Spark installation instructions

If you have any questions related to spark installation, please feel free to contact TA and mention MSBD5003 on the title. Also, you are encouraged to set up a discussion about your questions.

A pre-installed Spark version ubuntu has been launched at http://www.cse.ust.hk/msbd5003/ubuntu.ova (http://www.cse.ust.hk/msbd5003/ubuntu.ova) . If you don't get Spark installed properly, you can try this ova file by following the VirtualBox install instruction but switch the image to this one. Jupyter notebook with PySpark Kernel support and GraphFrame file are already installed in this system, you can modify pyspark driver and use the pyspark shell directly with GraphFrames as introduced below.

The login user is msbd5003 and password is HKUST.

Spark Installation

Windows 10

In this course, you are highly recommended to use Linux/Unix system to install Spark. You may use the "Windows Subsystem for Linux (WSL)" to get a Linux environment on your laptop if you are running Windows 10. Here is how to activate it.

1. Open PowerShell as **Administrator** and run:
   ```
   Enable-WindowsOptionalFeature -Online -FeatureName Microsoft-Windows-Subsystem-Linux
   ```
2. Restart your computer when prompted
3. Go to 'Control Panel > Programs > Turn Windows features on or off', and make sure "Windows Subsystem for Linux" is checked.
4. Search for and install a Linux distribution (e.g. Ubuntu) in Microsoft Store.
5. Follow the instructions for Linux to install Spark in WSL.
6. Jupyter Notebook+WSL may encouter a problem (https://github.com/jupyter/notebook/issues/4594) if it could not find your default browser. (Thanks Yingjie for finding this!) You may solve it with the following steps:

   (1) create an alias for google chrome: open ~/.bashrc and add the following line:

   alias chrome="/mnt/c/Program\ Files\ $x86$/Google/Chrome/Application/chrome.exe"

   (2) run jupyter notebook --generate-config , and open it with nano ~/.jupyter/jupyter_notebook_config.py .

   (3) find these two settings and set them to the values below:

   c.NotebookApp.browser = u'/mnt/c/Program\ Files\ $x86$/Google/Chrome/Application/chrome.exe %s'

   c.NotebookApp.use_redirect_file = False

   Alternatively, you can do a quick-fix: Press "Ctrl+C" and then enter "n" to not shut down the server but go back to ubuntu. Then you can copy the link with the token and open it in a browser to access the notebook.

Linux (Ubuntu)

All the following operations should be done under Terminal.

1. Download Spark 3.0:

```
wget https://downloads.apache.org/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz (https://downloads.apache.org/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz)
```

(You may go to **https://spark.apache.org/downloads.html** (https://spark.apache.org/downloads.html) to find other mirrors. If you are in mainland, try **https://mirrors.tuna.tsinghua.edu.cn/apache/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz** (https://mirrors.tuna.tsinghua.edu.cn/apache/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz))

2. Unpack:

   ```
   tar xf spark-3.0.3-bin-hadoop2.7.tgz
   ```

3. Check if Java8 is installed:

   ```
   java -version
   ```

   - If your java version is 1.8.xxx, you may skip this step.
   - If no java is installed, install as below.

     ```
     sudo add-apt-repository ppa:openjdk-r/ppa
     ```

     ```
     sudo apt-get update
     ```

     ```
     sudo apt-get install openjdk-8-jdk
     ```

   - If you have multiple java versions, choose the right version as below

     ```
     sudo update-java-alternatives --set java-1.8.0-openjdk-amd64
     ```

     then restart the terminal

4. Check if python3 is installed:

   ```
   python3 -V
   ```

   1. If not, install it by typing:

      ```
      sudo apt-get install python3
      ```

   2. Bind spark's python driver to python3:

      ```
      nano ~/.bashrc
      ```

      (You may use any text editor)

      Move to the end of the file, and add the following line: export PYSPARK_PYTHON=python3

      ```
      source ~/.bashrc
      ```

5. Now you are ready to start the Spark shell.

   ```
   cd spark-3.0.3-bin-hadoop2.7
   ```

   ```
   pyspark
   ```

6. In the python shell, input 'spark'. If you see a character painting of spark, your installation is successful.
7. If you see a connection error in Spark, the easiest fix is to turn off the firewall.

(**Optional**) If you want to use Spark more skillfully, it's better for you to get familiar with Basic Linux Commands and Basic Bash Operations. You can refer to the following book **http://linux-training.be/linuxfun.pdf** (http://linux-training.be/linuxfun.pdf) and learn Part III, Chapter 14 of Part IV, Part VIII and any other parts you are interested in. You're also encouraged to Google and learn. Most of these commands also work on MacOS.

