Using IPython Notebook with Apache Spark

In this tutorial we are going to configure IPython notebook with Apache Spark on YARN in a few steps.

IPython notebook is an interactive Python shell which lets you interact with your data one step at a time and also perform simple visualizations.

IPython notebook supports tab autocompletion on class names, functions, methods, variables. It also offers more explicit and colour-highlighted error messages than the command line python shell. It provides integration with basic UNIX shell allowing you can run simple shell commands such as cp, ls, rm, cp, etc. directly from the IPython. IPython integrates with many common GUI modules like PyQt, PyGTK, tkinter as well wide variety of data science Python packages.

Prerequisites

The only prerequiste for this tutorial is the latest Hortonworks Sandbox installed on your computer or on the cloud. In case you are running an Hortonworks Sandbox with an earlier version of Apache Spark, for the instruction in this tutorial, you need to install the Apache Spark 1.3.1.

Installing and configuring IPython

To begin, login in to Hortonworks Sandbox through SSH:The default password is `hadoop`.

```
1. root@sandbox:~ (ssh)
Last login: Sun Apr 12 17:56:45 on ttys001
Saptaks-MacBook-Pro:~ saptak$ ssh root@127.0.0.1 -p 2222;
root@127.0.0.1's password:
Last login: Fri Apr 10 01:42:11 2015 from 10.0.2.2
[root@sandbox ~]#
```

Now let's configure the dependencies by typing in the following command:

```
yum install nano centos-release-SCL zlib-devel \
bzip2-devel openssl-devel ncurses-devel \
sqlite-devel readline-devel tk-devel \
gdbm-devel db4-devel libpcap-devel xz-devel \
libpng-devel libjpg-devel atlas-devel
```

IPython has a requirement for Python 2.7 or higher. So, let's install the "Development tools"

```
1. root@sandbox:~ (ssh)

Dependency Installed:
  atlas.x86_64 0:3.8.4-2.el6
  fontconfig-devel.x86_64 0:2.8.0-5.el6
  freetype-devel.x86_64 0:2.3.11-15.el6_6.1
  keyutils-libs-devel.x86_64 0:1.4-5.el6
  krb5-devel.x86_64 0:1.10.3-37.el6_6
  libX11-devel.x86_64 0:1.6.0-2.2.el6
  libXau-devel.x86_64 0:1.0.6-4.el6
  libXft-devel.x86_64 0:2.3.1-2.el6
  libXrender-devel.x86_64 0:0.9.8-2.1.el6
  libcom_err-devel.x86_64 0:1.41.12-21.el6
  libgfortran.x86_64 0:4.4.7-11.el6
  libpcap.x86_64 14:1.4.0-1.20130826git2dbcaa1.el6
  libselinux-devel.x86_64 0:2.0.94-5.8.el6
  libsepol-devel.x86_64 0:2.0.41-4.el6
  libxcb-devel.x86_64 0:1.9.1-2.el6
  tcl.x86_64 1:8.5.7-6.el6
  tcl-devel.x86_64 1:8.5.7-6.el6
  tk.x86_64 1:8.5.7-5.el6
  xorg-x11-proto-devel.noarch 0:7.7-9.el6

Complete!
[root@sandbox ~]#
```

dependency for Python 2.7

yum groupinstall “Development tools“

Now we are ready to install Python 2.7.

yum install python27

Now the Sandbox has multiple versions of Python, so we have to select which version of Python we want to use in this session. We will choose to use Python 2.7 in this session.

source /opt/rh/python27/enable

Then we will download `easy_install` which we will use to configure `pip`, a Python package installer.

wget https://bitbucket.org/pypa/setuptools\
/raw/bootstrap/ez_setup.py

Now let’s configure `easy_install` with the following command:

python ez_setup.py

Now we can install `pip` with `easy_install` using the following command:

```
1. root@sandbox:~ (ssh)

Dependency Installed:
  atlas.x86_64 0:3.8.4-2.el6
  fontconfig-devel.x86_64 0:2.8.0-5.el6
  freetype-devel.x86_64 0:2.3.11-15.el6_6.1
  keyutils-libs-devel.x86_64 0:1.4-5.el6
  krb5-devel.x86_64 0:1.10.3-37.el6_6
  libX11-devel.x86_64 0:1.6.0-2.2.el6
  libXau-devel.x86_64 0:1.0.6-4.el6
  libXft-devel.x86_64 0:2.3.1-2.el6
  libXrender-devel.x86_64 0:0.9.8-2.1.el6
  libcom_err-devel.x86_64 0:1.41.12-21.el6
  libgfortran.x86_64 0:4.4.7-11.el6
  libpcap.x86_64 14:1.4.0-1.20130826git2dbcaa1.el6
  libselinux-devel.x86_64 0:2.0.94-5.8.el6
  libsepol-devel.x86_64 0:2.0.41-4.el6
  libxcb-devel.x86_64 0:1.9.1-2.el6
  tcl.x86_64 1:8.5.7-6.el6
  tcl-devel.x86_64 1:8.5.7-6.el6
  tk.x86_64 1:8.5.7-5.el6
  xorg-x11-proto-devel.noarch 0:7.7-9.el6

Complete!
[root@sandbox ~]#
```

