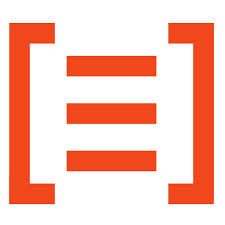
**Installation Guide** 

**Overview**:

In this lab, you will install all the necessary softwares for the following lab sessions. We provide three different programming environments (standalone, via docker, or cloude), feel free to chose your favorite.

Note: The standalone mode is only available for linux.

**What you need:**

* Internet access
* Access to a modern internet browser: Chrome/Firefox/Internet Explorer
* The lab files for this lab

**Standalone mode (Windows) :**

1. Install Java

Go to:

<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>

Download the right version of java and java jdk for your operating system and then install the java by following the instruction.

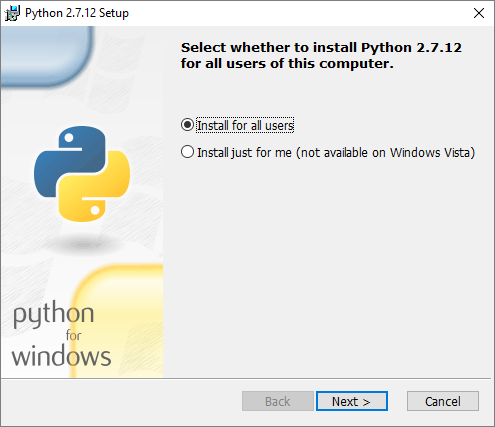
1. Open your web browser, and go to: <https://www.python.org/downloads/>
2. Download the python release for windows: Latest Python 2 Release - Python 2.7.12

Make sure to choose the right version for your operating system, else some of the compiled binaries and Python libraries don’t work well.

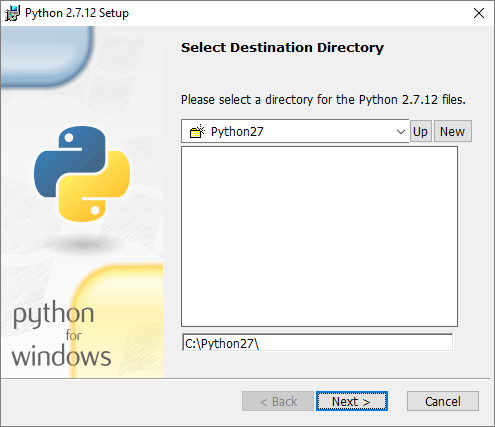
1. Installing Python:

Once you have downloaded the Python MSI, simply navigate to the download location on your computer, double clicking the file and pressing Run when the dialog box pops up.

If you are the only person who uses your computer, simply leave the “Install for all users” option selected. If you have multiple accounts on your PC and don’t want to install it across all accounts, select the “Install just for me” option then press “Next.”



If you want to change the install location, feel free to do so; however, it is best to leave it as is and simply select next.



Scroll down in the window and find the “Add Python.exe to Path” and click on the small red “x.” Choose the “Will be installed on local hard drive” option then press “Next.”

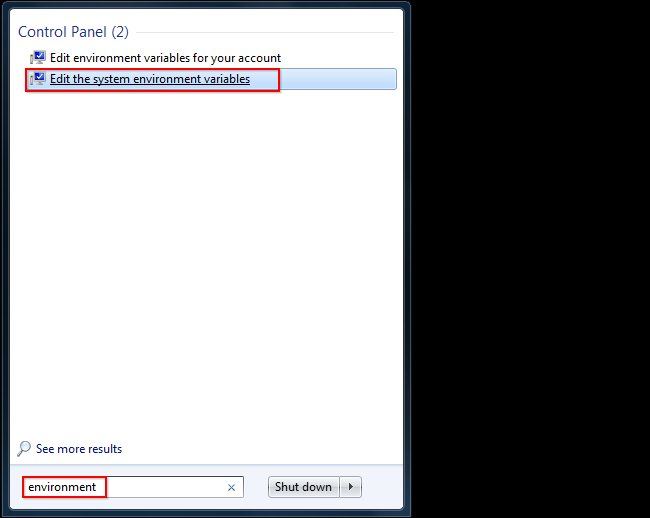


You will notice that the installation will bring up a command prompt window while Python downloads and installs “Pip.” Pip is just a package management tool. This will allow you to install all the additional Python packages that are available for download through [PyPI (Python Package Index)](https://pypi.python.org/pypi).

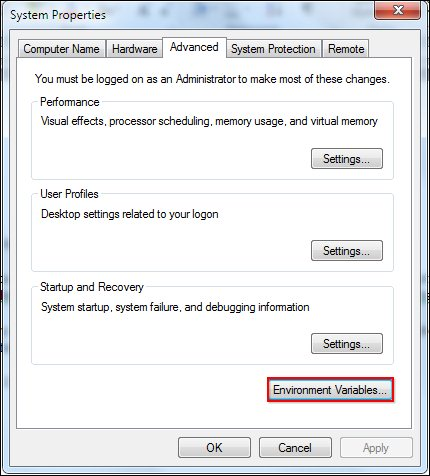
Now that you have completed the installation process, click on “Finish.”

