ukdataservice.ac.uk

Installing Spark on a Windows PC

UK Data Service

Author: UK Data Service
Updated: May 2016
Version: 1.0

Contents

1. Introduction

Apache Spark is an open source parallel processing framework that enables users to run large-scale data analytics applications across clustered computers. Spark might be considered as a one-stop tool for big data processing, providing data manipulation facilities to slice and dice datasets as well as statistical functionality and visualisation capabilities to present results.

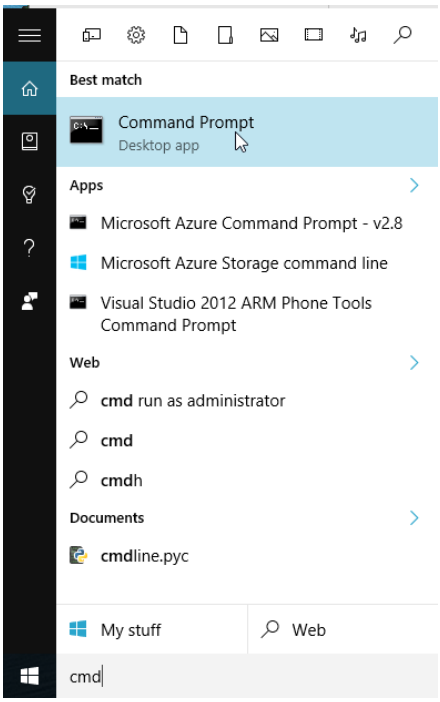
Although the real speed benefits of using Spark processing can only be realised in a clustered computing environment, Spark itself can be installed in a standalone environment on a variety of systems including a Windows PC. This guide provides instructions on installing on a Windows PC where it can be used for training and development purposes.

2. Step-by-step installation guide

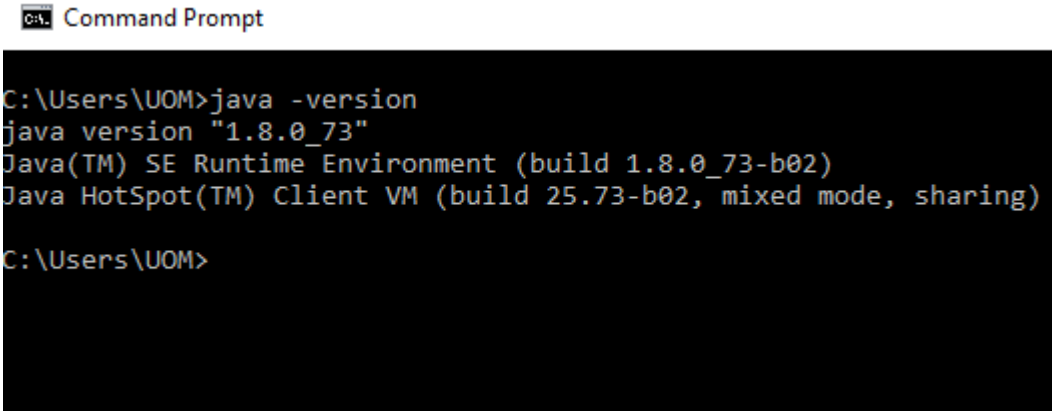
The Spark installation process is simple, although there are a few steps you have to follow to make the installation more user- friendly. The steps below will guide you through the whole process.

Step 1 – Make sure Java is installed

The first step is to make sure that Java is installed on your PC by typing *cmd* into the search panel on the start menu, and clicking on the *Command Prompt* Desktop app. The screenshot below shows this step on a Windows 10 machine.