easy_install-2.7 pip

`pip` makes it really easy to install the Python packages. We will use `pip` to install the data science packages we might need using the following command:

pip install numpy scipy pandas \
scikit-learn tornado pyzmq \
pygments matplotlib jinja2 jsonschema

Finally, we are ready to install IPython notebook using `pip` using the following command:

pip install “ipython[notebook]“

Configuring IPython

Since we want to use IPython with Apache Spark we have to use the Python interpreter which is built with Apache Spark, `pyspark`, instead of the default Python interpreter.

As a first step of configuring that, let’s create a IPython profile for `pyspark`

ipython profile create pyspark

Within the this newly minted IPython profile for `pyspark` found at `~/.ipython/profile_pyspark/`, edit the file `ipython_notebook_config.py` with text editor like `nano` and change the values in the file to resemble values below:

```
1. root@sandbox:~ (ssh)
Installed:
  python27.x86_64 0:1.1-16.el6.centos.alt

Dependency Installed:
  python27-python.x86_64 0:2.7.5-10.el6.centos.alt
  python27-python-babel.noarch 0:0.9.6-7.el6.centos.alt
  python27-python-devel.x86_64 0:2.7.5-10.el6.centos.alt
  python27-python-docutils.noarch 0:0.11-1.el6.centos.alt
  python27-python-jinja2.noarch 0:2.6-10.el6.centos.alt
  python27-python-libs.x86_64 0:2.7.5-10.el6.centos.alt
  python27-python-markupsafe.x86_64 0:0.11-11.el6.centos.alt
  python27-python-nose.noarch 0:1.3.0-1.el6.centos.alt
  python27-python-pygments.noarch 0:1.5-2.el6.centos.alt
  python27-python-setuptools.noarch 0:0.9.8-2.el6.centos.alt
  python27-python-simplejson.x86_64 0:3.2.0-1.el6.centos.alt
  python27-python-sphinx.noarch 0:1.1.3-7.el6.centos.alt
  python27-python-sqlalchemy.x86_64 0:0.7.9-3.el6.centos.alt
  python27-python-virtualenv.noarch 0:1.10.1-2.el6.centos.alt
  python27-python-werkzeug.noarch 0:0.8.3-5.el6.centos.alt
  python27-runtime.x86_64 0:1.1-16.el6.centos.alt
  scl-utils.x86_64 0:20120927-27.el6_6

Complete!
[root@sandbox ~]#
```

c.NotebookApp.ip = '*'
c.NotebookApp.open_browser = False
c.NotebookApp.port = 8889
c.NotebookApp.notebook_dir = u'/usr/hdp/current/spark-client/'

Note the port we are using for IPython. Ensure this port is not already being used. The default port for IPython notebook is `8888`, which is also being used by Sandbox as it's welcome page. So we are changing it to `8889`. We are going to forward this port in the next section to ensure we can access IPython notebook from the host machine.

Next we are going to create a shell script to set the appropriate values everytime we want to start IPython.

Create a shell script with the following command:

nano ~/start_ipython_notebook.sh

Then copy the following lines into the file:

#!/bin/bash
source /opt/rh/python27/enable
IPYTHON_OPTS="notebook –profile pyspark" pyspark

Finally we need to make the shell script we just created executable:

chmod +x start_ipython_notebook.sh

```
1. root@sandbox:~ (ssh)
warning: no previously-included files found matching '.mailmap'
warning: no previously-included files found matching '.travis.yml'
warning: no previously-included files found matching 'pip/_vendor/Makefile'
warning: no previously-included files found matching 'tox.ini'
warning: no previously-included files found matching 'dev-requirements.txt'
no previously-included directories found matching '.travis'
no previously-included directories found matching 'docs/_build'
no previously-included directories found matching 'contrib'
no previously-included directories found matching 'tasks'
no previously-included directories found matching 'tests'
creating /opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.0-py2.7.e
gg
Extracting pip-7.1.0-py2.7.egg to /opt/rh/python27/root/usr/lib/python2.7/site-p
ackages
Adding pip 7.1.0 to easy-install.pth file
Installing pip script to /opt/rh/python27/root/usr/bin
Installing pip2.7 script to /opt/rh/python27/root/usr/bin
Installing pip2 script to /opt/rh/python27/root/usr/bin

Installed /opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.0-py2.7.
egg
Processing dependencies for pip
Finished processing dependencies for pip
[root@sandbox ~]#
```

Port Forwarding

We need to forward the port `8889` from the guest VM, Sandbox to the host machine, your desktop for IPython notebook to be accessible from a browser on your host machine.

