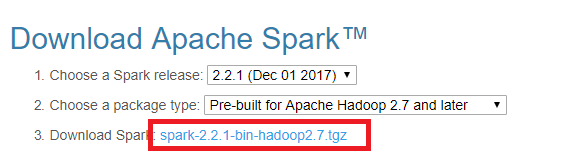
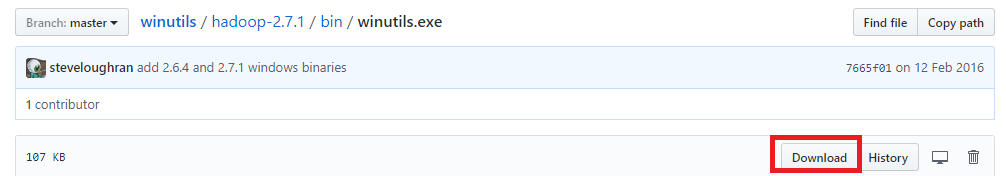
### A. Items needed

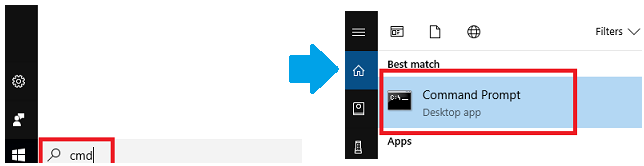
1. Spark distribution from [spark.apache.org](https://spark.apache.org/downloads.html)



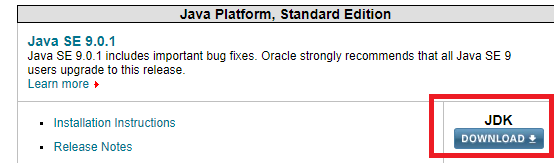
1. Python and Jupyter Notebook. You can get both by installing the Python 3.x version of [Anaconda distribution](https://www.anaconda.com/download/).
2. winutils.exe — a Hadoop binary for Windows — from Steve Loughran’s [GitHub repo](https://github.com/steveloughran/winutils/). Go to the corresponding 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> .



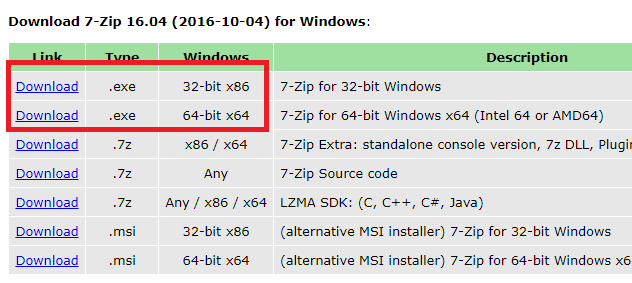
1. The findspark Python module, which can be installed by running python -m pip install findspark either in Windows command prompt or Git bash if Python is installed in item 2. You can find command prompt by searching cmd in the search box.



1. If you don’t have Java or your Java version is 7.x or less, download and install Java from [Oracle](http://www.oracle.com/technetwork/java/javase/downloads/index.html). I recommend getting the latest JDK (current version 9.0.1).



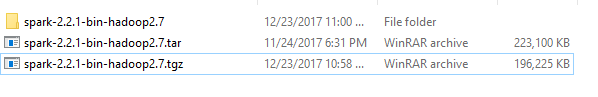
1. If you don’t know how to unpack a .tgz file on Windows, you can download and install [7-zip](http://www.7-zip.org/download.html) on Windows to unpack the .tgz file from Spark distribution in item 1 by right-clicking on the file icon and select 7-zip > Extract Here.



## B. Installing PySpark

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

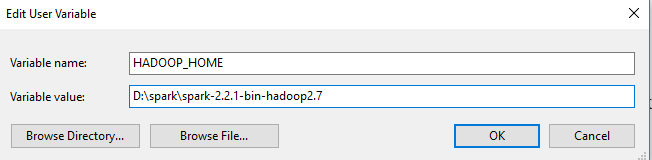
1. Unpack the .tgz file. For example, I unpacked with 7zip from step A6 and put mine under D:\spark\spark-2.2.1-bin-hadoop2.7



1. 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
2. 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 |

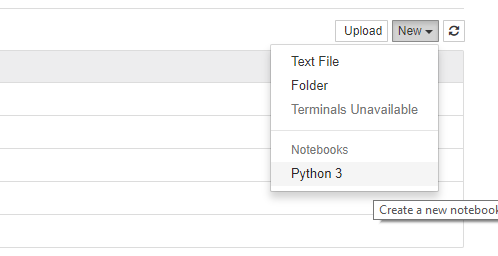


1. In the same environment variable settings window, look for the Path or PATH variable, click edit and add D:\spark\spark-2.2.1-bin-hadoop2.7\bin to it. In Windows 7 you need to separate the values in Path with a semicolon ; between the values.
2. If JDK is installed under \Program Files (x86), then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

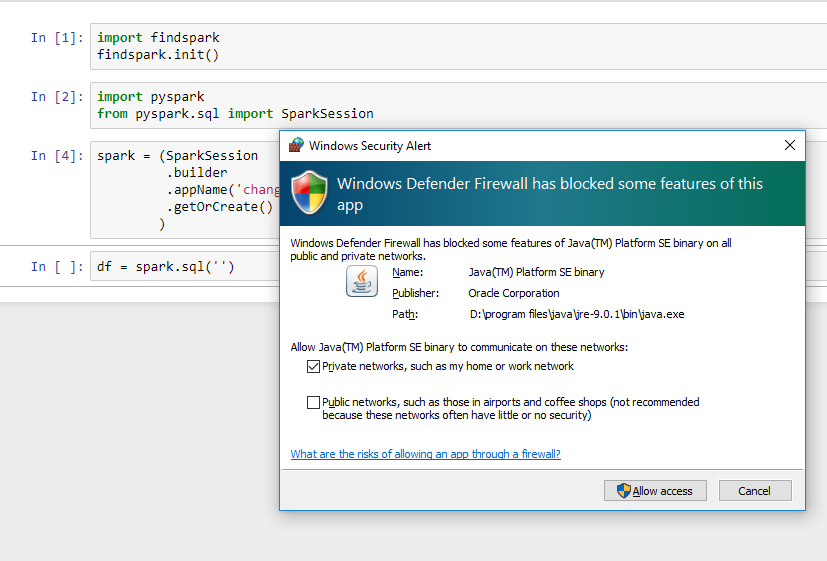
from pyspark.sql import SparkSession

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.



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

