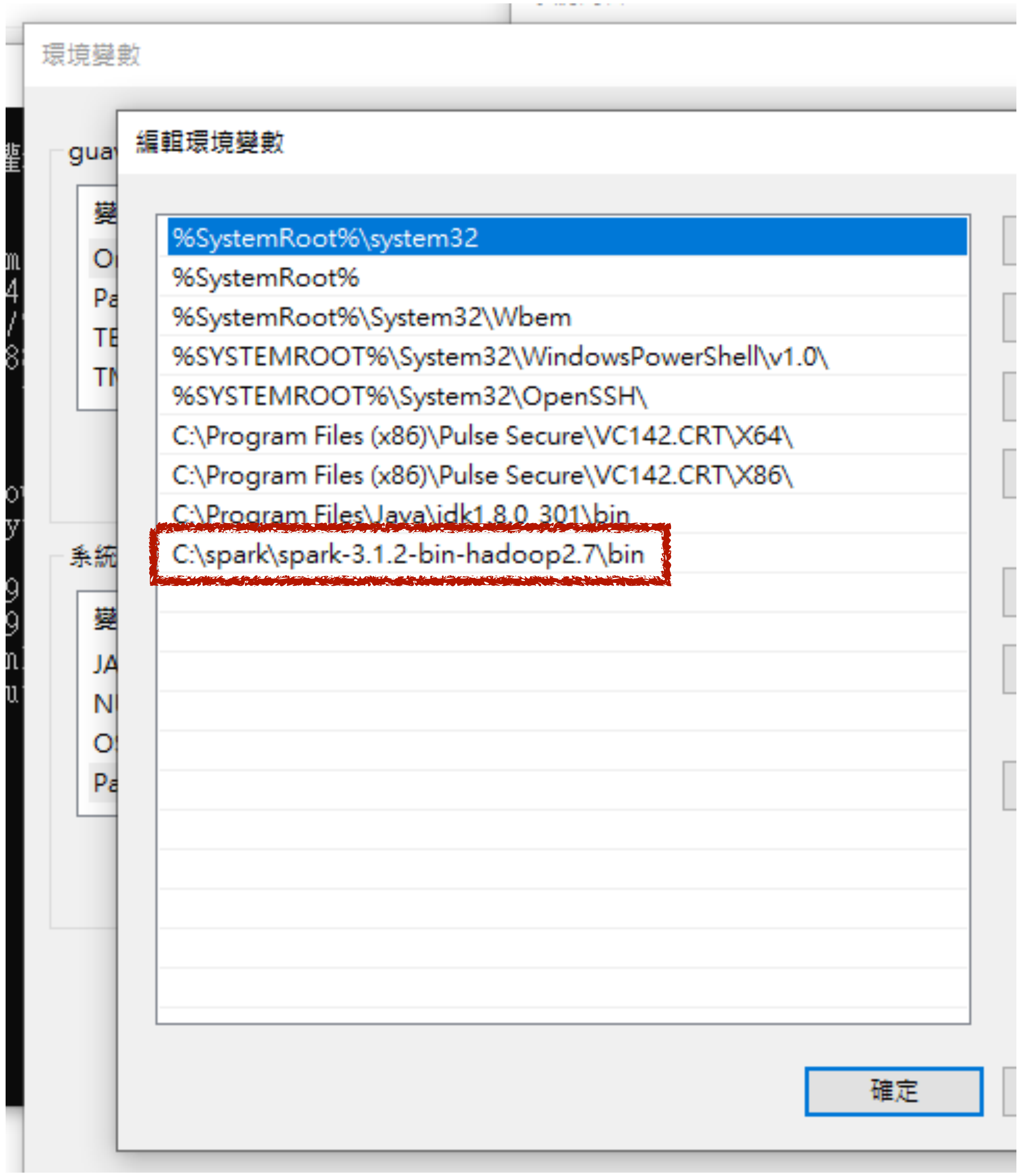
系統變數(S)

變數	值
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
HADOOP_HOME	C:\spark\spark-3.1.2-bin-hadoop2.7
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_301