Open the VirtualBox App and open the settings page of the Sandbox VM by right clicking on the Sandbox VM and selecting settings.

Then select the networking tab from the top:

Then click on the port forwarding button to configure the port. Add a new port configuration by clicking the `+` icon on the top right of the page.

Input a name for application, IP and the guest and host ports as per the screenshot below:

Then press `OK` to confirm the change in configuration.

Now we are ready to test IPython notebook.

Running IPython notebook

Execute the shell script we created before from the sandbox command prompt using the command below:

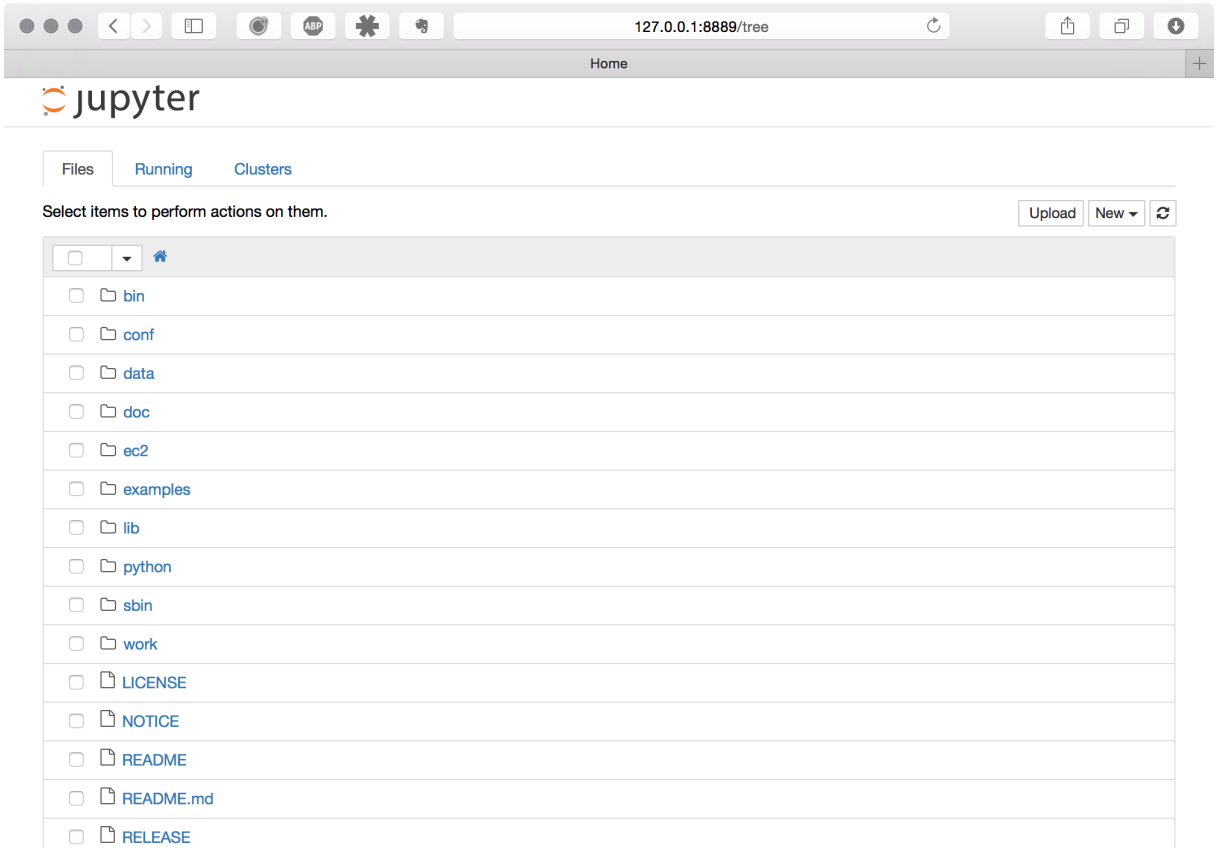
./start_ipython_notebook.sh

```
1. root@sandbox:~ (ssh)
  Downloading functools32-3.2.3-2.tar.gz
Collecting funcsigs (from mock->matplotlib)
  Downloading funcsigs-0.4-py2.py3-none-any.whl
Collecting pbr>=0.11 (from mock->matplotlib)
  Downloading pbr-1.3.0-py2.py3-none-any.whl (83kB)
    100% |████████████████████████████████| 86kB 631kB/s
Installing collected packages: numpy, scipy, six, python-dateutil, pytz, pandas,
 scikit-learn, backports.ssl-match-hostname, certifi, tornado, pyzmq, pyparsing,
 funcsigs, pbr, mock, matplotlib, functools32, jsonschema
  Running setup.py install for numpy
  Running setup.py install for scipy
  Running setup.py install for pandas
  Running setup.py install for scikit-learn
  Running setup.py install for backports.ssl-match-hostname
  Running setup.py install for tornado
  Running setup.py install for pyzmq
  Running setup.py install for matplotlib
  Running setup.py install for functools32
Successfully installed backports.ssl-match-hostname-3.4.0.2 certifi-2015.4.28 fu
ncsigs-0.4 functools32-3.2.3.post2 jsonschema-2.5.1 matplotlib-1.4.3 mock-1.2.0
numpy-1.9.2 pandas-0.16.2 pbr-1.3.0 pyparsing-2.0.3 python-dateutil-2.4.2 pytz-2
015.4 pyzmq-14.7.0 scikit-learn-0.16.1 scipy-0.15.1 six-1.9.0 tornado-4.2.1
[root@sandbox ~]#
[root@sandbox ~]#
```

