COMP9313 Lab1 Installation Instructions

Introduction

In this lab, we will install and configure the programming environment for **MapReduce**
For COMP9313-20T2, we will be using the following Python and Java packages (Please note the corresponding versions):

1. `Python 3.6.5`
2. `PySpark 2.4.6`
3. `Spark 2.4.6`
4. `Hadoop 2.7`
5. `Jdk 1.8`

The students willing to setup the programming environment in their Personal Computers/ Laptops, should:

1. **MAKE SURE THAT YOU INSTALL THE APPROPRIATE VERSIONS (mentioned above).**

- **Your working directory should not contain `SPACE` , otherwise, PySpark may not function appropriately.**

1. Install JDK

1. Use the following link to install jdk 1.8:
   https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html (https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html)
2. You need to set the environment variable `JAVA_HOME` as follows:

   - Windows users:
     - Follow https://javatutorial.net/set-java-home-windows-10 (https://javatutorial.net/set-java-home-windows-10), and make sure your JDK directory is correct.
   - Mac users:
     - Follow https://mkyong.com/java/how-to-set-java_home-environment-variable-on-mac-os-x/ (https://mkyong.com/java/how-to-set-java_home-environment-variable-on-mac-os-x/), and set the environment variable in `.bash_profile` .
   - Linux users:
     - Follow https://stackoverflow.com/questions/9612941/how-to-set-java-environment-path-in-ubuntu (https://stackoverflow.com/questions/9612941/how-to-set-java-environment-path-in-ubuntu), and set the environment variable in `~/.bashrc` .

Finally, use your terminal/windows shell/linux command line, and type `java -version` to check whether your installation is correct.
If you can see **java version 1.8**, it means you have successfully installed the `jdk` .

2. Install Python (Anaconda) + Jupyter Notebook

1. Use the following link to download the latest version of Anaconda for Python3+.
   https://www.anaconda.com/products/individual (https://www.anaconda.com/products/individual)

   - Mac/Linux Users:
     - Follow the link to install Anaonda: https://medium.com/@menuram1126/how-to-install-anaconda-on-ubuntu-16-04-538009ca7936 (https://medium.com/@menuram1126/how-to-install-anaconda-on-ubuntu-16-04-538009ca7936)
   - Window Users:
     - Install via .exe file.
2. Once installed, Anaconda lets you use the `jupyter notebook` . You should be able to run the following command in the terminal to open the Jupyter Notebook.

   - `jupyter notebook`

3. Set Conda Environment

1. After installing the Python (i.e., Anaconda), use the following commands in the terminal to create python programming environment named: `COMP9313`

   - `conda create -n COMP9313 python=3.6.5`

1. And, activate the environment by using the following command:

   - `conda activate COMP9313`

- Note: After activating the environment, you should be able to see `(COMP9313)` in your terminal screen.

4. Install PySpark

1. Once you have activated your environment `(COMP9313)` , you can use the corresponding `pip` to install the required Python Packages.
2. For example, you can use `pip install pyspark==2.4.6` to install the required PySpark Package.

Use terminal/windows shell/linux command line, and type `pyspark --version` . You should be able to see: **SPARK version 2.4.6**, which implies you have successfully installed PySpark.

5. Install Spark & Hadoop

1. Use the following link to download and install Spark and Hadoop
   https://spark.apache.org/downloads.html (https://spark.apache.org/downloads.html)

   - Choose the Spark release: 2.4.6 (June 05 2020)
   - Choose the package type: Pre-built for Apache Hadoop 2.7
   - Download Spark: `spark-2.4.6-bin-hadoop2.7.tgz`

- You also need to set the environment variable: `export SPARK_HOME=YOUR_SPARK_DIRECTORY_PATH`
- Windows users: Download `hadoop winutils` and put it into `YOUR_SPARK_DIRECTORY_PATH/bin` , otherwise, you may not run Hadoop.

6. Link Conda Environment with Jupyter Notebook

1. Once, we have setup the python environment and installed all the Packages, we need to link the environment `COMP9313` with the `jupyter notebook` . For this, you need to run following commands in the terminal:

   - `conda activate COMP9313`
   - `conda install -n COMP9313 ipykernel`
   - `python -m ipykernel install --user --name COMP9313`

1. You should be able to see `COMP9313` in the dropdown menu of Jupyter Notebook: `Kernel > Change kernel > COMP9313` .
2. Select `COMP9313` as your `jupyter notebook` Kernel and run the following script in the notebook to test your environment.

Test your Environment

In [1]:

```
# In this section, we play around a simple Hello-World example.

from pyspark import SparkConf, SparkContext

# create SparkConf and SparkContext
conf = SparkConf().setMaster("local").setAppName("hellow_9313")
sc = SparkContext(conf = conf)

# your input data
data = ["hello", "world", "hello", "world", "word", "count", "hello"]

# transform data into spark Rdd
rdd = sc.parallelize(data)

# use map to form each element x become (x, 1)
# then use reduceByKey to aggregate intermediate result from rdd.map()
result = rdd.map(lambda word:(word, 1)).reduceByKey(lambda a, b:a + b)

# use rdd.collect() to transfrom into collecton and print your final result
collection = result.collect()
for line in collection:
    print(line)

sc.stop()
```

```
('hello', 3)
('world', 2)
('word', 1)
('count', 1)
```

If you can run this program without any error, you have completed this lab. **CONGRATULATIONS...!**

FAQs

If your program shows: *Python in worker has different version 2.7 than that in driver 3.6, PySpark cannot run with different minor versions. Please check environment variables PYSPARK_PYTHON and PYSPARK_DRIVER_PYTHON are correctly set.*, this occurs because you may have multiple versions of python in your machine. For this, you need to set two environment variables:

- export PYSPARK_PYTHON = python3
- export PYSPARK_DRIVER_PYTHON = python3

In [ ]:

```
```