

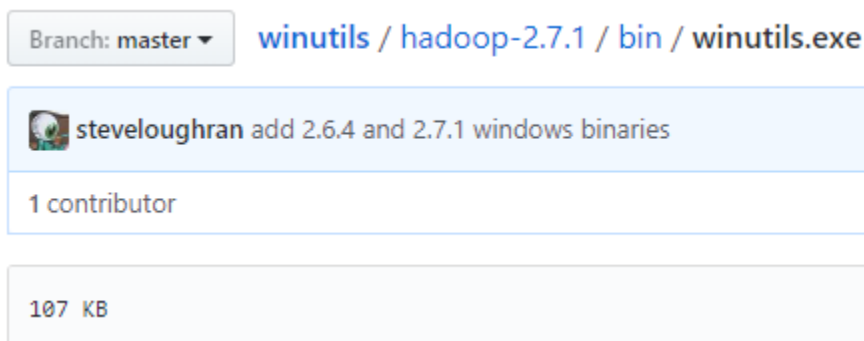
## A. Items needed

1. Spark distribution from [spark.apache.org](http://spark.apache.org)

### Download Apache Spark™

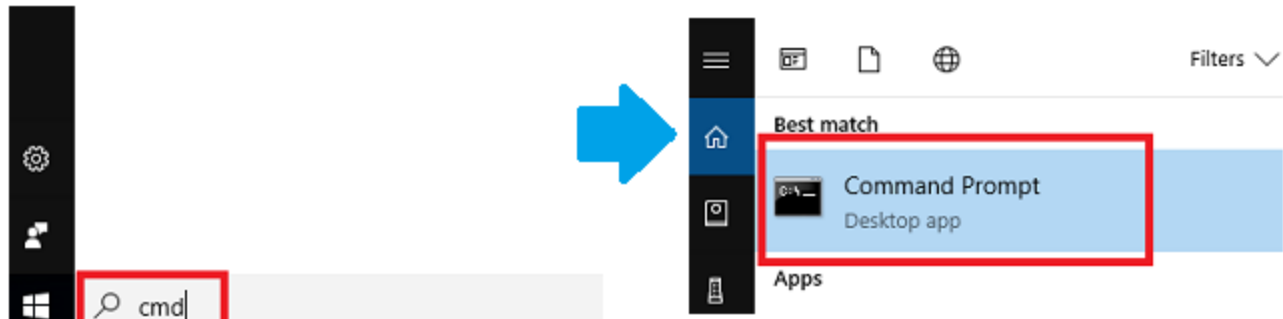
1. Choose a Spark release: 2.2.1 (Dec 01 2017) ▼
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later ▼
3. Download Spark: [spark-2.2.1-bin-hadoop2.7.tgz](#)

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>.



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
<a href="#">Download</a>	.exe	32-bit x86	7-Zip for 32-bit Windows
<a href="#">Download</a>	.exe	64-bit x64	7-Zip for 64-bit Windows x64 (Intel 64 or AMD64)
<a href="#">Download</a>	.7z	x86 / x64	7-Zip Extra: standalone console version, 7z DLL, Plugin
<a href="#">Download</a>	.7z	Any	7-Zip Source code
<a href="#">Download</a>	.7z	Any / x86 / x64	LZMA SDK: (C, C++, C#, Java)
<a href="#">Download</a>	.msi	32-bit x86	(alternative MSI installer) 7-Zip for 32-bit Windows
<a href="#">Download</a>	.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.

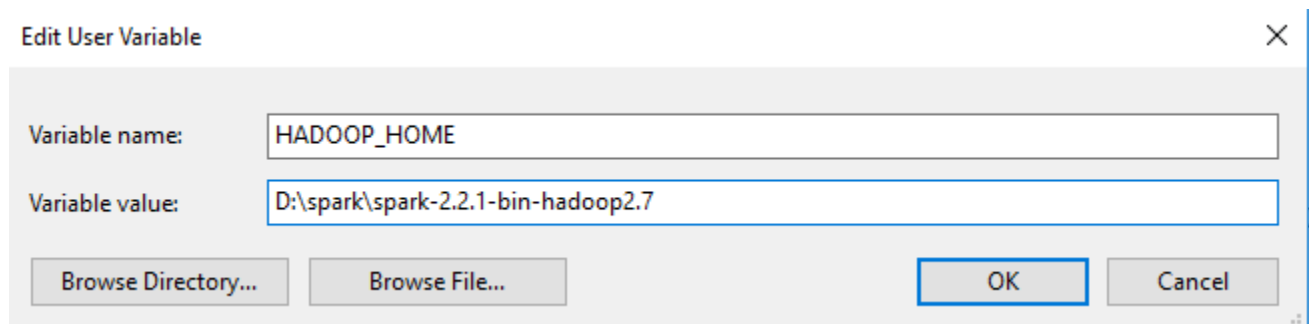
1. Unpack the .tgz file. For example, I unpacked with 7zip from step A6 and put mine under D:\spark\spark-2.2.1-bin-hadoop2.7

