



Michael Galarnyk

Data Scientist at Daymon Interactions as well as Data Science Masters Student at UC San Diego.

Apr 2, 2017 · 2 min read

Install Spark on Windows (PySpark)

This embedded content is from a site that does not comply with the Do Not Track (DNT) setting now enabled on your browser.

Please note, if you click through and view it anyway, you may be tracked by the website hosting the embed.

Install PySpark on Windows

The video above walks through installing spark on windows following the set of instructions below. You can either leave a comment here or leave me a comment on [youtube](#) (please subscribe if you can) if you have any questions!

Prerequisites: Anaconda and GOW. If you already have anaconda and GOW installed, skip to step 5.

1. Download and install Gnu on windows (GOW) from the following [link](#). Basically, GOW allows you to use linux commands on windows. In this install, we will need curl, gzip, tar which GOW provides.

```
C:\Users\mgalarny>gow --list
Available executables:
awk, basename, bash, bc, bison, bunzip2, bzip2, bzip2recover, cat,
chgrp, chmod, chown, chroot, cksum, clear, cp, csplit, curl, cut, dc,
dd, df, diff, diff3, dirname, dos2unix, du, egrep, env, expand, expr,
factor, fgrep, flex, fmt, fold, gawk, gfind, gow, grep, gsar, gsort,
gzip, head, hostid, hostname, id, indent, install, join, jwhois, less,
lesskey, ln, ls, md, make, md5sum, mkdir, mkfifo, mknod, mv, nano,
ncftp, nl, od, pageant, paste, patch, pathchk, plink, pr, printenv,
printf, pscp, psftp, putty, puttygen, pwd, rm, rmdir, scp, sdiff, sed,
seq, sftp, shasum, shar, sleep, split, ssh, su, sum, sync, tac, tail,
tar, tee, test, touch, tr, uname, unexpand, uniq, unix2dos, unlink,
unrar, unshar, uudecode, uuencode, vim, wc, wget, whereis, which,
whoami, xargs, yes, zip
```

Linux Commands on Windows

2. Download and install Anaconda. If you need help, please see this [tutorial](#).
3. Close and open a new command line (CMD).
4. Go to the Apache Spark website ([link](#))

Download Apache Spark™

1. Choose a Spark release: **2.1.0 (Dec 28 2016)**
2. Choose a package type:
Pre-built for Hadoop 2.7 and later
3. Choose a download type: **Direct Download**
4. Download Spark: [spark-2.1.0-bin-hadoop2.7.tgz](#)
5. Verify this release using the [2.1.0 signatures and checksums](#) and [project release KEYS](#).

Download Apache Spark

- a) Choose a Spark release
- b) Choose a package type
- c) Choose a download type: (Direct Download)
- d) Download Spark

5. Move the file to where you want to unzip it.

```
mkdir C:\opt\spark
```

```
mv C:\Users\mgalarny\Downloads\spark-2.1.0-bin-hadoop2.7.tgz
C:\opt\spark\spark-2.1.0-bin-hadoop2.7.tgz
```

6. Unzip the file. Use the bolded commands below

```
gzip -d spark-2.1.0-bin-hadoop2.7.tgz
```

```
tar xvf spark-2.1.0-bin-hadoop2.7.tar
```

7. Download winutils.exe into your **spark-2.1.0-bin-hadoop2.7\bin**

```
curl -k -L -o winutils.exe
```

```
https://github.com/steveloughran/winutils/blob/master/hadoop-2.6.0/bin/winutils.exe?raw=true
```

8. Make sure you have Java 7+ installed on your machine.

9. Next, we will edit our environmental variables so we can open a spark notebook in any directory.

```
setx SPARK_HOME C:\opt\spark\spark-2.1.0-bin-hadoop2.7
```

```
setx HADOOP_HOME C:\opt\spark\spark-2.1.0-bin-hadoop2.7
```

```
setx PYSARK_DRIVER_PYTHON ipython
```


```
setx PYSARK_DRIVER_PYTHON_OPTS notebook
```

Add ;C:\opt\spark\spark-2.1.0-bin-hadoop2.7\bin to your path.

Notes on the setx command: <https://ss64.com/nt/set.html>

See the video if you want to update your path manually.

10. Close your terminal and open a new one. Type the command below.



```
C:\Users\mgalarny\Desktop>pyspark --master local[2]
```

pyspark local

Notes: The PYSARK_DRIVER_PYTHON parameter and the PYSARK_DRIVER_PYTHON_OPTS parameter are used to launch the PySpark shell in Jupyter Notebook. The —master parameter is used for setting the master node address. Here we launch Spark locally on 2 cores for local testing.

Done! Please let me know if you have any questions. You can view the ipython notebook used in the video to test PySpark [here](#)!