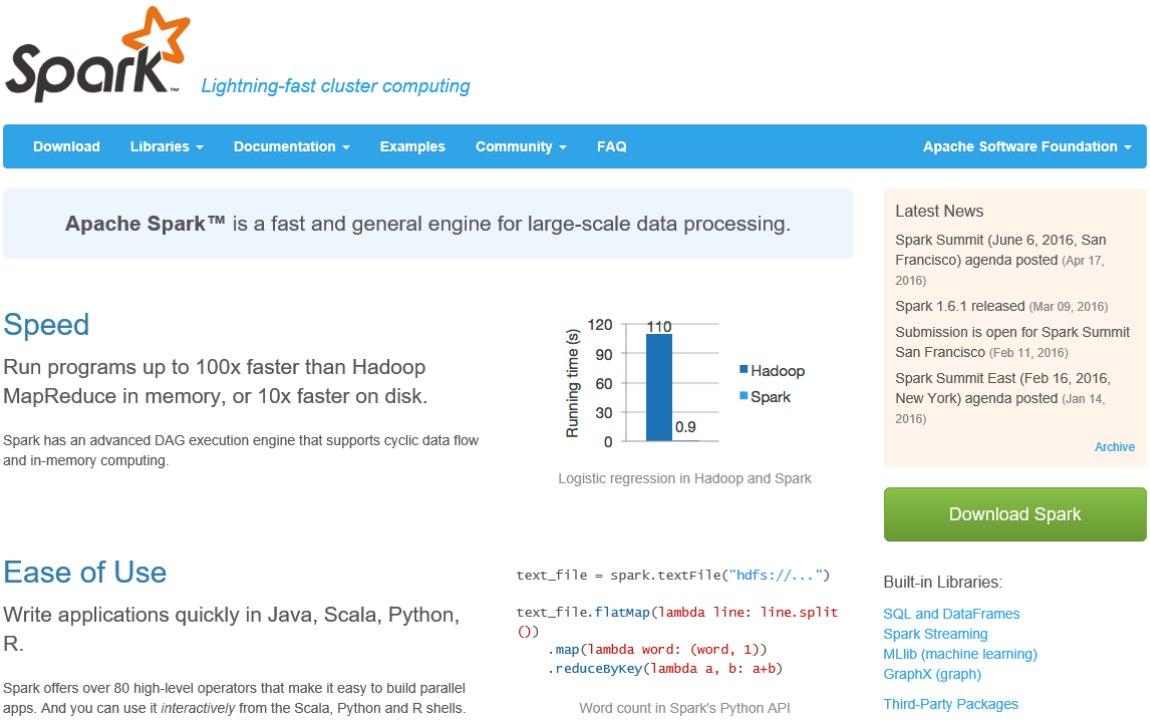
This will open a command line terminal window as shown below into which you type:
`Java-version`
The resulting display should be similar to the screenshot below:



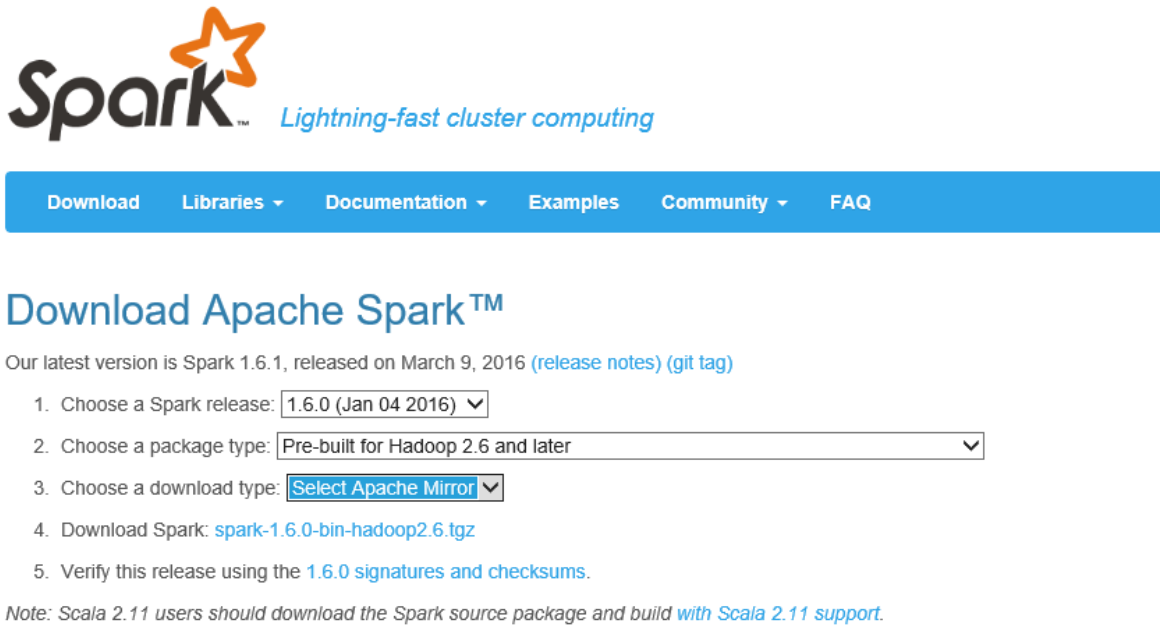
The actual Java version number doesn't matter as long as it is above 1.7.
If you don't have Java installed, then you will get a message along the lines of 'Java is not recognized as a command'. This is unlikely to be the case on Windows 7 or above but if this is the case, please download Java.

Step 2 - Download the Spark software

You can download Spark from the Apache Spark website. To do so, click on the *Download Spark* button on the right hand side of the webpage.