1. Adding Python to System Path Variable

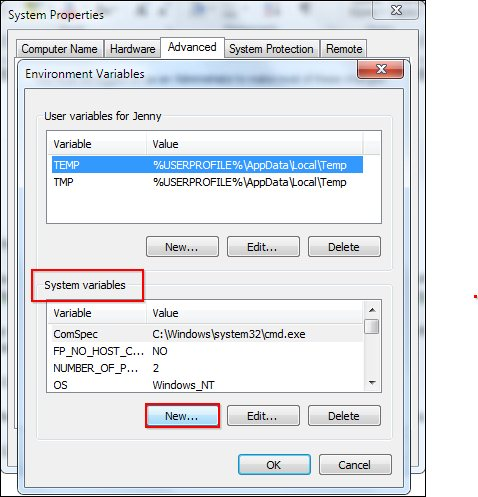
Begin by opening the start menu and typing in “environment” and select the option called “Edit the system environment variables.”



When the “System Properties” window appears, click on “Environment Variables…”

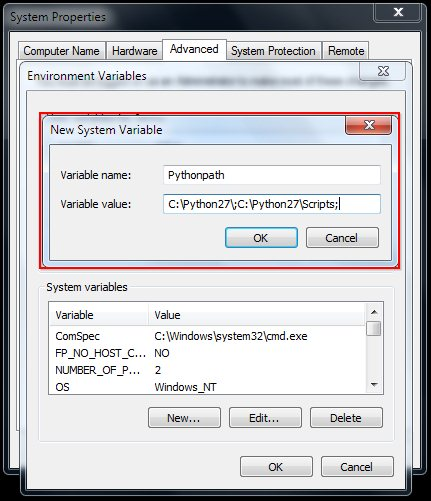


Once you have the “Environment Variables” window open, direct your focus to the bottom half. You will notice that it controls all the “System Variables” rather than just this associated with your user. Click on “New…” to create a new variable for Python.



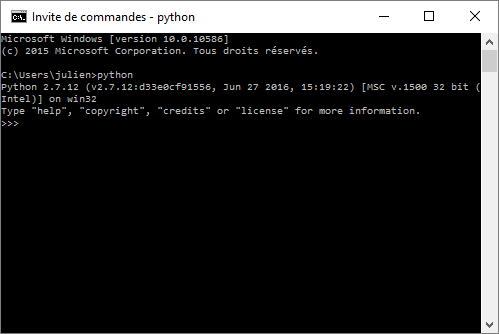
Simply enter a name for your Path and the code shown below. For the purposes of this example we have installed Python 2.7.12, so we will call the path: “Pythonpath.”

The string that you will need to enter is: “C:\Python27\;C:\Python27\Scripts;”



Press “OK,” then “OK,” then “OK,” then the red “X” to accept all changes and exit the “System Properties” window.

In the command line, tape “python”. If the installation is correct, the following figure should appear. Tape “quit()” to exit.



1. Open console (“cmd” in application), run the following commands:

python -m pip install --upgrade pip

python -m pip install jupyter

python -m pip install scikit-learn

python -m pip install matplotlib

python -m pip install pandas

python -m pip install numpy

1. Install H2o directly from Python: <http://www.h2o.ai/download/h2o/python>

python -m pip install requests

python -m pip install tabulate

python -m pip install future

python -m pip install six

python -m pip install http://h2o-release.s3.amazonaws.com/h2o/rel-turing/3/Python/h2o-3.10.0.3-py2.py3-none-any.whl

