Installation of Spark in Ubunbu in virtual box

- Installing Virtual Box. Choose fixed size disk, relatively large memory usage... Check the current installation. Also choose the proper display driver. If the resolution is small, then Jupyter notebook will not work.
- Installing Ubuntu.
- Make sure python3 is already there (come with Ubuntu). If so, install Jupyter notebook.
  - Run 'pip3 install jupyter'
  - Run 'jupyter note book' and make sure it is working.
  - Ctrl-C to terminate the jupyter notebook server
- Install Java, scala, py4j
  - Install Java. First run 'sudo apt-get update', then run 'sudo apt-get install default-jre', then test with 'java -version'
  - Install Scala by 'sudo apt-get install scala', then try 'scala -version'
  - Install py4j by 'pip3 install py4j', which connects python to java, scala
- Install Spark and tell python where to find Spark
  - Searching Apache Spark from web, and download and save to file. Then move the file to home folder, and unzip with 'sudo tar zxvf filename'.
  - Run: export SPARK_HOME='home/ubuntu/spark-2.4.0-bin-hadoop2.7'
  - Run: export PATH=$SPARK_HOME$:PATH
  - Run: export PYTHONPATH=$SPARK_HOME/python$:PYTHONPATH
  - Run: export PYSPARK_DRIVER_PYTHON="jupyter"
  - Run: export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
  - Run: export PYSPARK_PYTHON=python3
- Handling permissions of folder
  - In home folder: sudo chmod 777 spark-2.4.0-bin-hadoop2.7/.
  - cd spark-2.4.0-bin-hadoop2.7 folder, run: sudo chmod 777 python. This changes permission of python folder within the above hadoop2.7 folder.
  - cd .../python folder, run: sudo chmod 777 pyspark
- Setting up pyspark
  - Now if we run python3 at home, and run import pyspark within python, it will give error. However, if we go to the folder .../spark2.4.0phadoop2.7/python, and then run python3, and then import pyspark, then there is no error. Now we fix this.
  - Go to home directory, run: pip3 install findspark
  - cd /home/spark-2.4.0-bin-hadoop2.7/, run pwd will give a path, copy that path to the following code in python script.
  - Run python3 in home folder, and then run: import findspark findspark.init('/home/spark-2.4.0-bin-hadoop2.7/') import pyspark Now we don't have error anymore.

    - Note we can also do the above with other way. For example export our path to bash.

Install pyspark in other places

- AWS EC2 PySpark setup
- Databricks Setup
- AWS EMR Cluster Setup
- https://medium.com/@ashish1512/how-to-setup-apache-spark-pyspark-on-jupyter-ipython-notebook-3330543ab307 (https://medium.com/@ashish1512/how-to-setup-apache-spark-pyspark-on-jupyter-ipython-notebook-3330543ab307) Following this link to install on Windows with Jupyter notebook. However, during the installation, I have some issues and thus it might not work properly. Check it out in the future.

Check details in the course video.