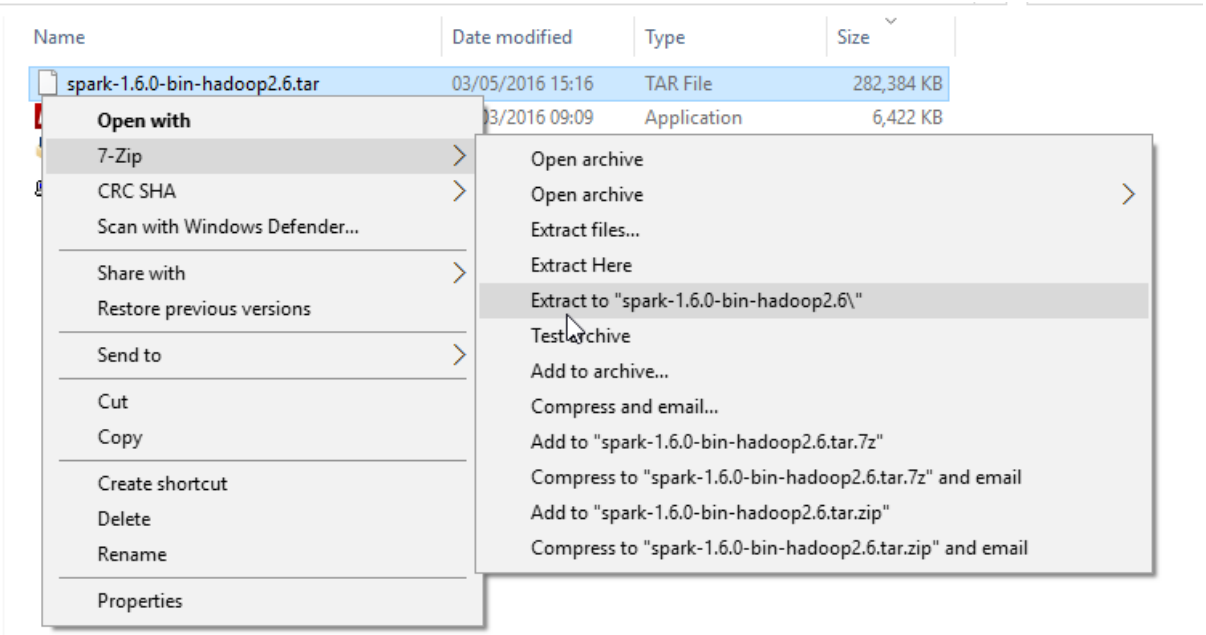


On the download page you can select the version of Spark you want, along with a package type and a download type. The options selected and shown in the screenshot below are suitable.
The fourth line - Download Spark - provides a link for you to click on (the link changes dynamically based on your choices for 1 & 2). Once you click on it, you will be asked to select a suitable download site from a list. Any of the sites in the list should be OK but the download may be quicker if you choose a local (i.e. same country) site. The download size is approx. 275Mb.

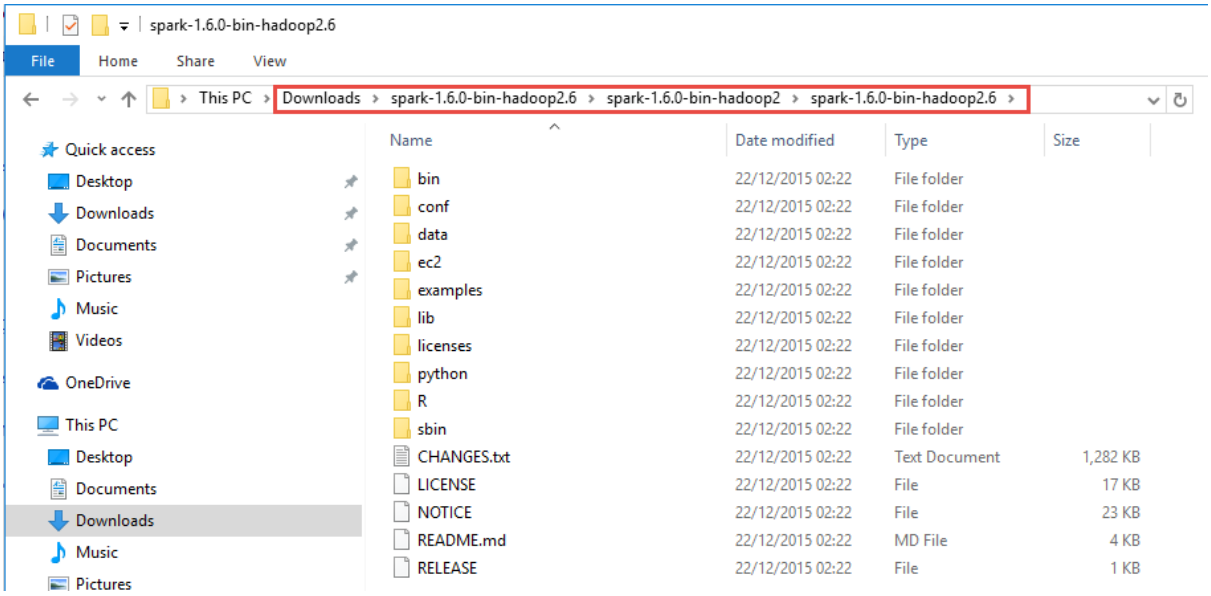


Step 3 - Uncompress the file

The downloaded compressed file will have a *.tar* extension. Unfortunately Windows File Explorer is unable to uncompress such files so it is necessary to download and install a 3rd party application to perform the uncompression – unless you already have one installed. One of the open source applications available is 7-zip and it can be downloaded from their official site http://www.7-zip.org/.
If you have 7-zip installed then right mouse clicking on the downloaded file in File Explorer should allow you to select 7-zip and to specify where the file is to be un-compressed (extracted) to.



