**How To Install Apache Spark On Windows**

[Apache Spark](https://spark.apache.org/) is a powerful framework to utilise cluster-computing for data procession, streaming and machine learning. Its native language is [Scala](https://en.wikipedia.org/wiki/Scala_(programming_language)). It also has multi-language support with Python, Java and R. Spark is easy to use and comparably faster than MapReduce. For example, you can write Spark on the Hadoop clusters to do transformation and machine learning easily.

If you have a large binary data streaming into your Hadoop cluster, writing code in Scala might be the best option because it has the native library to process binary data. For most of the Big Data use case, you can use other supported languages. In terms of machine learning, I found the performance and development experience of [MLlib](https://spark.apache.org/mllib/) (Spark’s machine learning library) is very good, but the methods you can choose are limiting. However, as Spark goes through more releases, I think the machine learning library will mature given its popularity.

To play with Spark, you do not need to have a cluster of computers. You can simply install it on your machine. When I develop with Spark, I typically write code on my local machine with a small dataset before testing in on a cluster. It is also handy for debugging if you can just run it on your local machine. You can also run Spark code on [Jupyter](http://jupyter.org/) with Python on your desktop.

Spark installation can be tricky and the other web resources seem to miss steps. If you are stuck with Spark installation, try to follow the steps below. It always works for me!

**Installation Steps**

1. Go to the [official download page](official%20download%20page) and choose the latest release. For the package type, choose ‘Pre-built for Apache Hadoop’.

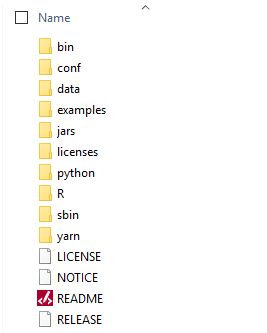
Make sure to put pre-built as below.



To unzip .tgz file. If you have Cygwin or Git Bash, you can use the command below. Otherwise you can use WinZip or WinRAR.

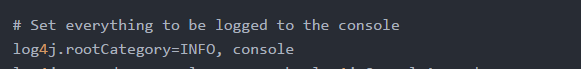
* tar -xzf spark-2.2.0-bin-hadoop2.7.tgz

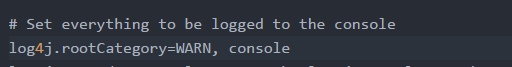
1. Create a folder called Spark under your user folder (C:\Users\<user name>\Spark) and copy and paste the unzipped content



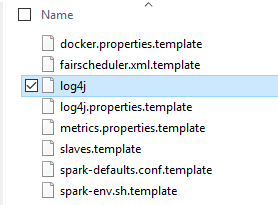
1. Go to the Conf folder and open log file called, log4j.properties.template.

Change INFO to WARN (It can be ERROR to reduce the log)



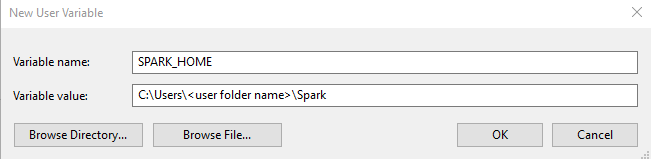


1. Remove .template so that Spark reads the file



1. We need to configure path. Go to Control Panel 🡪 System 🡪 Advanced system settings

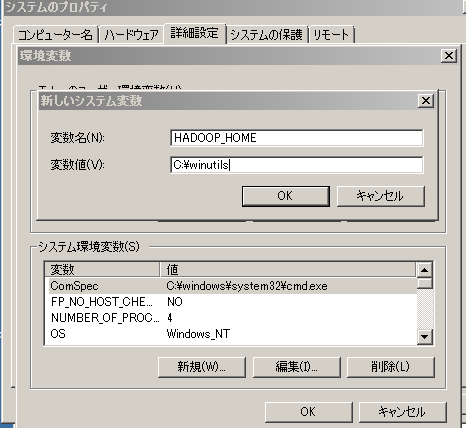
Add SPARK\_HOME in the user variable



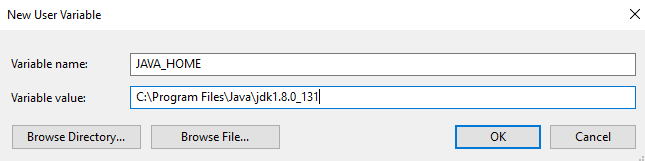
1. Select Path in the above screen shot and press Edit. Then add ;%SPARK\_HOME%\bin



1. For Hadoop 2.7, install winutils.exe from [here](https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe).
2. Create a folder called winutils in C drive and create a folder called bin inside.
3. Add the user variable %HADOOP\_HOME%



If you are getting java.lang.UnsupportedClassVersionError: JVMCFRE003 bad major version, you need to define a user variable called JAVA\_HOME so that Spark can use the correct version of Java.



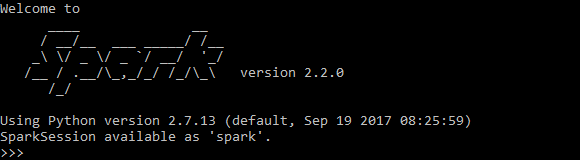
In the end, the environment variable have 3 new paths.



1. Finally chmod on /tmp/hive folder (This step was not necessary).

C:\winutils\bin\winutils.exe chmod 777 /tmp/hive

1. pysark will start Pyspark.



1. Run below so that we can use pyspark on Jupyter notebook. Set two variables and execute pyspark. The jupyter notebook will open and you can start writing pyspark code.

set PYSPARK\_DRIVER\_PYTHON=jupyter

set PYSPARK\_DRIVER\_PYTHON\_OPTS=notebook

pyspark