Running BigDL (Python) on Spark on your Local Machine

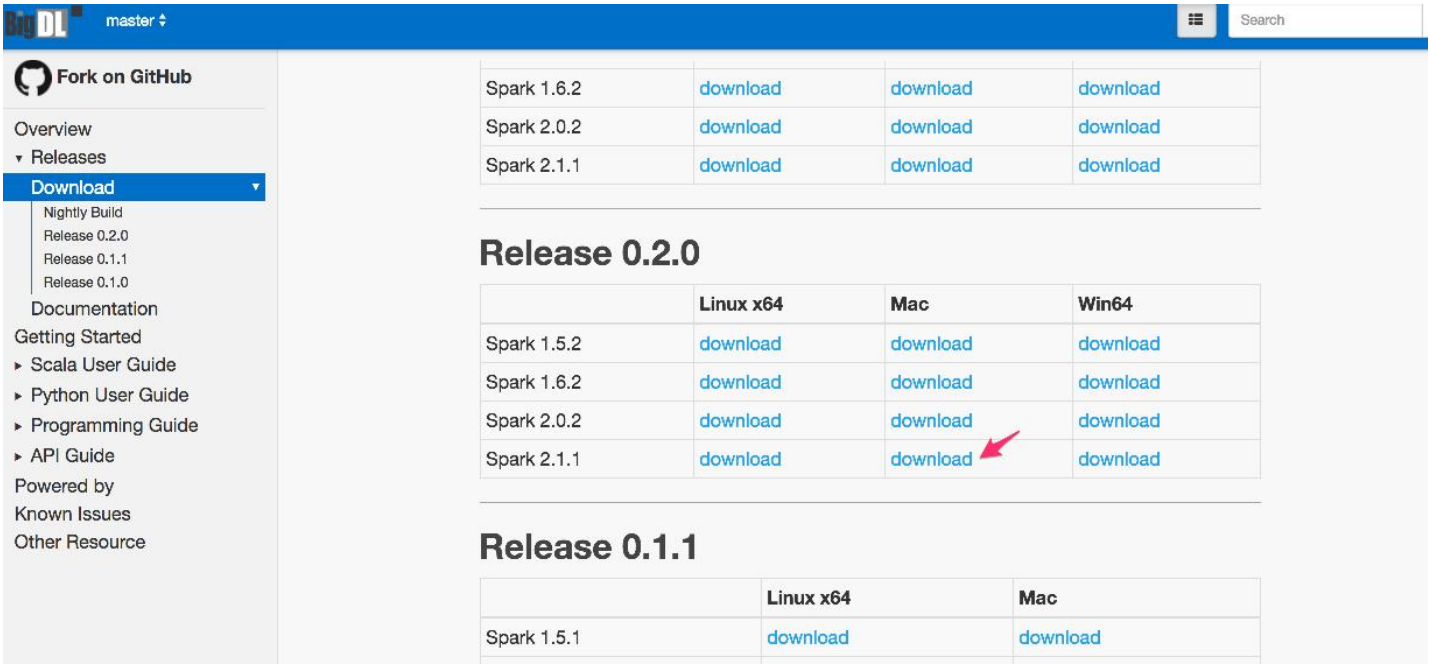
Updated Aug 31, 2017

This tutorial will teach you how to set up BigDL on your Mac and run the MNIST digit recognition BigDL Python application.

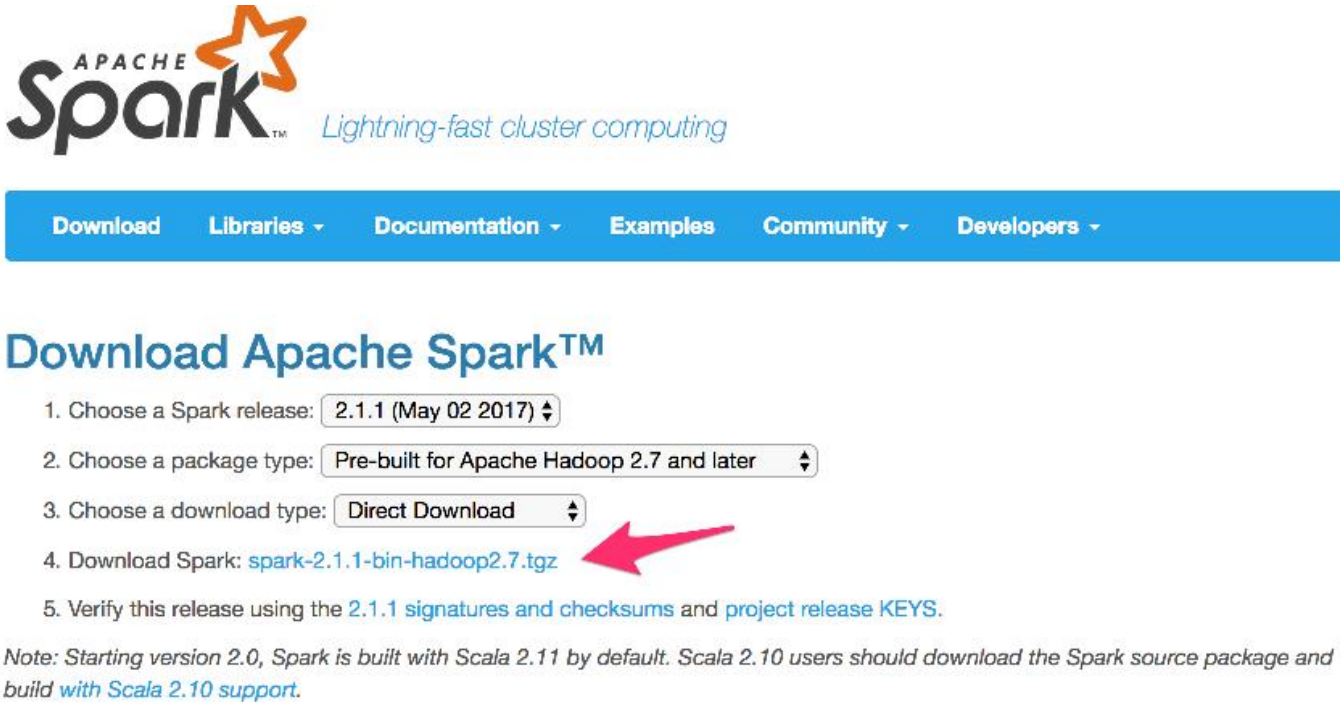
Prerequisites:
-Mac/Linux OS
-Java latest version (if you don't have it run: conda install java)
-Jupyter Notebook (your pyspark driver)
-Python 2.7
-scikit-learn, spicy, numpy, pandas, matplotlib (pip install these in your environment)