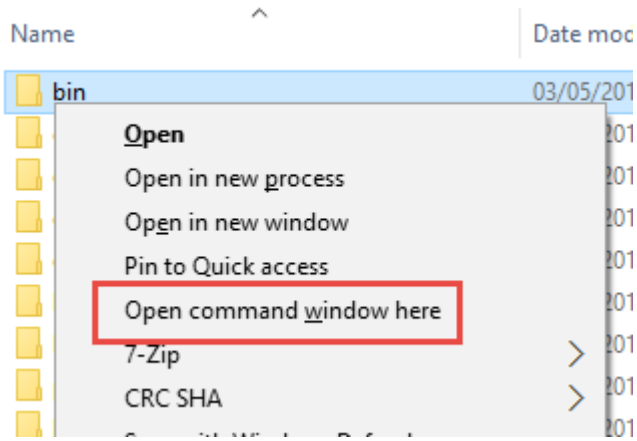
The uncompressed file is actually a folder containing another compressed file. You can uncompress this file exactly the same way and this time the resulting folder will contain a set of uncompressed folders and files.



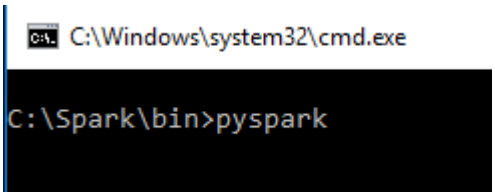
Because of the long path names you now have it is convenient to simply create a folder called *Spark* (c:\spark) and to copy all of the above uncompressed folders and files into it. You can then delete everything apart from your original downloaded file to save space.

Step 4 - Test run Spark

You are now in a position to try a test run in the Spark system.
Navigate to the folder you have copied the files and folders into and right mouse click on the 'bin' folder whilst holding down the Shift key.



Select the 'Open command window here' option. This will open a command line window with the prompt indicating that you are in the c:\spark\bin folder. From here you can type the command `pyspark`.



A rather large number of messages will scroll across the screen but at the end of them you should see something similar to the screenshot below.

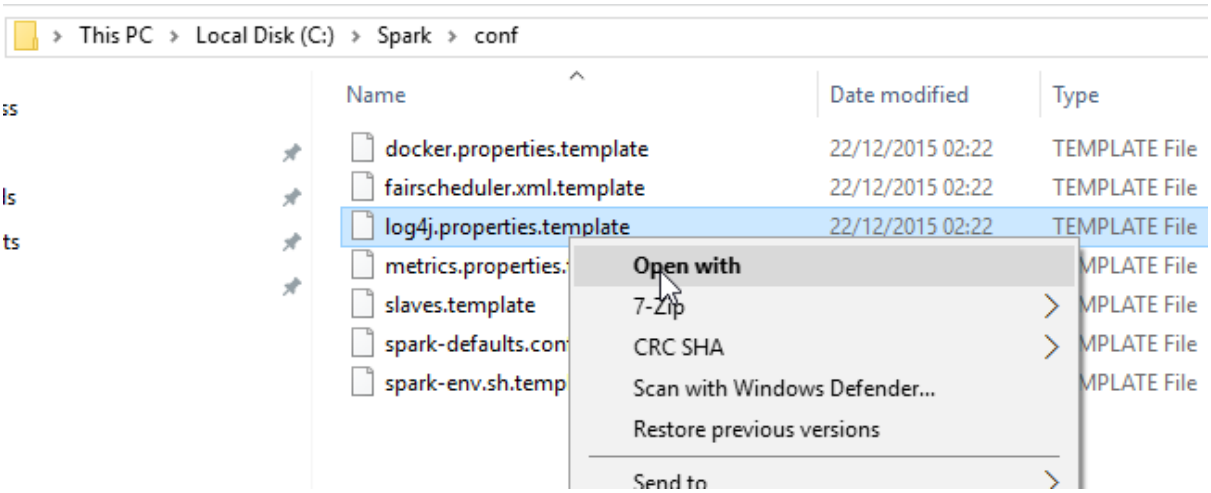
This tells you that Spark has loaded successfully. The three chevrons at the bottom indicate that you are in the PySpark shell and if you wish you could start typing PySpark code here. For now we just want to exit the shell using the `'quit();'` command.

