

Chang Hsin Lee
Committing my thoughts to words.

Posts YouTube 中文 About

How to Install and Run PySpark in Jupyter Notebook on Windows

When I write PySpark code, I use Jupyter notebook to test my code before submitting a job on the cluster. In this post, I will show you how to install and run PySpark locally in Jupyter Notebook on Windows. I've tested this guide on a dozen Windows 7 and 10 PCs in different languages.

A. Items needed


1. Spark distribution from spark.apache.org

Download Apache Spark™



1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.2.1-bin-hadoop2.7.tgz](#)

2. Python and Jupyter Notebook. You can get both by installing the Python 3.x version of [Anaconda distribution](#).
3. `winutils.exe` — a Hadoop binary for Windows — from Steve Loughran's [GitHub repo](#). Go to the corresponding Hadoop version in the Spark distribution and find `winutils.exe` under `/bin`. For example,
<https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>

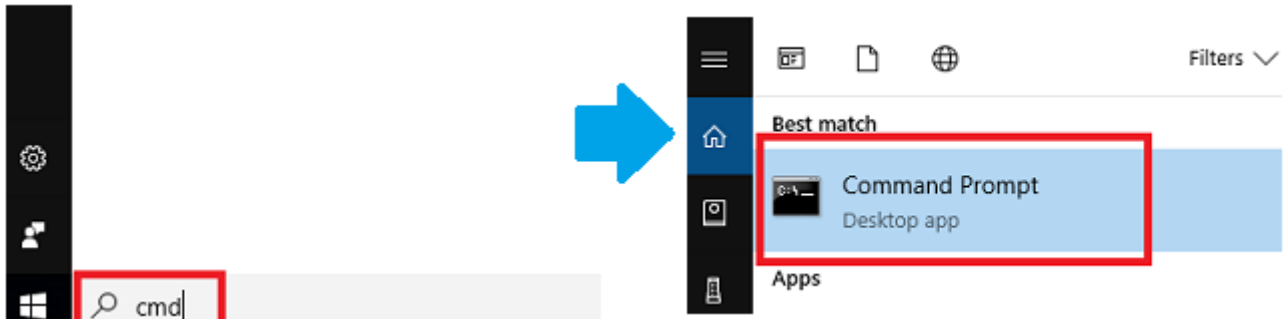
Branch: master [winutils / hadoop-2.7.1 / bin / winutils.exe](#) Find file Copy path

 **steveloughran** add 2.6.4 and 2.7.1 windows binaries 7665f01 on 12 Feb 2016

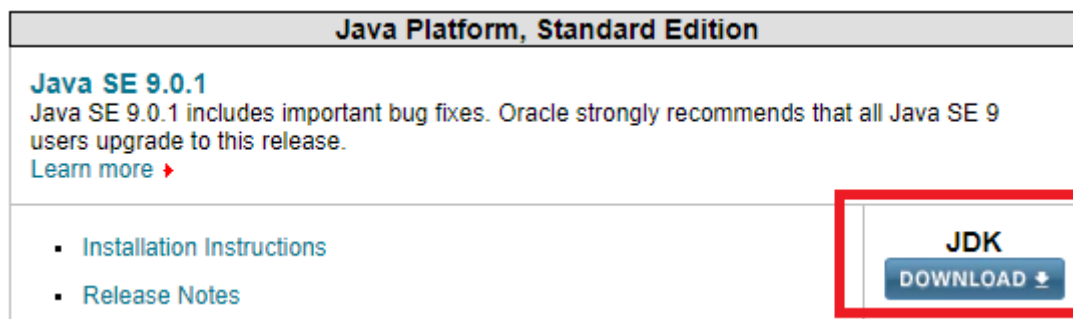
1 contributor

107 KB Download History  

4. The findspark Python module, which can be installed by running `python -m pip install findspark` either in Windows command prompt or Git bash if Python is installed in item 2. You can find command prompt by searching cmd in the search box.



5. If you don't have Java or your Java version is 7.x or less, download and install Java from [Oracle](#). I recommend getting the latest JDK (current version 9.0.1).