Step 3 - Spark Install and Setting

Setting environment variables

- At Environment Variables > System Variables
- Modify and add **PATH**
- Fill in the **actual address** of the **bin** file in your spark folder



環境變數

編輯環境變數

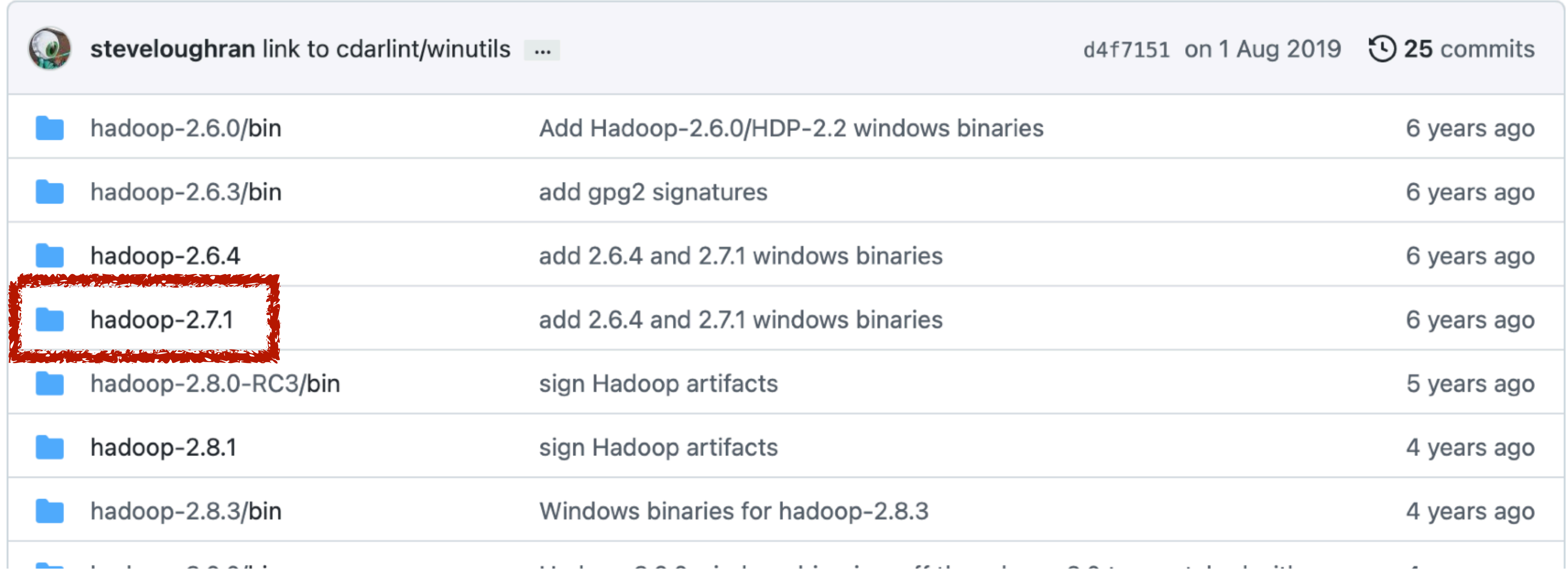
```
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files (x86)\Pulse Secure\VC142.CRT\X64\
C:\Program Files (x86)\Pulse Secure\VC142.CRT\X86\
C:\Program Files\Java\jdk1.8.0_301\bin
C:\spark\spark-3.1.2-bin-hadoop2.7\bin
```