Step 5 - Completing the configuration

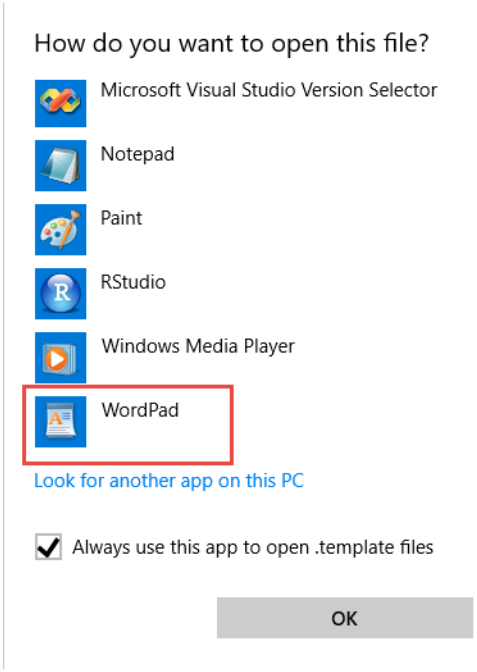
Although all of the messages that appear when running Spark are intended to be informative, they are generally rather lengthy. In addition to these verbose information messages, there is also a genuine error message when you start Spark on Windows. This is a known problem and there is an easy fix to it which we can apply.

Step 5.1 - Dealing with the information messages

In File Explorer navigate to the 'conf' folder within your Spark folder and right mouse click the 'log4j.properties.template' file. Select the 'Open with' from the menu.



From the list of Apps, choose *WordPad*.



You can also use other editors, such as *Notepad++* or *PSPad*, but do not use *Notepad*.

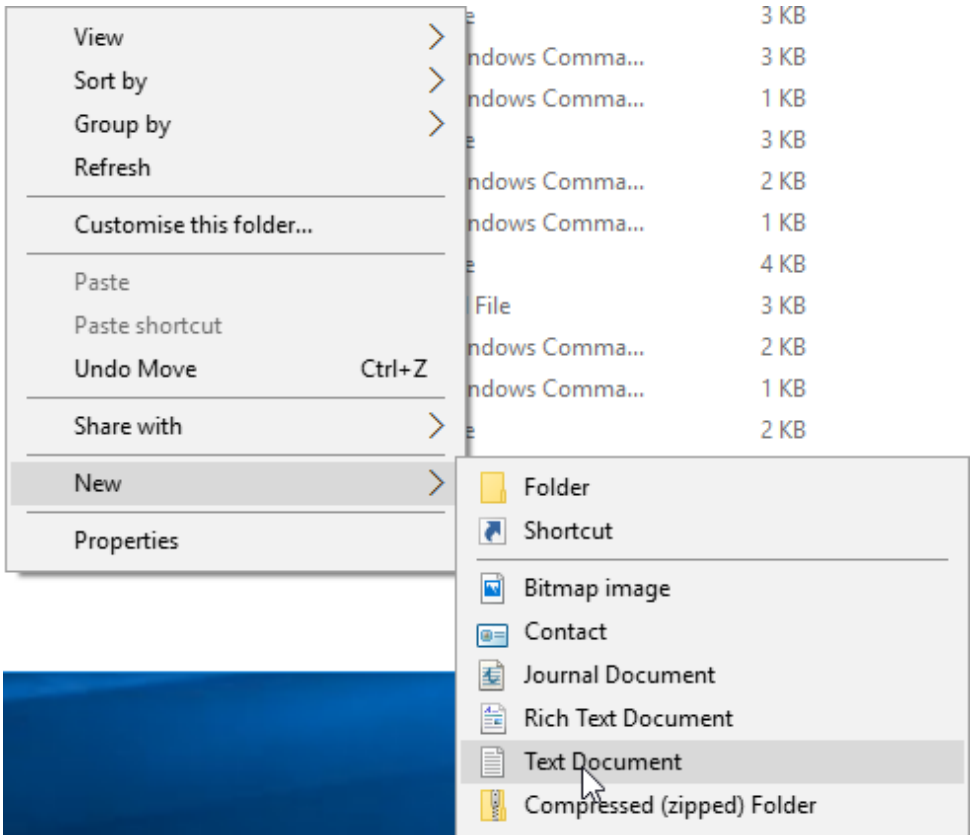
```
# Set everything to be logged to the console
log4j.rootCategory=INFO, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n

# Settings to quiet third party logs that are too verbose
log4j.logger.org.spark-project.jetty=WARN
log4j.logger.org.spark-project.jetty.util.component.AbstractLifeCycle=ERROR
log4j.logger.org.apache.spark.repl.SparkIMain$exprTyper=INFO
log4j.logger.org.apache.spark.repl.SparkILoop$SparkILoopInterpreter=INFO
log4j.logger.org.apache.parquet=ERROR
log4j.logger.parquet=ERROR
```