STEP 1: Set up your BigDL folder.
Create a new folder to store the necessary bigdl/spark requirements. We are going to put number of things in there:
1) BigDL 0.2.0 Package
2) Spark 2.1.1
3) start_notebook.sh
4) cnn.ipynb
5) utils.py

#1: Download the built BigDL Package under Release 0.2.0, for Spark 2.1.1 here: [https://bigd project.github.io/master/#release-download/](https://bigdl-project.github.io/master/#release-download/). There is a version for Mac and Linux. This ha all BigDL dependencies and python files too.



#2: Download Spark version 2.1.1 here: https://spark.apache.org/downloads.html This unzip into a folder with all spark related contents.



#3: Download this starter script. start_notebook.sh 912 bytes

Note: Remember to change **<path to folder you created>** after the **SPARK_HOME** and **BigDL_HOME** environment variables to the absolute path to your bigdl folder you just created. This is the script that you will run to start a spark session and use BigDL. It sets th correct parameters and environment for your spark session to run BigDL applications.

#4: Download cnn.ipynb here. cnn.ipynb 132.7 KB

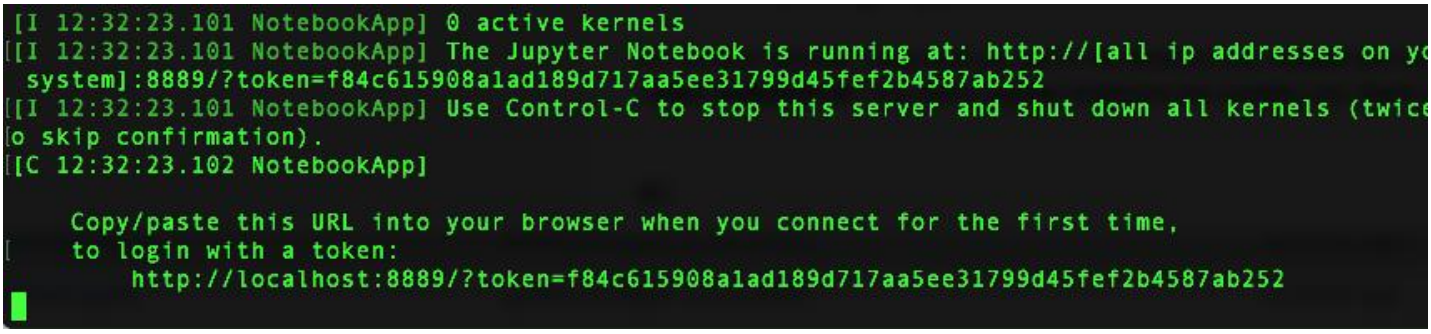
Note: There are more example notebooks in BigDL's tutorial repository. Feel free to clone and download these example notebooks here: https://github.com/intel-analytics/BigDL-Tutorials/tree/branch-0.2/notebooks/neural_networks

#5: Download utils.py here. utils.py 1.2 KB

#6 Verify that all items listed above are in your bigdl folder. You are ready to run your application now.

STEP 2: Run your BigDL application

#1: On your terminal, make sure you are in the directory you created and run the command **./start_notebook.sh**. This starts Jupyter Notebook and creates your Spark Context (sc).



#2: Copy and Paste the URL into your browser to start Jupyter Notebook.
Note: If you need to, change the access permission of start_notebook.sh script using this command: chmod +x <your path>/start_notebook.sh

#3: Run cnn.ipynb file in Jupyter Notebooks **using a kernel that runs Python 2.7**. It should match the output on the BigDL tutorials.
Note: If your default kernel in Jupyter Notebook is Python 3, you have to change it to Python 2. You can create a virtual environment that uses Python 2.7 by following this guid https://uoa-eresearch.github.io/eresearch-cookbook/recipe/2014/11/20/conda/
Note: You can only run **one** .ipynb at a time because you have only one spark session, remember to shutdown the rest. Also make sure your kernel is running Python 2 (Python 3 is still being tested).

Thats it!