1. Launch a Jupyter Notebook in the command line, tape:

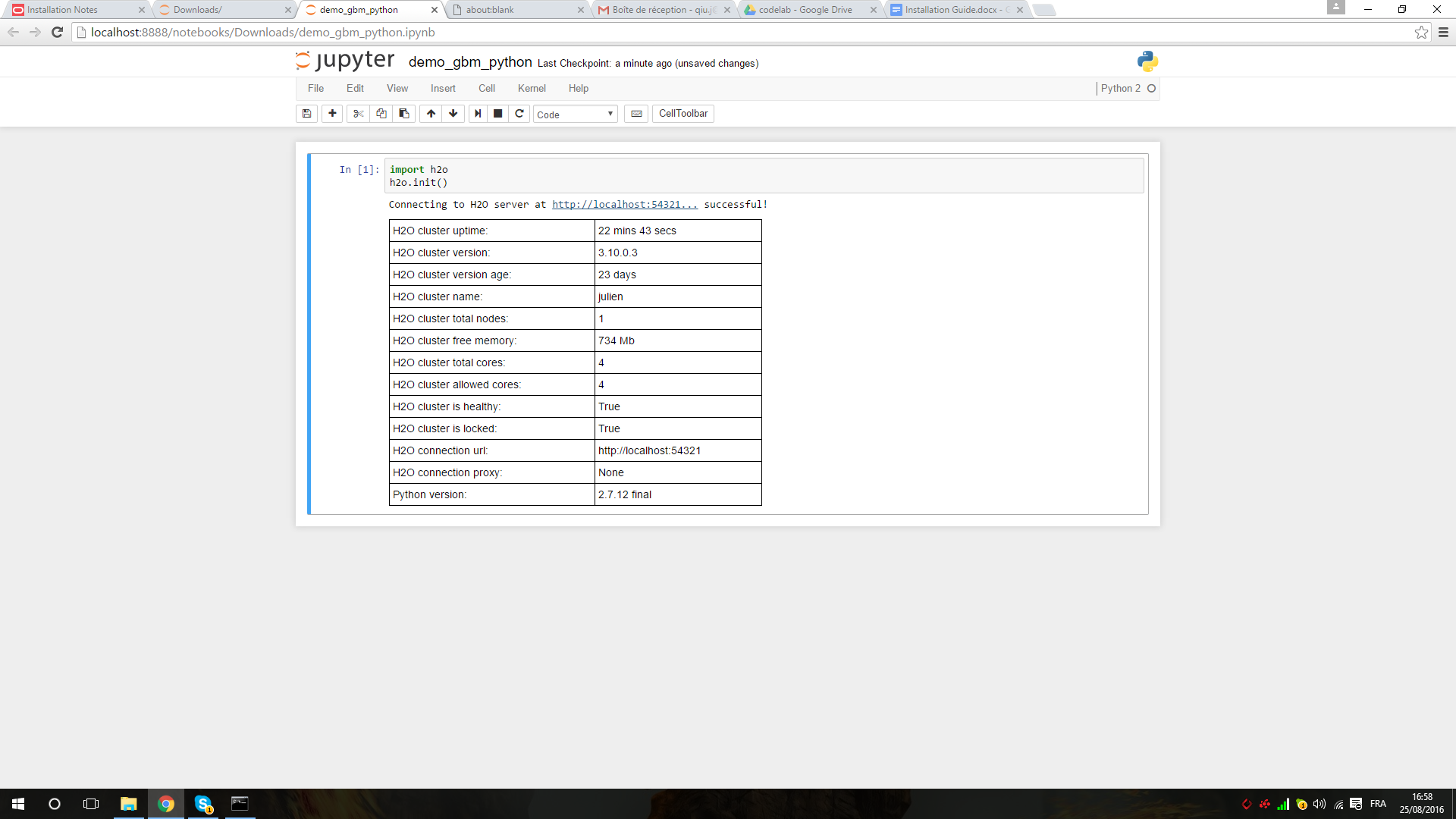
jupyter notebook

And open a new jupyter notebook, the tape:

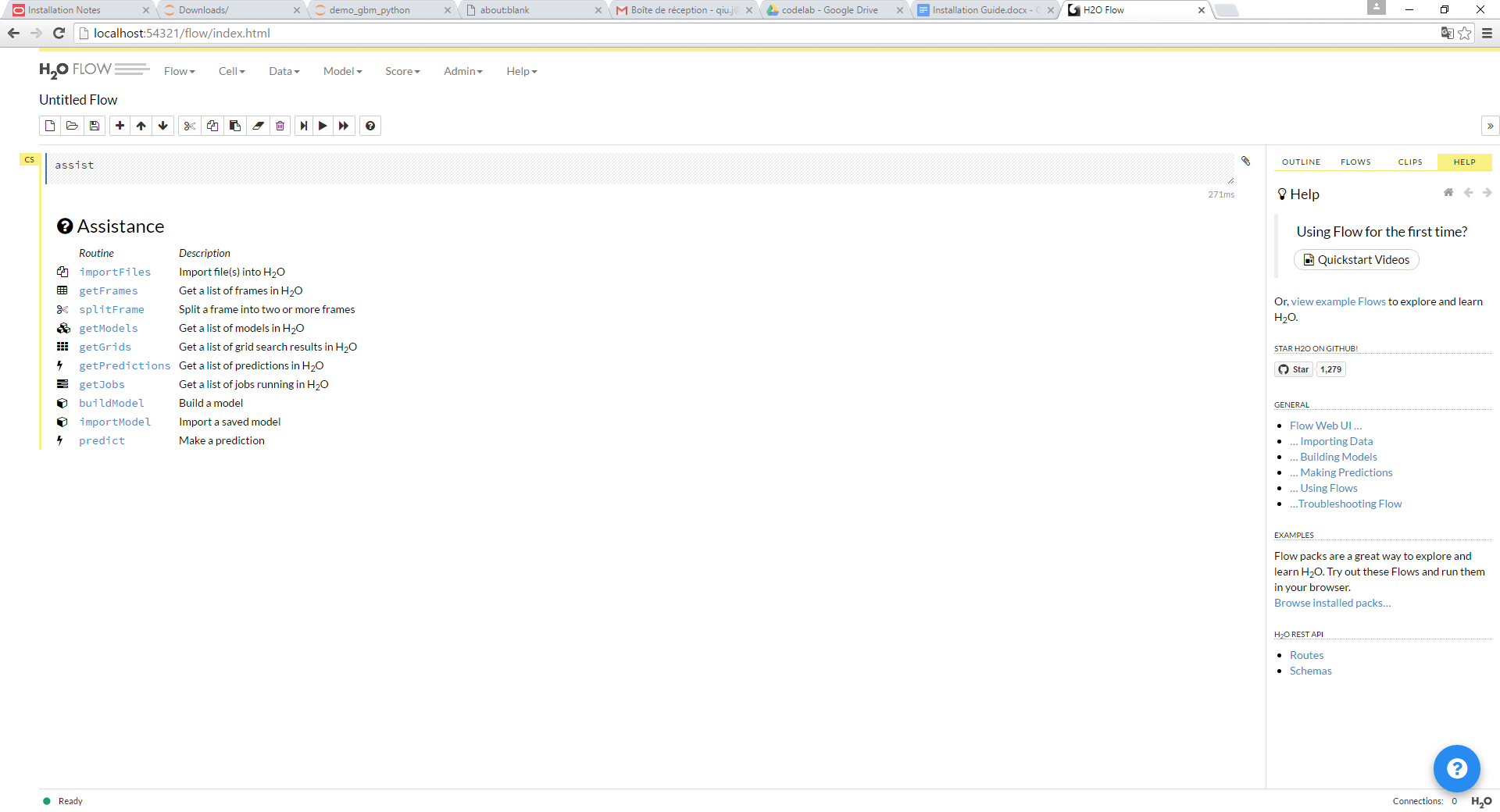
import h2o

h2o.init()

If the following figure is observed, then the installation is correct.



Open a new tab in the web browser and go to : <http://localhost:54321/>



**Standalone mode (Ubuntu after 14.04 or Mac) :**

1. Open console, run the following commands (prepending with ‘sudo’ if needed):

[sudo] apt-get update

[sudo] apt-get install python-dev python-pip

[sudo] pip install jupyter

[sudo] pip install scikit-learn

1. Install H2o directly from Python

[sudo] pip install -U requests

[sudo] pip install -U tabulate

[sudo] pip install -U future

[sudo] pip install -U six

[sudo] pip install matplotlib

[sudo] pip install pandas

[sudo] pip install numpy

[sudo] pip install scipy

[sudo] pip install test\_helper

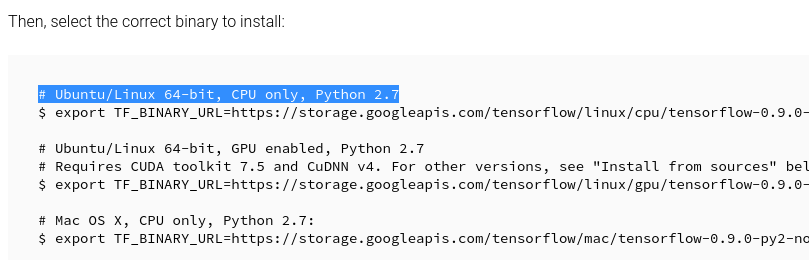
At the command line, copy and paste these commands one line at a time:

[sudo] pip install pip install http://h2o-release.s3.amazonaws.com/h2o/rel-turing/3/Python/h2o-3.10.0.3-py2.py3-none-any.whl

1. Install TensorFlow from [https://www.tensorflow.org](https://www.tensorflow.org/)

For the details installation, use the following link: <https://www.tensorflow.org/versions/r0.9/get_started/os_setup.html#pip-installation>

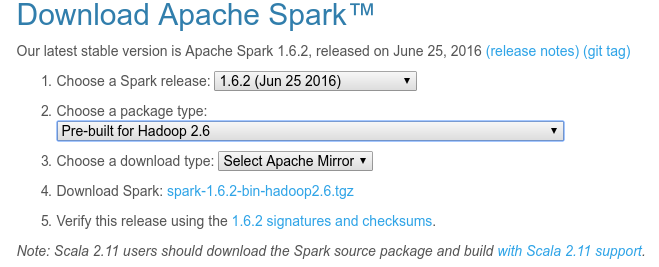
For this lab, please install the CPU version for ubuntu/linux 64-bits and python 2.7.



1. Download Spark from: <http://spark.apache.org/downloads.html>

Choose the following configuration:

* Spark release:1.6.2
* Package type: Pre-build for Hadoop 2.6



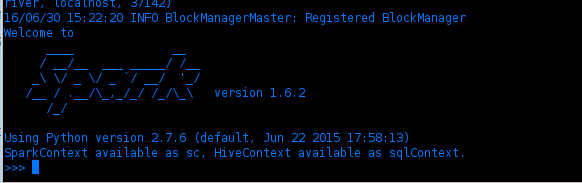
Point SPARK\_HOME to the existing installation of Spark in the ~/.bashrc by adding:

export SPARK\_HOME=/path/to/spark/installation

It should look like: export SPARK\_HOME="/\*\*\*/spark-1.6.2-bin-hadoop2.6"

To test the installation of spark, tape the following command in the console:

$SPARK\_HOME/bin/pyspark



You can also launch a Jupyter Notebook that connects to PySpark:

IPYTHON\_OPTS="notebook" $SPARK\_HOME/bin/pyspark

1. Download Sparkling Water from: <http://www.h2o.ai/download/sparkling-water/spark16>

Point SPARKLING\_HOME to the existing installation of Sparkling in the ~/.bashrc by adding:

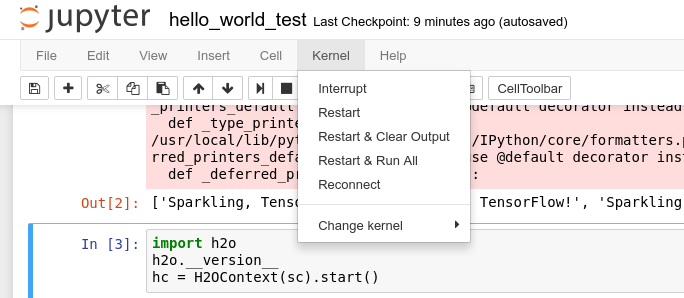
export SPARKLING\_HOME=/path/to/spark/installation

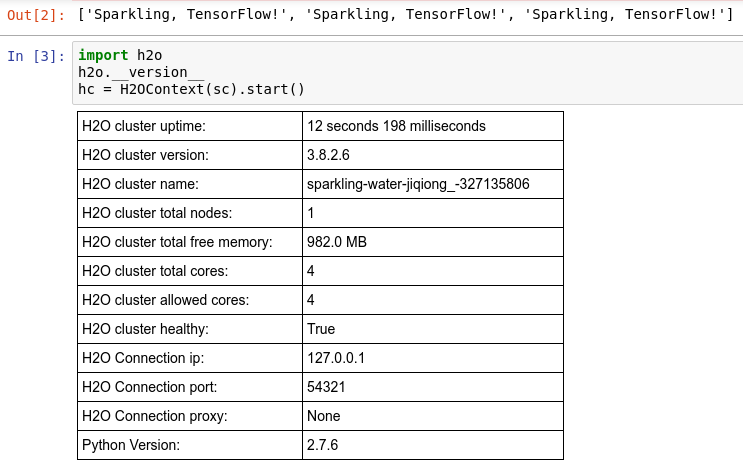
It should look like: export SPARKLING\_HOME="/\*\*/sparkling-water-1.6.5"

1. Launch a Jupyter Notebook that connects to PySparkling:

IPYTHON\_OPTS="notebook" $SPARKLING\_HOME/bin/pysparkling

1. Upload the “hello\_world\_test.ipynb” in the jupyter notebook, then go to Kernel -> Restart & Run all:



If the installation is correct, you will see:

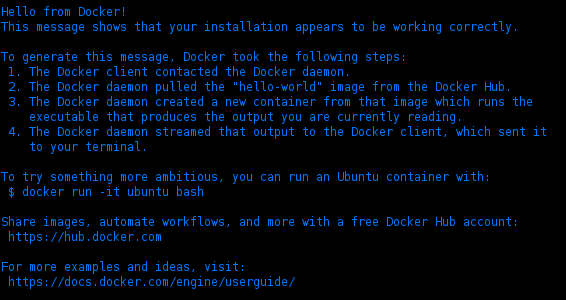
**via Docker:**

1. Go to the docker homepage and install docker: <https://www.docker.com/>
2. Launch docker and tape:

docker pull hello-world

docker run hello-world

If the installation is correct, we should obtain the following image:

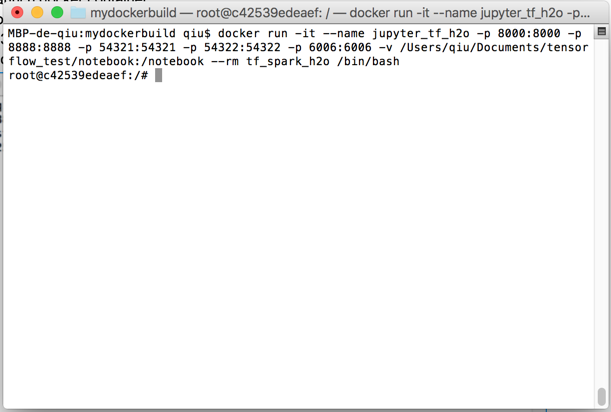


1. Download the Dockerfile, put it into a folder called mydockerbuild and go to mydockerbuild floder
2. Launch docker-build and create the docker image by tapping in the console:

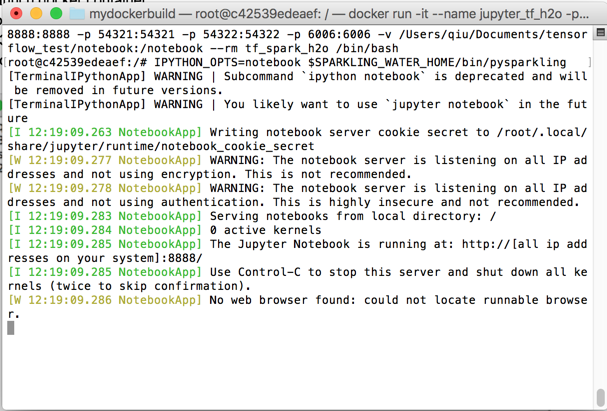
docker build -t name\_docker\_image .

1. Launch docker container:

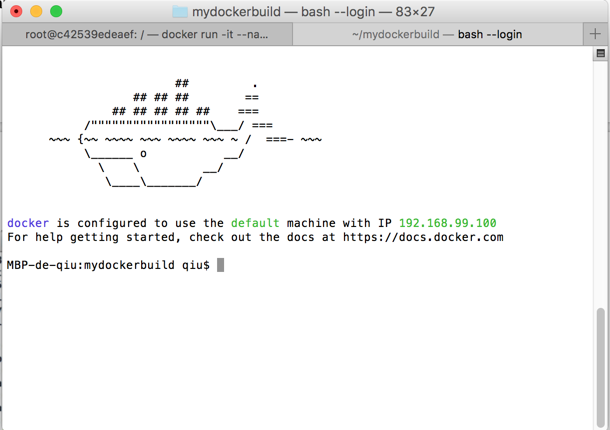
docker run -it --name name\_of\_your\_container -p 8000:8000 -p 8888:8888 -p 54321:54321 -p 54322:54322 -p 6006:6006 -v /path\_to\_your\_notebooke\_location/notebook:/notebook --rm name\_docker\_image /bin/bash



1. Launch sparkling in notebook mode:

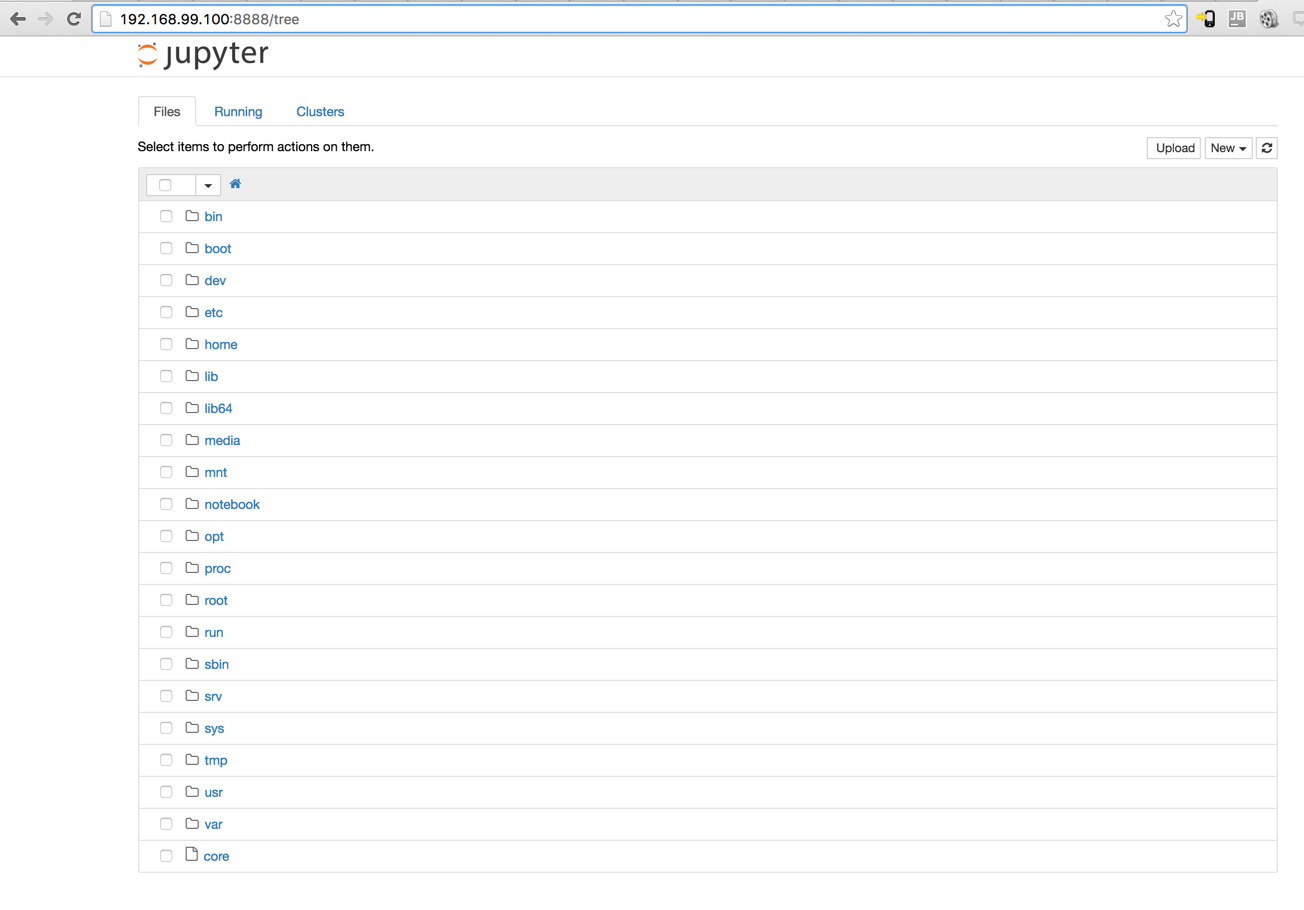


1. Open another docker console:

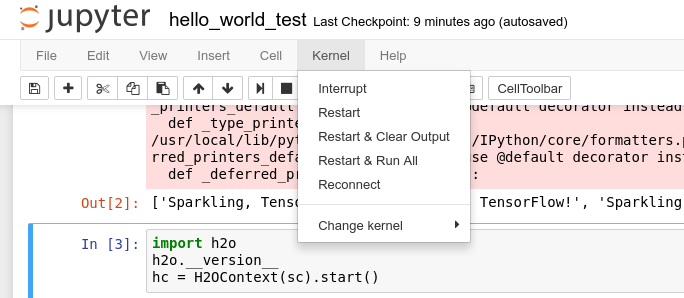


The 192.168.99.100 is the ip of your docker container.

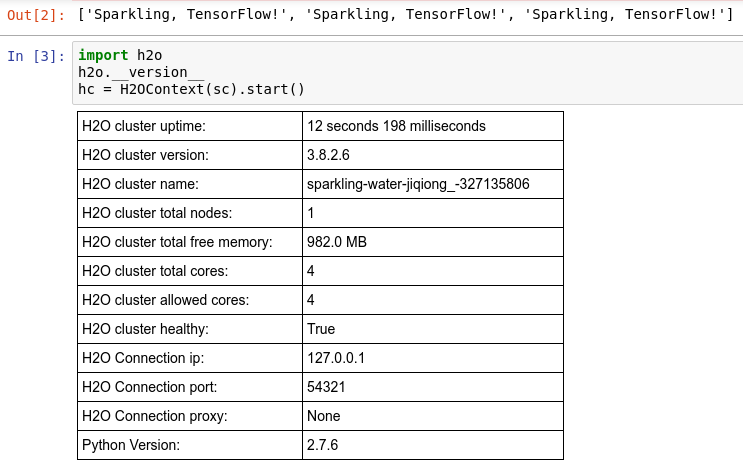
1. Open the bowser, and entrer yourip:8888. In the showing case, it should be: 192.168.99.100:8888



1. Go to the notebook folder. Upload the “hello\_world\_test.ipynb” in the jupyter notebook, then go to Kernel -> Restart & Run all:



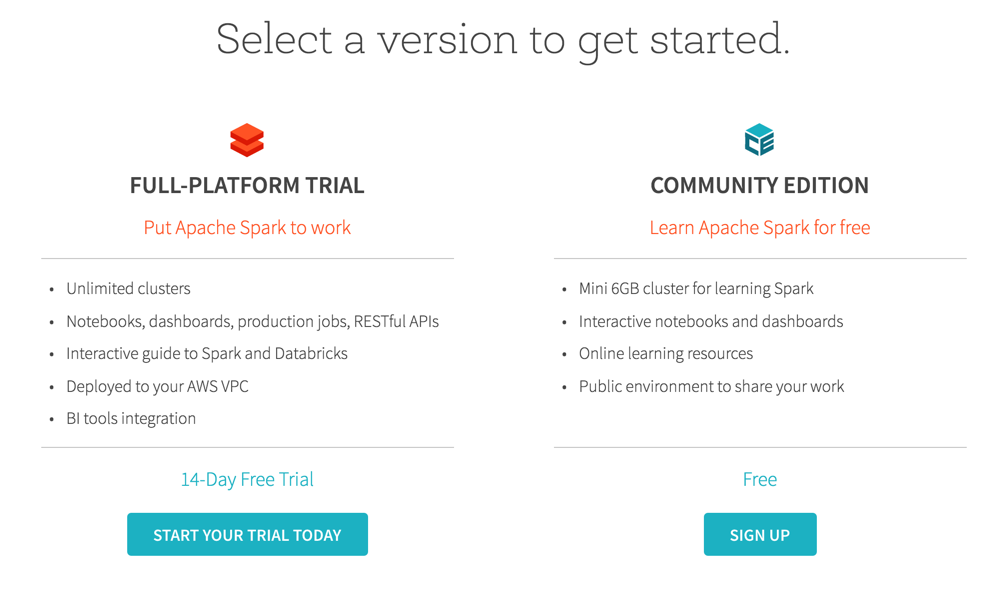
If the installation is correct, you will see:



**Use Spark in Databricks:**

In this course, you can also use [Databricks Community Edition](https://databricks.com/ce), a free online Spark software development environment. If you do not already have an account, you can register for one at this link: <https://databricks.com/ce>.

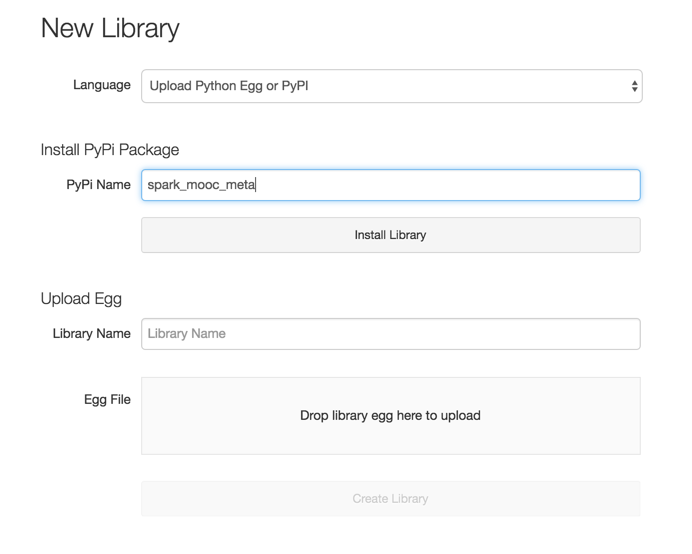
Note: You should **not** sign up for a "14-Day Free Trial" because that is a paid service. If you end up on a page that asks for your credit card information, then you have used the wrong registration link. [*Databricks Community Edition*](https://databricks.com/ce) *is a completely free service.*



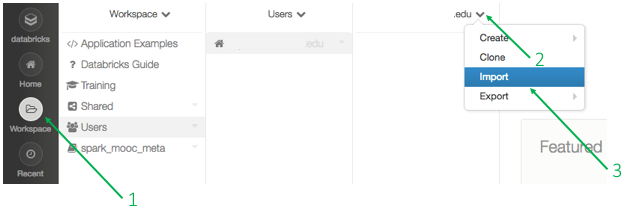
Login to your [Databricks Community Edition account](https://community.cloud.databricks.com/).

Import library by selecting “Workspace”, clicking on the dropdown next to your username, and selecting “Create” ->”Library”.

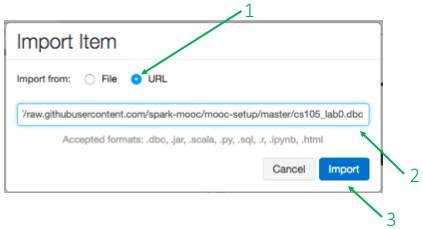
Then select “Upload Python Egg or PyPI” with the name: “spark\_mooc\_meta”, and click “Install Library”.



Import your Notebook by selecting "Workspace", clicking on the dropdown next to your username, and selecting "Import".



Select the URL radio button, copy the address of the your notebook , paste the copied link into the textbox, and click "Import" to import the notebook. You can also click “File” and select the notebook on your computer.

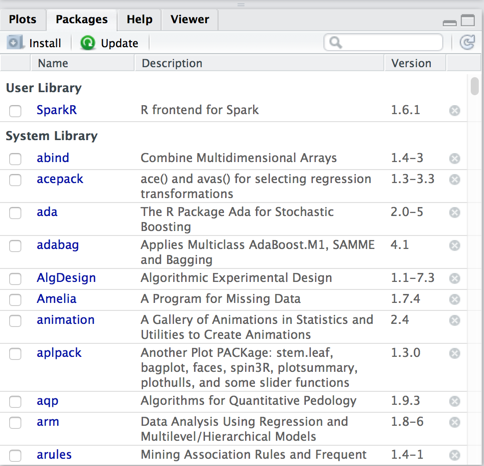


**Use SparkR in Rstudio:**

Open Rstudio, and run the following script:

Sys.setenv(SPARK\_HOME="/Path\_to\_your\_spark\_installation/spark-1.6.1-bin-hadoop2.6")

.libPaths(c(file.path(Sys.getenv("SPARK\_HOME"),"R","lib"),.libPaths()))



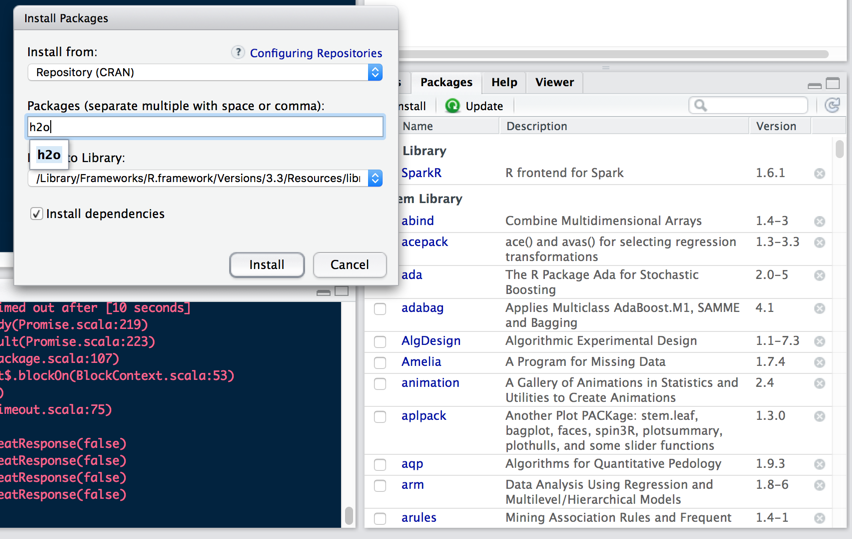
If it is corrected loaded, you should see “SparkR” in your user Library. You can test a GLM algorithm by using the provided R file in the Lab Spark.

**Install H2o/H2o ensembl in R:**

Install the dependencies for H2O.



Open Rstudio, go to “Package” -> “Install”:



And also install devtools.

To install h2oensemble:

