**CSCI316 Software and Programming Packages**
**Installation Guide (for Windows)**

(1) Download Java 8 and set environment variable JAVA_HOME

(2) Download Anaconda (with Python 3) and set PYTHON_HOME

- Note. use “where python” in Anaconda Prompt to check the path.

(3) Download Apache Spark 2.x (https://spark.apache.org/downloads.html, select the “pre-built for Apache Hadoop 2.7).

- After uncompressing the file, a folder such as spark2.4.4-bin-hadoop-2.7 is created. Can move this folder to a convenient location, say, the root directory of C Drive.

(4) Download the correct version of winutils.exe from https://github.com/steveloughran/winutils. For example, for hadoop 2.7, download winutils.exe from https://github.com/steveloughran/winutils/tree/master/hadoop-2.7.1/bin. Save it to C:/spark2.4.4-bin-hadoop-2.7

(5) Set environment variables:

- SPARK_HOME = C:\spark2.4.4-bin-hadoop-2.7
- HADOOP_HOME = C:\spark2.4.4-bin-hadoop-2.7
- PYSPARK_PYTHON = PYTHON_HOME
- PYSPARK_DRIVER_PYTHON = jupyter
- PYSPARK_DRIVER_PYTHON_OPTS = 'notebook'

Add C:\spark2.4.4-bin-hadoop-2.7 to Path (for system environments).

(6) Use conda to install findspark:

- conda install -c conda-forge ~~finspark~~ findspark

(7) Also use conda to install the following libraries (Note. Some libraries are installed with Anaconda):

- TensorFlow2
- Jupyter (Jupyter notebook)
- IPython3
- Scikit-Learn
- Scientific computing libraries: SciPy, NumPy, Pandas, Matplotlib
- Some dependence libraries: Py4J, pyarrow, psutil, cairocffi

Use references:
https://changhsinlee.com/install-pyspark-windows-jupyter/
https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c