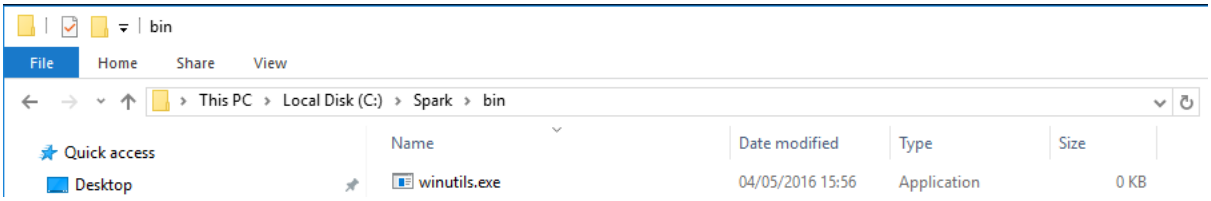
Notepad was use in the screenshot above, simply because it provides a clearer view of the file. If you use *WordPad*, the text will be the same, but in a different font and more spread out. The three highlighted 'INFO' words and the one 'WARN' all need to be changed to 'ERROR'. The file should then be saved with the name 'log4j.properties'.

Step 5.2 - Add the winutils file

The error message you get when you run Spark in Windows is caused by Spark being unable to locate a file called 'winutils.exe'. This is a known problem and will no doubt be fixed in some future release of PySpark for Windows. In the meantime, the simple workaround is to create a file with the name 'winutils.exe' in the 'bin' folder.
Navigate to the 'bin' folder in File Explorer, right mouse click and select 'New' and 'Text Document'.

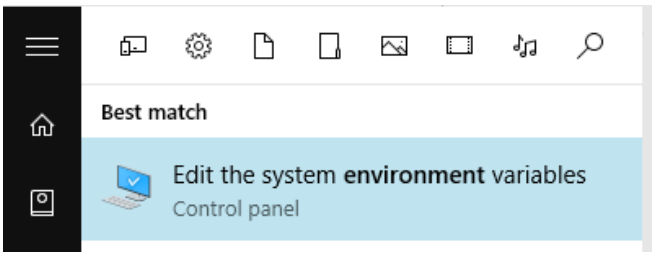


Rename the created document as 'winutils.exe'. You will get a warning about changing the file extension which you can accept. The result will be a file called winutils.exe with a length of zero bytes.