6. If you don't know how to unpack a .tgz file on Windows, you can download and install 7-zip on Windows to unpack the .tgz file from Spark distribution in item 1 by right-clicking on the file icon and select 7-zip > Extract Here.

Download 7-Zip 16.04 (2016-10-04) 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
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

B. Installing PySpark

After getting all the items in section A, let's set up PySpark.

1. Unpack the .tgz file. For example, I unpacked with 7zip from step A6 and put mine under D:\spark\spark-2.2.1-bin-hadoop2.7

spark-2.2.1-bin-hadoop2.7	12/23/2017 11:00 ...	File folder	
spark-2.2.1-bin-hadoop2.7.tar	11/24/2017 6:31 PM	WinRAR archive	223,100 KB
spark-2.2.1-bin-hadoop2.7.tgz	12/23/2017 10:58 ...	WinRAR archive	196,225 KB

2. Move the winutils.exe downloaded from step A3 to the \bin folder of Spark distribution. For example, D:\spark\spark-2.2.1-bin-hadoop2.7\bin\winutils.exe
3. Add environment variables: the environment variables let Windows find where the files are when we start the PySpark kernel. You can find the environment variable settings by putting “environ...” in the search box.

The variables to add are, in my example,

Name	Value
SPARK_HOME	D:\spark\spark-2.2.1-bin-hadoop2.7
HADOOP_HOME	D:\spark\spark-2.2.1-bin-hadoop2.7
PYSPARK_DRIVER_PYTHON	jupyter
PYSPARK_DRIVER_PYTHON_OPTS	notebook

Edit User Variable

Variable name: HADOOP_HOME

Variable value: D:\spark\spark-2.2.1-bin-hadoop2.7

Browse Directory... Browse File... OK Cancel

4. In the same environment variable settings window, look for the Path or PATH variable, click edit and add D:\spark\spark-2.2.1-bin-hadoop2.7\bin to it. In Windows 7 you need to separate the values in Path with a semicolon ; between the values.
5. (Optional, if see Java related error in step C) Find the installed Java JDK folder from step A5, for example, D:\Program Files\Java\jdk1.8.0_121, and add the following environment variable

Name	Value
JAVA_HOME	D:\Progra~1\Java\jdk1.8.0_121

If JDK is installed under \Program Files (x86), then replace the Progra~1 part by Progra~2 instead. In my experience, this error only occurs in Windows 7, and I think it's because Spark couldn't parse the space in the folder name.

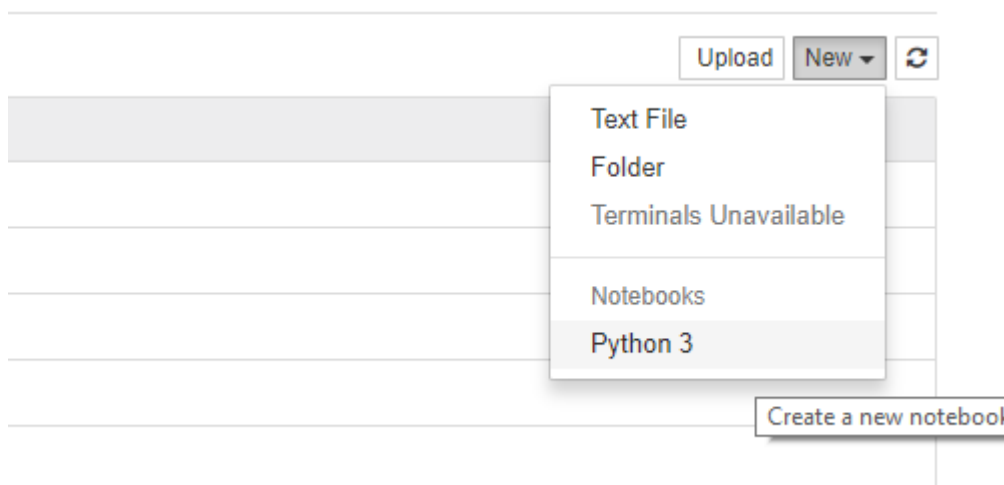
Edit (1/23/19): You might also find Gerard's comment helpful:

<http://disq.us/p/1z5qou4>

C. Running PySpark in Jupyter Notebook

To run Jupyter notebook, open Windows command prompt or Git Bash and run `jupyter notebook`. If you use Anaconda Navigator to open Jupyter Notebook instead, you might see a Java gateway process exited before sending the driver its port number error from PySpark in step C. Fall back to Windows cmd if it happens.