Step 4 - Hadoop winutils

Download winutils.exe

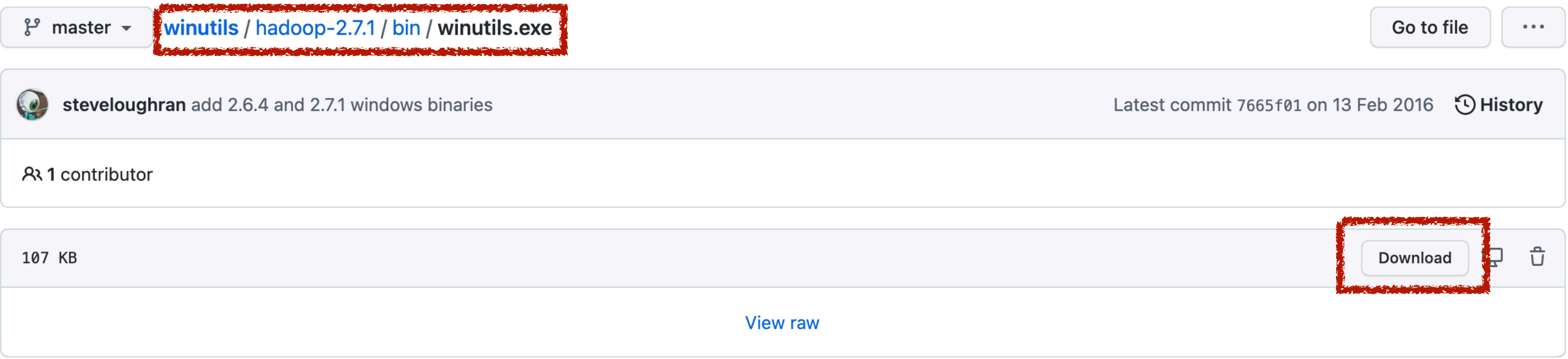
- Available download page: https://github.com/steveloughran/winutils

- Download **winutils.exe** in hadoop.2.7.1



steveloughran link to cdarlint/winutils ... d4f7151 on 1 Aug 2019 **25** commits

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	6 years ago
hadoop-2.7.1	add 2.6.4 and 2.7.1 windows binaries	6 years ago
hadoop-2.8.0-RC3/bin	sign Hadoop artifacts	5 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