 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

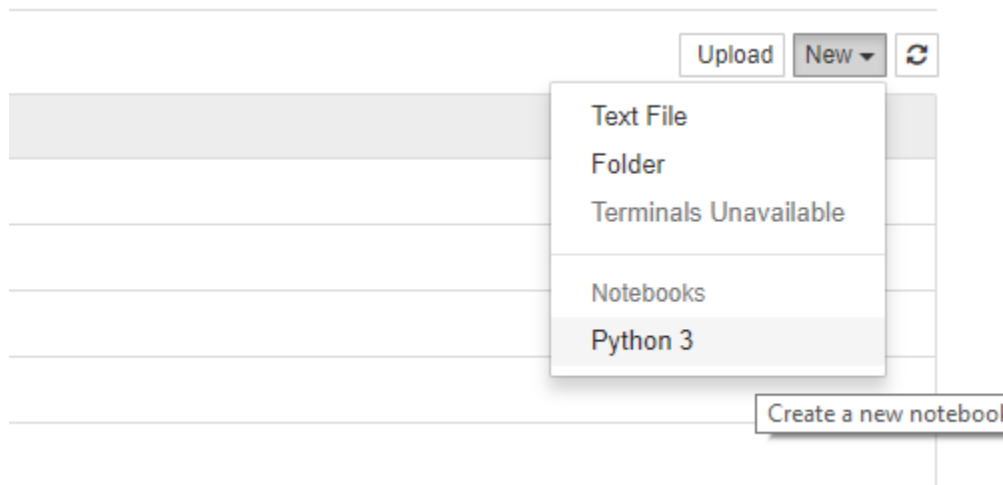


4. In the same environment variable settings window, look for the Path or PATH variable, click edit and add D:\spark\spark-2.2.1-bin-hadoop2.7\bin to it. In Windows 7 you need to separate the values in Path with a semicolon ; between the values.
5. If JDK is installed under \Program Files (x86), then 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.

## 6. 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()
```

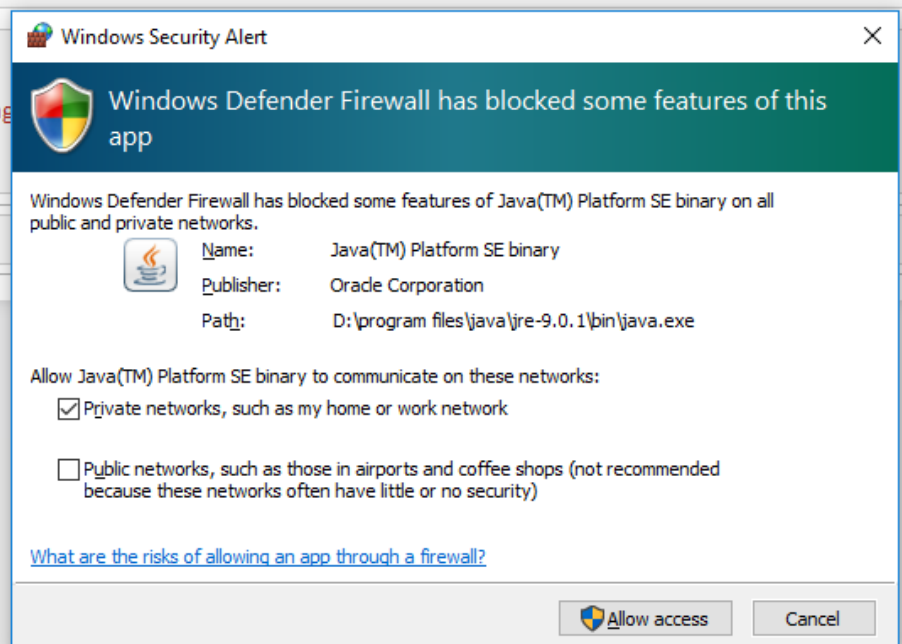
When you press run, it might trigger a Windows firewall pop-up. I pressed cancel on the pop-up as blocking the connection doesn't affect PySpark.

```
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|
+-----+
```