Spark Environment Set up

Setting Up Environment

- Following slides contain instructions to install Spark and Machine Learning environment on MAC and Windows
- NOTE – Spark will run in standalone mode (not distributed)
- Consists of 3 steps:
    1. Install Anaconda (Jupyter notebook)
    2. Install Spark
    3. Install “findSpark” package

Setting Up Environment - Anaconda

- https://www.anaconda.com/distribution/ - use the link to download the distribution for your machine (Windows or MAC).
- https://docs.anaconda.com/anaconda/install/ - follow the instructions for the machine type ("Installing on Windows", or "Installing on macOS")

Setting Up Environment – Spark on MAC

- Follow the instructions in the following link to install Spark on MAC
  - JAVA, Spark, Scala and SBT
  - Do not install python since it is already installed by Anaconda in the pervious step
  - Make sure to set up environmental variables in the "bashrc" file as defined in the instructions

- https://medium.com/luckspark/installing-spark-2-3-0-on-macos-high-sierra-276a127b8b85

Setting Up Environment – Spark on Windows

- Follow the instructions in the following link to install Spark on Windows
  - Spark, JAVA and Winutils will be installed
  - Do not install python since it is already installed by Anaconda in the pervious step
  - Make sure to set up environmental variables correctly

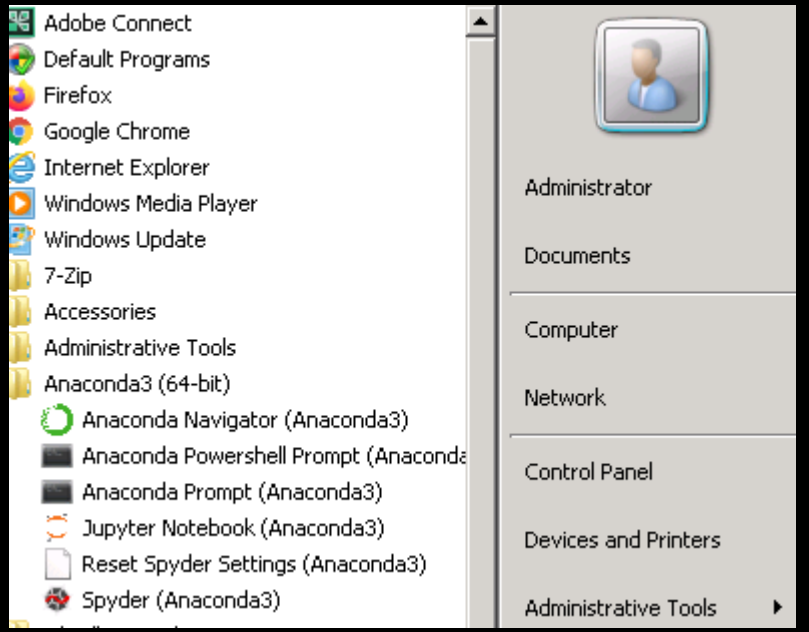
- https://www.knowledgehut.com/blog/big-data/how-to-install-apache-spark-on-windows

“findSpark” installation

```
# Set up the environment for using pyspark
import findspark

findspark.init()
```

- Need to install “findSpark” utility to set up environment variables correctly so that Jupyter notebook can execute Spark code
- Open “Anaconda Prompt”
- At the prompt type:
  - conda install –c conda-forge findspark



Jupyter Notebook with Spark (Linux/MAC)

Information in the .bashrc file for Linux Spark runs locally

```
function snotebook()
{
SPARK_HOME=/usr/local/spark/spark-2.4.2-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH

export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'

$SPARK_HOME/bin/pyspark --master local[2]
}
```