master ▾ winutils / hadoop-2.7.1 / bin / **winutils.exe**

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

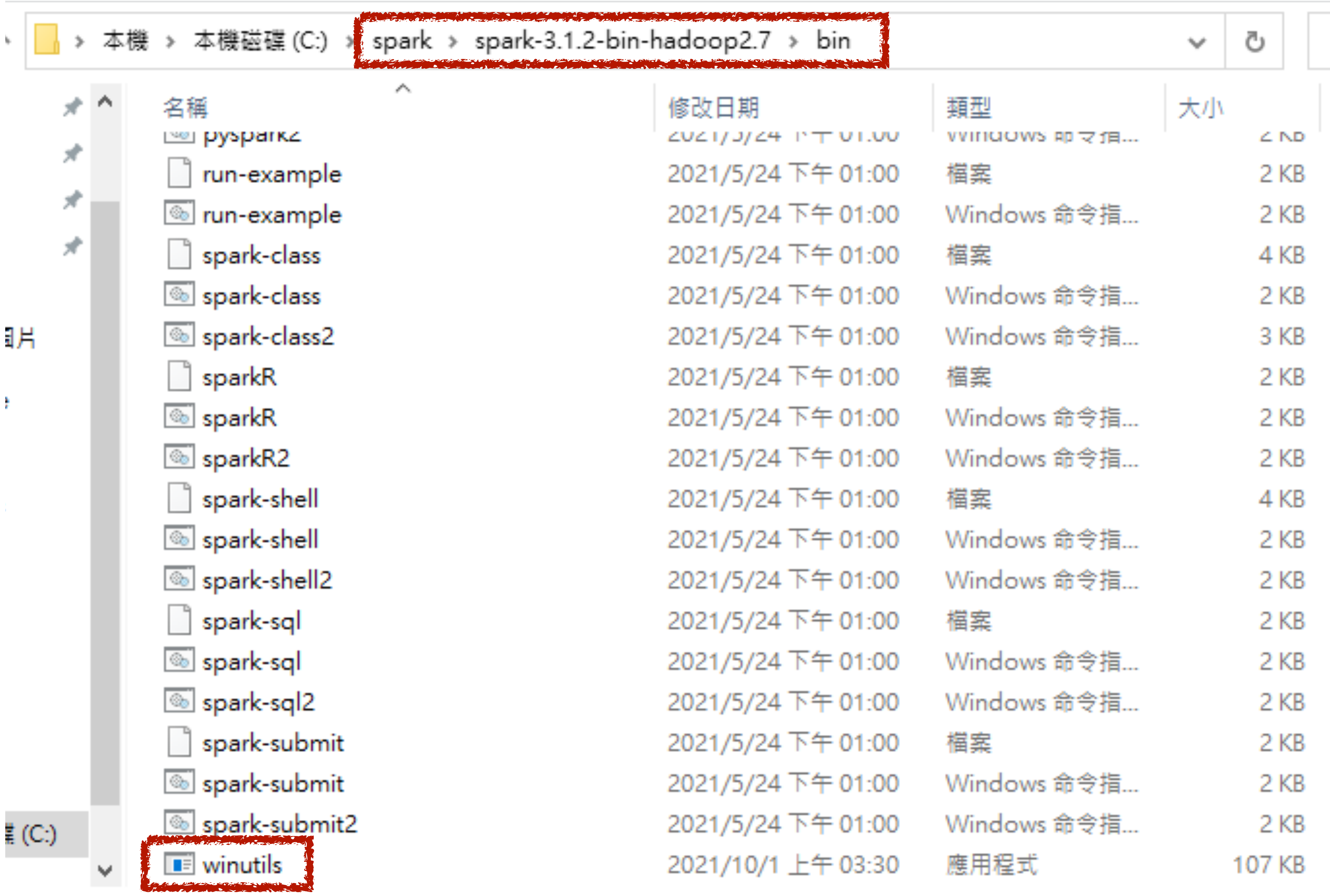
1 contributor

107 KB

Step 4 - Hadoop winutils

Set winutils.exe

- Put it into the bin folder in Spark



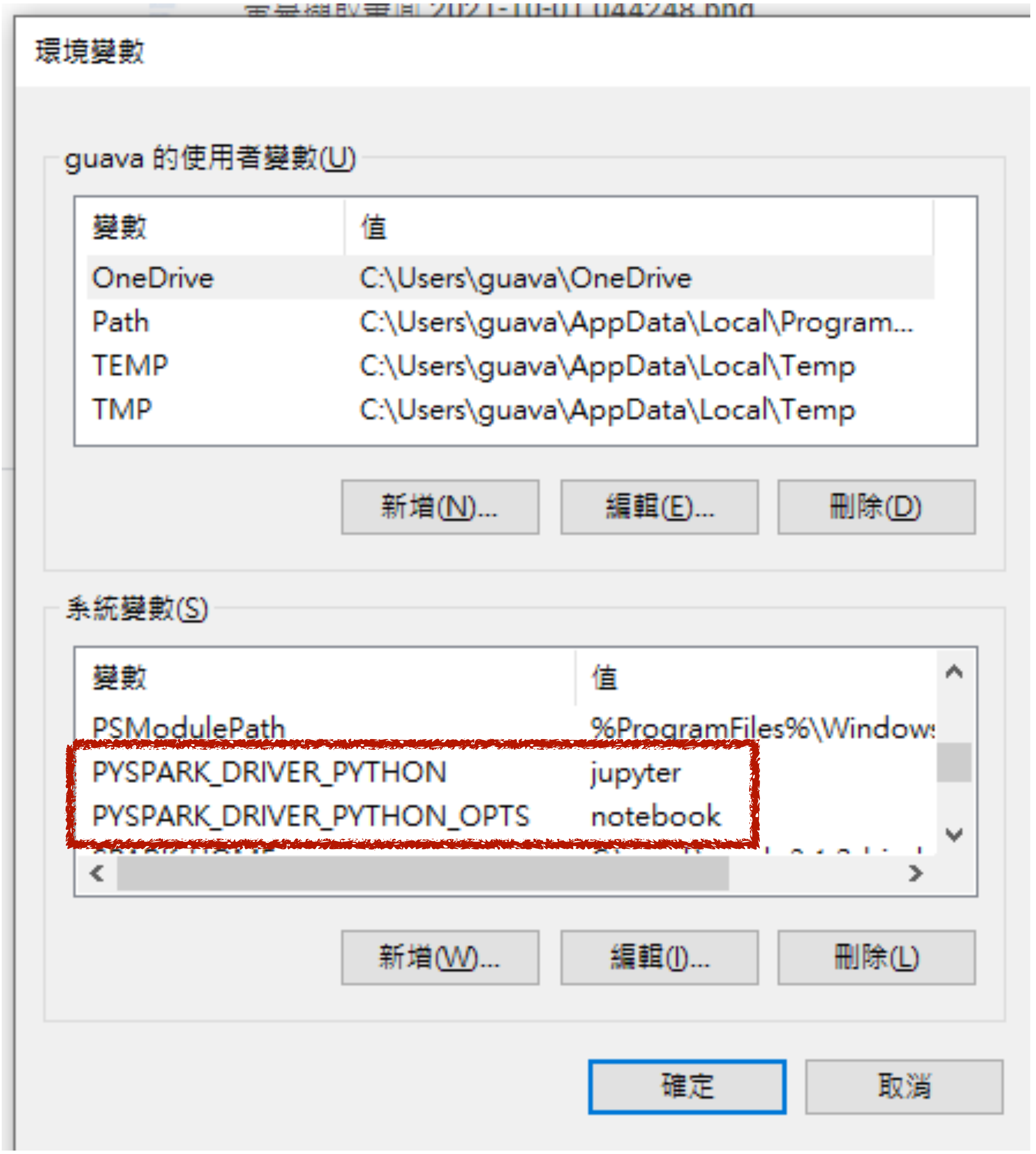
本機 › 本機磁碟 (C:) › spark › spark-3.1.2-bin-hadoop2.7 › bin