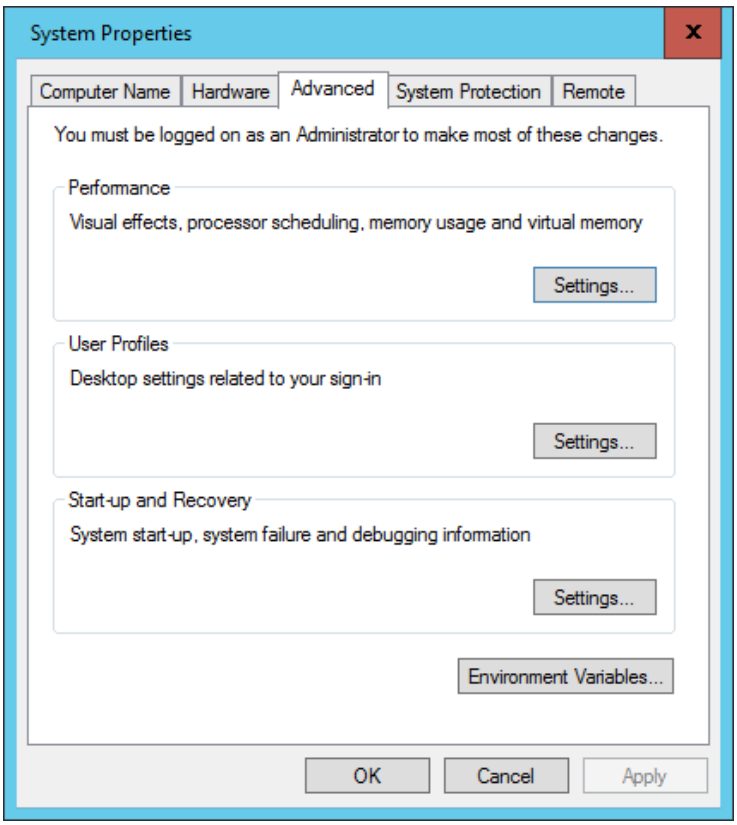


Step 5.3 - Add environment variables

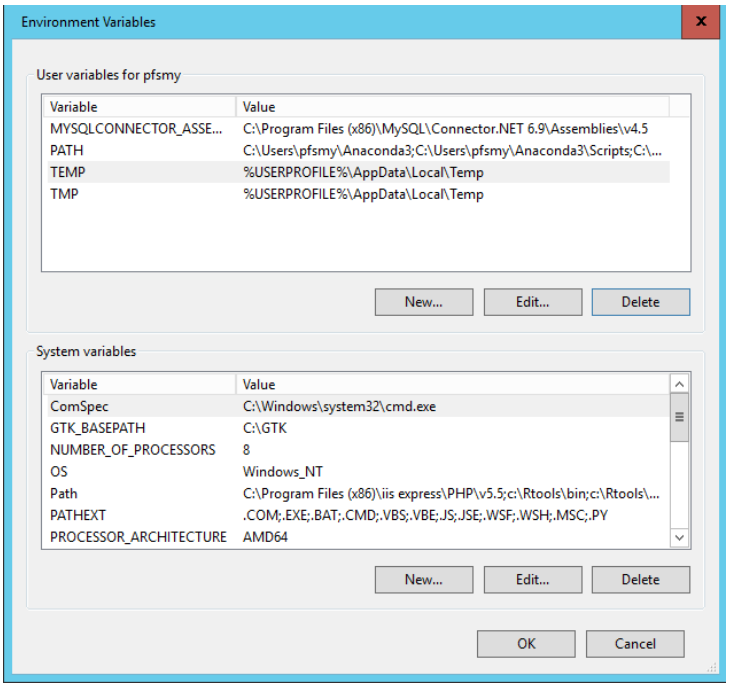
The last step is to add environment variables *HADOOP_HOME* and *SPARK_Home* to the Windows environment.
In the Windows Start search panel type in 'environment' and select the option 'Edit the system environment variables'.



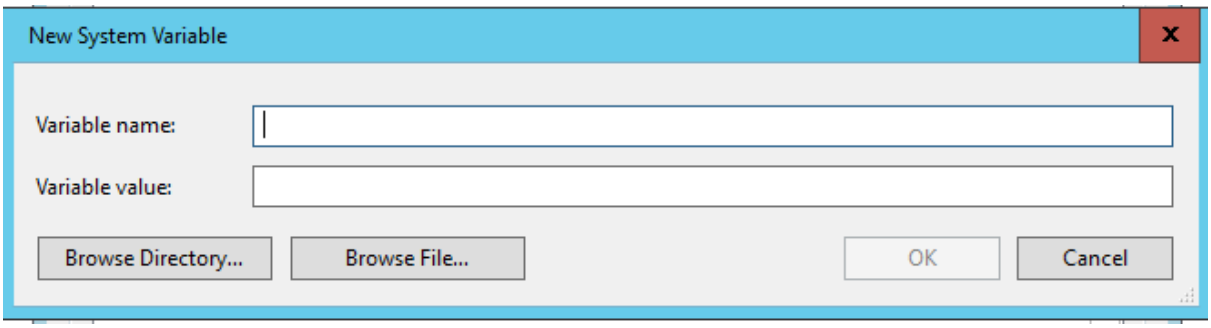
In the next window click on 'Environment Variables' at the bottom.



Next, click on 'New' at the bottom of the 'System variables' panel in the lower half of the screen.