Once inside Jupyter notebook, open a Python 3 notebook



In the notebook, run the following 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()
```

When you press run, it might trigger a Windows firewall pop-up. I pressed cancel on the pop-up as blocking the connection doesn't affect PySpark.

```

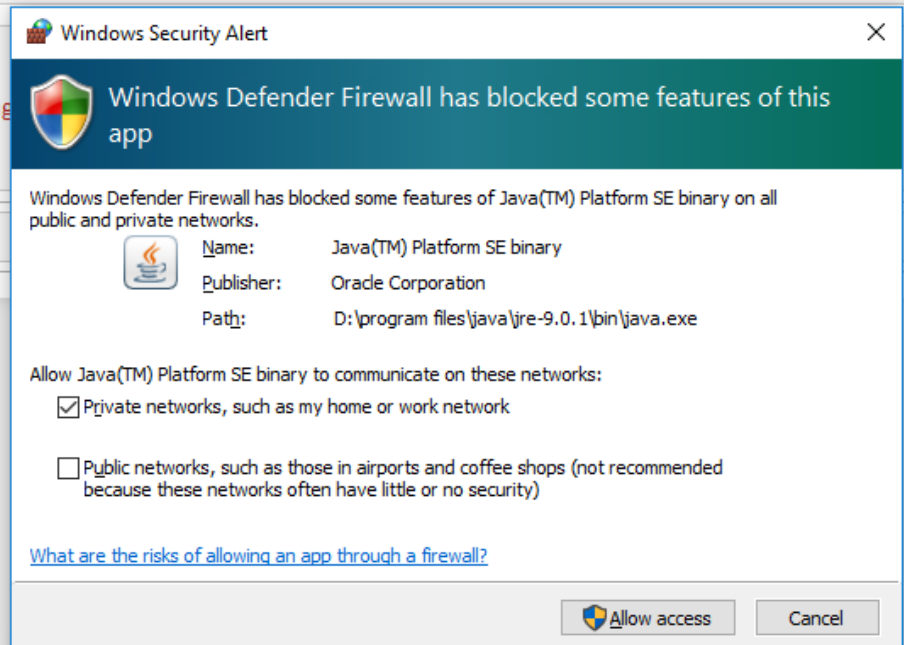
In [1]: import findspark
        findspark.init()

In [2]: import pyspark
        from pyspark.sql import SparkSession

In [4]: spark = (SparkSession
              .builder
              .appName('chang')
              .getOrCreate())

In [ ]: df = spark.sql('')

```



If you see the following output, then you have installed PySpark on your Windows system!

```

In [4]: # test spark.sql
        df = spark.sql('select 'spark' as hello ')
        df.show()

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

```

Misc

- Update (10/30/19): Tip from Nathaniel Anderson in comments: you might want to install Java 8 and point JAVA_HOME to it if you are seeing this error: "Py4JJavaError: An error occurred..." [StackOverflow Answer](#)

Please leave a comment in the comments section or tweet me at [@ChangLeeTW](#) if you have any question.

Other PySpark posts from me (last updated 3/4/2018) —

- [How to Turn Python Functions into PySpark Functions \(UDF\)](#)
- [PySpark Dataframe Basics](#)

Written on *December 30, 2017*

Share via

 facebook

 twitter

 linkedin

82 Comments

Chang Hsin Lee

1


Login ▾

Recommend 9

Tweet

Share

Sort by Best ▾



Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS 

Name

Load more comments

ALSO ON CHANG HSIN LEE