名稱	修改日期	類型	大小
pyspark2	2021/5/24 下午 01:00	Windows 命令指...	2 KB
run-example	2021/5/24 下午 01:00	檔案	2 KB
run-example	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-class	2021/5/24 下午 01:00	檔案	4 KB
spark-class	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-class2	2021/5/24 下午 01:00	Windows 命令指...	3 KB
sparkR	2021/5/24 下午 01:00	檔案	2 KB
sparkR	2021/5/24 下午 01:00	Windows 命令指...	2 KB
sparkR2	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-shell	2021/5/24 下午 01:00	檔案	4 KB
spark-shell	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-shell2	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-sql	2021/5/24 下午 01:00	檔案	2 KB
spark-sql	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-sql2	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-submit	2021/5/24 下午 01:00	檔案	2 KB
spark-submit	2021/5/24 下午 01:00	Windows 命令指...	2 KB
spark-submit2	2021/5/24 下午 01:00	Windows 命令指...	2 KB
winutils	2021/10/1 上午 03:30	應用程式	107 KB

Step 5 - jupyter notebook

Install Jupyter and set environment variables

- Type `pip install jupyter` in cmd
- Add **PYSPARK_DRIVER_PYTHON**: **jupyter**
- Add **PYSPARK_DRIVER_PYTHON_OPTS**: **notebook**



環境變數

guava 的使用者變數(U)

變數	值
OneDrive	C:\Users\guava\OneDrive
Path	C:\Users\guava\AppData\Local\Program...
TEMP	C:\Users\guava\AppData\Local\Temp
TMP	C:\Users\guava\AppData\Local\Temp

系統變數(S)

變數	值
PSModulePath	%ProgramFiles%\Window:
PYSPARK_DRIVER_PYTHON	jupyter
PYSPARK_DRIVER_PYTHON_OPTS	notebook

Step 6 - Testing

- Type `pyspark` in cmd
- The jupyter notebook page will open in browser
-

Some do's and don'ts

- Avoid setting **Chinese words**, **space** or **special marks** in environment variables

- Make sure all addresses exist in your own device (Copy addresses on picture may not work)

- **Restart cmd** after updating environment variables