MAC OS

1. Install Homebrew:

   ```
   /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
   ```

   (If you are in mainland you may encounter a connection error. You may follow this **link (https://www.jianshu.com/p/ed35ce1981e8)** instead.)
2. Install Java8:

   ```
   brew tap homebrew/cask-versions
   ```

   ```
   brew install --cask adoptopenjdk8
   ```

3. Install python3:

   ```
   brew install python3
   ```

4. Install Scala:

   ```
   brew install scala@2.12
   ```

5. Download Spark **https://mirror-hk.koddos.net/apache/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz (https://mirror-hk.koddos.net/apache/spark/spark-3.0.3/spark-3.0.3-bin-hadoop2.7.tgz)**
6. Install Spark:

   ```
   mkdir ~/hadoop/spark-3.0.3
   ```

   ```
   tar -xvzf ~/Downloads/spark-3.0.3-bin-hadoop2.7.tgz -C ~/hadoop/spark-3.0.3 --strip 1
   ```

7. Setup environment variables:

   ```
   vi ~/.bashrc
   ```

   ```
   export SPARK_HOME=~/hadoop/spark-3.0.3
   ```

   ```
   export PATH=$SPARK_HOME/bin:$PATH
   ```

   ```
   source ~/.bashrc
   ```

8. Start spark python shell:

   ```
   pyspark
   ```

VirtualBox

If you're not using Linux/*Unix systems, I suggest that you install a Linux virtual machine (using VirtualBox or VMware Player), and then install Spark on the VM. See instructions below on installing VirtualBox.

1. Download & Install:
   - Windows **https://download.virtualbox.org/virtualbox/5.2.6/VirtualBox-5.2.6-120293-Win.exe (https://download.virtualbox.org/virtualbox/5.2.6/VirtualBox-5.2.6-120293-Win.exe)**
   - Mac OS **https://download.virtualbox.org/virtualbox/5.2.6/VirtualBox-5.2.6-120293-OSX.dmg (https://download.virtualbox.org/virtualbox/5.2.6/VirtualBox-5.2.6-120293-OSX.dmg)**
2. Download Ubuntu Image:
   - **http://www.cse.ust.hk/msbd5003/ubuntu.ova (http://www.cse.ust.hk/msbd5003/ubuntu.ova)**
3. Install Ubuntu Image on VirtualBox
   - Open VirtualBox
   - Load Ubuntu Image into VirtualBox: File->Import Virtual Appliance
4. Run Ubuntu (Click "Start" button)
   - username/password: ubuntu/reverse

5. Install spark
    - Follow instructions in previous part.
6. You should enable the virtualization in BIOS, otherwise the virtual machine might be unable to start.

Jupyter installation (Linux & Mac OS):

1. Install Anaconda.
    1. Download the script:
        - Linux

          ```
          wget https://repo.anaconda.com/archive/Anaconda3-2020.07-Linux-x86_64.sh (https://repo.anaconda.com/archiv
          e/Anaconda3-2020.07-Linux-x86_64.sh)
          ```

        - Mac OS:

          ```
          wget https://repo.anaconda.com/archive/Anaconda3-2020.07-MacOSX-x86_64.sh (https://repo.anaconda.com/archi
          ve/Anaconda3-2020.07-MacOSX-x86_64.sh)
          ```

       Again, if you are in mainland, try **https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/Anaconda3-2020.07-Linux-x86_64.sh (https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/Anaconda3-2020.07-Linux-x86_64.sh)** and **https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/Anaconda3-2020.07-MacOSX-x86_64.sh (https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/Anaconda3-2020.07-MacOSX-x86_64.sh)**.
    2. Install:

       ```
       bash Anaconda3-2020.07-Linux-x86_64.sh
       ```

       (Use corresponding file)
    3. Modify pyspark driver: Add the following lines to '~/.bashrc' or '~/.bash_profile':
        - export PYSPARK_DRIVER_PYTHON="jupyter"
        - export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
    4. Update $PATH variable (the file you modifed in previous step)

       ```
       source ~/.bashrc
       ```

       ```
       source ~/.bash_profile
       ```

2. Start spark python shell (in the spark directory):

   ```
   pyspark
   ```

   You should get a link directing you to localhost:8888

Notes:

1. you can execute "unset PYSPARK_DRIVER_PYTHON PYSPARK_DRIVER_PYTHON_OPTS" to run normal pyspark shell
2. If you find this error "I couldn't find a kernel matching PySpark. Please select a kernel:" after you upload notebooks from lecture notes, you just choose Python3 kernel which already supports pyspark kernel.
3. After you finish the steps, create a new notebook, type "spark" and run it. If you see "**SparkSession - hive**" in the output, the installation should be successful.

GraphFrames:

- The latest version of GraphFrames requires numpy to run, to install numpy, run "pip3 install numpy"
- For pre-installed Spark version ubuntu, to use GraphFrames:
    1. get the jar file:

```
wget https://repos.spark-packages.org/graphframes/graphframes/0.8.1-spark3.0-s_2.12/graphframes-0.8.1-spark3.0-s_2.12.jar
```

2. Load the jar file in the Jupyter notebook

   ```
   sc.addPyFile('path_to_the_jar_file')
   ```

You can also refer to "~/Untitled.ipynb".

- Using the pyspark shell directly with GraphFrames:

  ```
  pyspark --packages graphframes:graphframes:0.8.1-spark3.0-s_2.12
  ```

- Demo Program

```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
v = sqlContext.createDataFrame([("a", ),("b", ),], ["id", ])
e = sqlContext.createDataFrame([("a", "b"),], ["src", "dst"])
from graphframes import *
g = GraphFrame(v, e)
g.inDegrees.show()
```

And the correct output will be:

```
    +---+--------+
    | id|inDegree|
    +---+--------+
    |  b|       1|
    +---+--------+
```