

[Get unlimited access](#)[Open in app](#)

Michael Galarnyk

[Follow](#)Apr 3, 2017 · 2 min read · [Listen](#)[Save](#)

## Install Spark on Windows (PySpark)

Install Spark on Windows (PySpark) + Configure Jupyter Notebook



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, m4, 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, uuencode, 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](#).





# 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. Keep in mind if you download a newer version, you will need to modify the remaining commands for the file you downloaded.
5. Move the file to where you want to unzip it.

`mkdir C:\opt\spark`

`mv C:\Users\mgalarnyk\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 PYSPARK_DRIVER_PYTHON ipython`

`setx PYSPARK_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.



Get unlimited access

Open in app

pyspark local

Notes: The PYSPARK\_DRIVER\_PYTHON parameter and the PYSPARK\_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 here or through [Twitter](#). You can view the ipython notebook used in the video to test PySpark [here](#)!

More from Michael Galarnyk

Data Scientist <https://www.linkedin.com/in/michaelgalarnyk/>

Published in **Towards Data Science** · Mar 6, 2017

**Exploratory Data Analysis (JHU Coursera, Course 4)**

Data Science 2 min read

Feb 28, 2017

**Making a github.io Website without Knowing Git**

Web Development 2 min read

Feb 16, 2017

**Recording Movies on Mac (Youtube)**

Live Streaming 1 min read

Published in **Towards Data Science** · Feb 16, 2017

**Getting and Cleaning Data (JHU Coursera, Course 3)**

Data Science 2 min read

Jan 30, 2017

**Linear Regression using R**

Machine Learning 1 min read

Follow

Love podcasts or audiobooks? Learn on the go with our new app.

Try Knowable