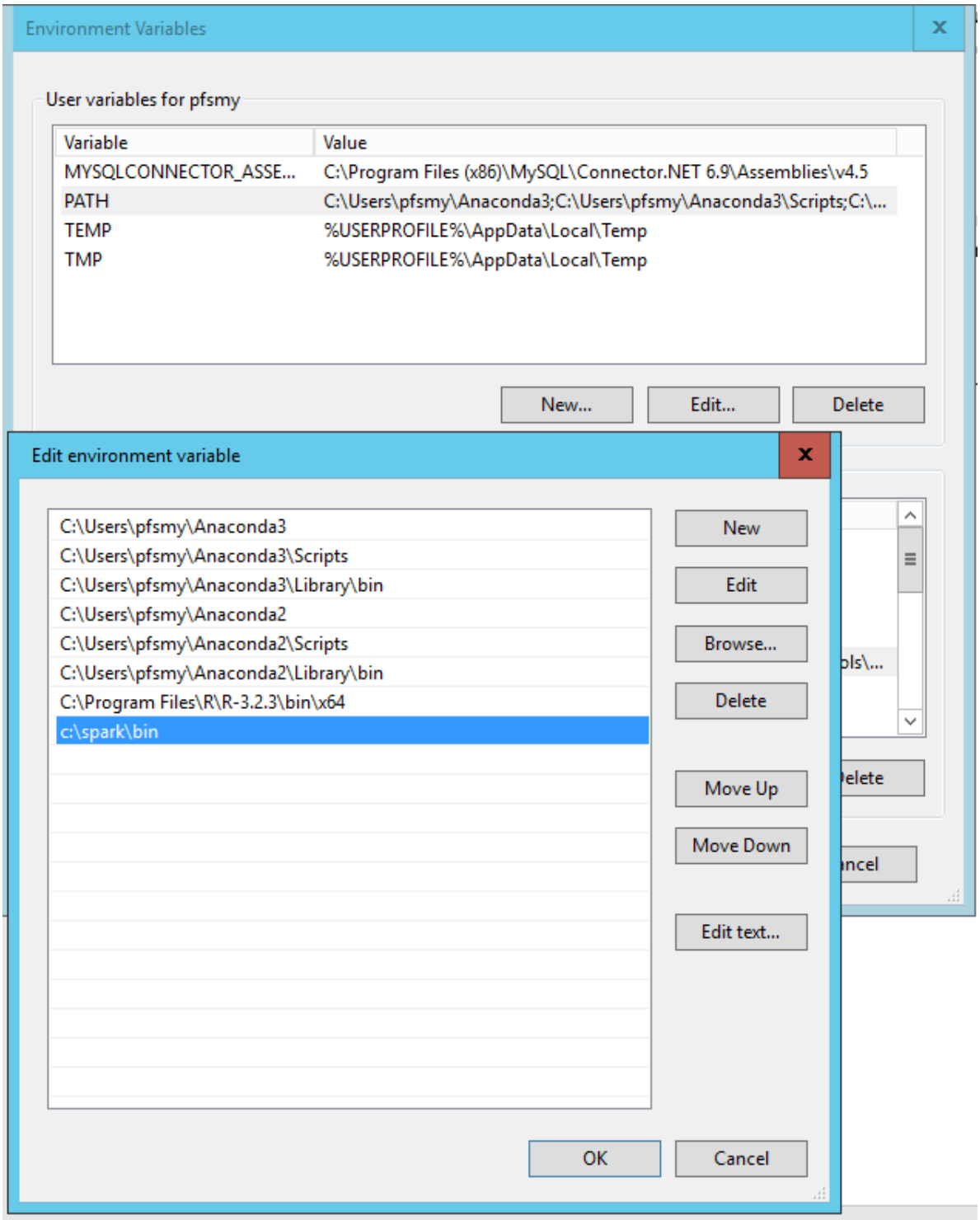


In the window that pops up type *HADOOP_HOME* as the 'Variable name' and *c:\spark* (assuming this is what you called your spark folder) in the 'Variable value' box and click OK. Repeat the process to add another new variable called SPARK_HOME using the save variable value as before.



Step 5.4 - Add Spark to the path

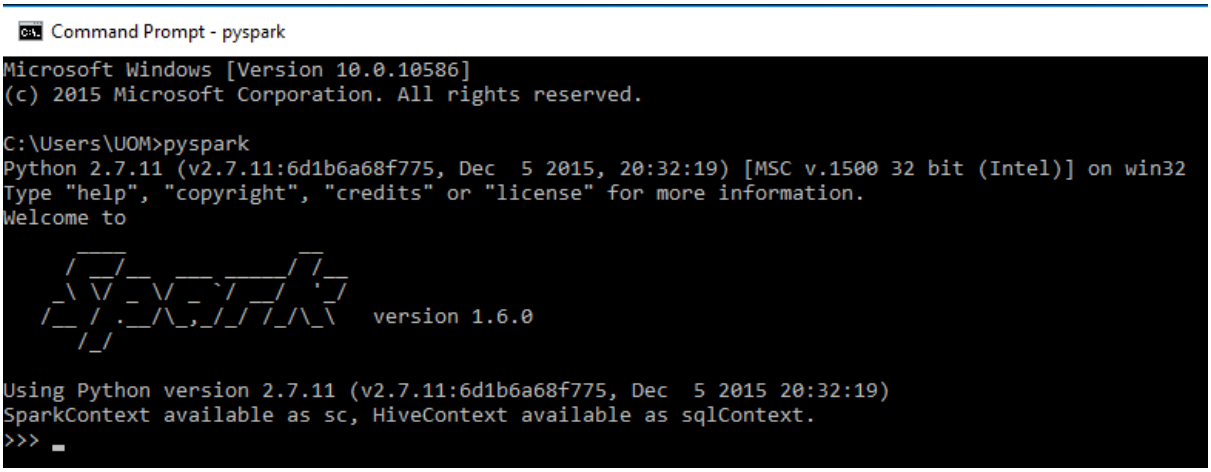
One final thing to do is to add the location of the Spark 'bin' folder to the path statement for the user. To do this select the PATH variable in the user section (top half) of the 'Environment Variables' panel and click on 'Edit'.



In the new window there will be a list of all of the currently defined variables (your list may differ from that shown in the screenshot above). Click on 'New' and you will be able to add a new entry at the bottom of the list as shown. You may need to type something different if you have a different location for Spark. Click OK to finish.
The advantage of adding the Spark library to the path statement is that you can run Spark from any Command Prompt. You will however have to restart the PC before this takes effect.

Step 6 - Re-Test Spark

With all of the configuration complete, and the PC restarted, you should be able to run Spark and type *pyspark* into the command line window. At this point there should be no error messages showing.



May 2016

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service provides the UK's largest collection of social, economic and population data resources

UK Data Service

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL