**Spark – Python**

* Why I choose ipython notebook (now it is Jupyter)?

You will be able to explore any dataset visually using matplotlib/plotly modules after importing them.

Good way to document things while working on any dataset in Spark.

**Linking Spark to Ipython**

* Need to make sure Spark and Python are pre-installed.
* Install jupyter package.

pip install jupyter

If you get error related to pip, that it is not installed. We need to install pip and then install python packages.

yum install –y python-pip

* In the terminal execute the below three commands.

echo "export PATH=$PATH:<path to spark install directory>/bin" >> .profile

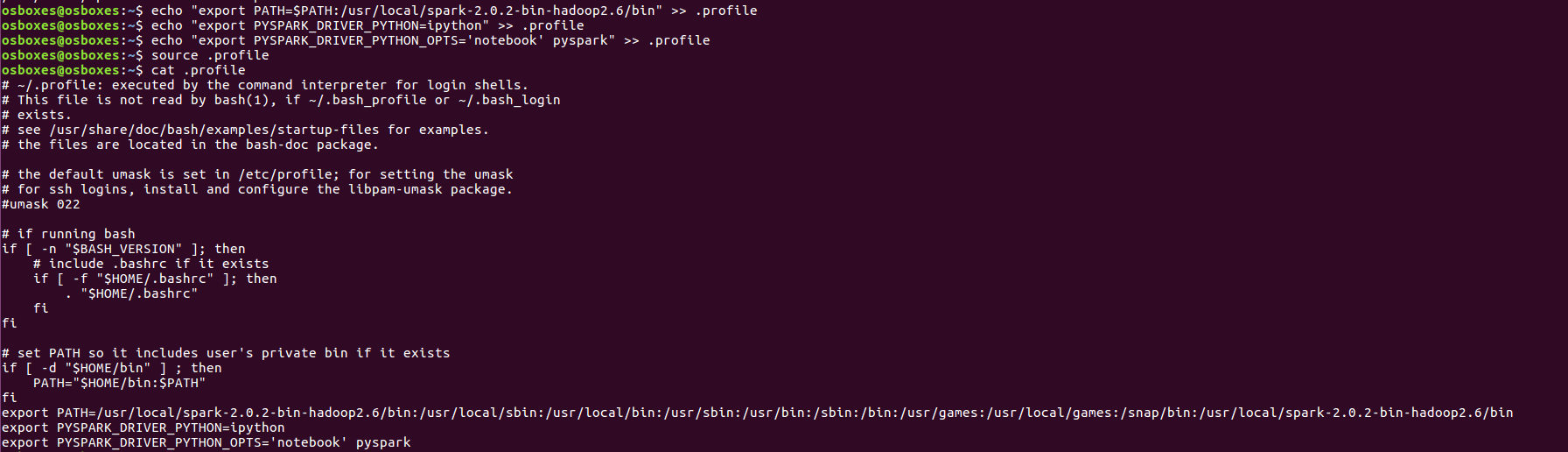
echo "export PYSPARK\_DRIVER\_PYTHON=ipython" >> .profile

echo "export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook' pyspark" >> .profile

source .profile #which saves the configuration file.

sudo pip install -U "ipython[notebook]" #for installing ipython package, need to use sudo

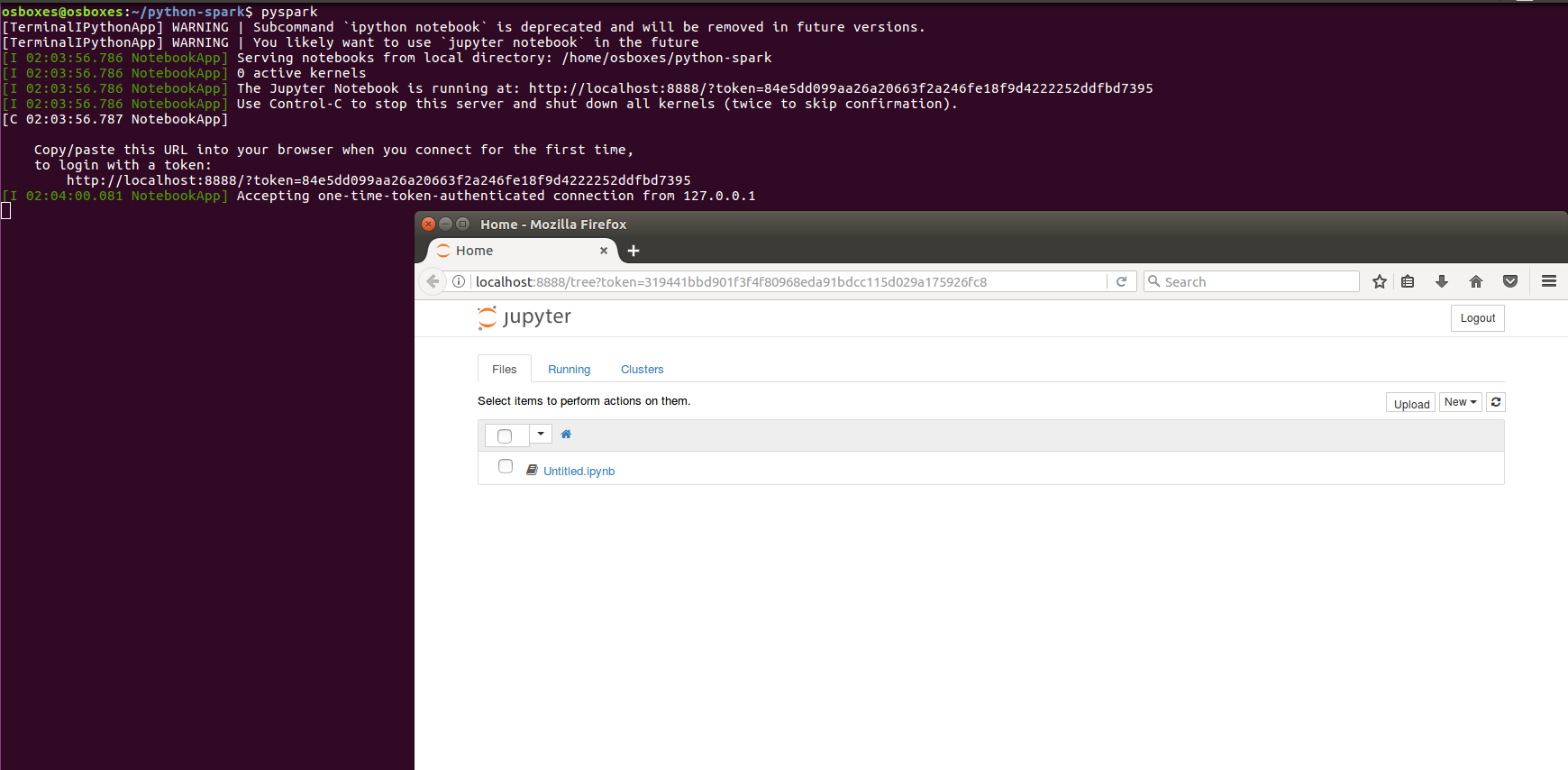
admin rights.



* Create a directory where you want to save the python notebook files. For example, as below:



* Then execute pyspark, it will open the jupyter notebook. We can write our pyspark programs.



**MLLib using Python on HVAC data and Human Activity Recognition Using Smartphones Data Set - UCI Data repository**

* When running machine-learning algorithms in Spark using Python, you need to be very careful about the errors. You need to refer to the documents.
* I have run MLlib on the abovie mentioned datasets. I came across many issues and resolved accordingly referring to the Spark documentation and googling for the respective error.
* Please refer the ipython notebooks, shared along with this document about the implementation.