Now, open a browser on your host machine and navigate to the URI http://127.0.0.1:8889 and you should see the screen below:

1. root@sandbox:~ (ssh)

```
Collecting terminado>=0.3.3 (from ipython[notebook])
  Downloading terminado-0.5.tar.gz
Collecting mistune>=0.5 (from ipython[notebook])
  Downloading mistune-0.7-py2.py3-none-any.whl
Requirement already satisfied (use --upgrade to upgrade): tornado>=4.0 in /opt/r
h/python27/root/usr/lib64/python2.7/site-packages (from ipython[notebook])
Requirement already satisfied (use --upgrade to upgrade): pyzmq>=13 in /opt/rh/p
ython27/root/usr/lib64/python2.7/site-packages (from ipython[notebook])
Requirement already satisfied (use --upgrade to upgrade): functools32 in /opt/rh
/python27/root/usr/lib/python2.7/site-packages (from jsonschema>=2.0->ipython[no
tebook])
Collecting ptyprocess (from terminado>=0.3.3->ipython[notebook])
  Downloading ptyprocess-0.5.tar.gz
Requirement already satisfied (use --upgrade to upgrade): backports.ssl-match-ho
stname in /opt/rh/python27/root/usr/lib/python2.7/site-packages (from tornado>=4
.0->ipython[notebook])
Requirement already satisfied (use --upgrade to upgrade): certifi in /opt/rh/pyt
hon27/root/usr/lib/python2.7/site-packages (from tornado>=4.0->ipython[notebook]
)
Installing collected packages: ptyprocess, terminado, mistune, ipython
  Running setup.py install for ptyprocess
  Running setup.py install for terminado
Successfully installed ipython-3.2.1 mistune-0.7 ptyprocess-0.5 terminado-0.5
[root@sandbox ~]#
```



```
1. root@sandbox:~ (ssh)
Collecting ptyprocess (from terminado>=0.3.3->ipython[notebook])
  Downloading ptyprocess-0.5.tar.gz
Requirement already satisfied (use --upgrade to upgrade): backports.ssl-match-ho
stname in /opt/rh/python27/root/usr/lib/python2.7/site-packages (from tornado>=4
.0->ipython[notebook])
Requirement already satisfied (use --upgrade to upgrade): certifi in /opt/rh/pyt
hon27/root/usr/lib/python2.7/site-packages (from tornado>=4.0->ipython[notebook]
)
Installing collected packages: ptyprocess, terminado, mistune, ipython
```

Voila! you have just configured IPython notebook with Apache Spark on you Sandbox.

In the next few tutorials we are going to explore how we can use IPython notebook to analyze and visualize data.

Get notified of new tutorials :

Leave your email for updates

Comments

John Kerley-Weeks | August 24, 2015 at 9:21 am | **Reply**

```
1. root@sandbox:~ (ssh)
  GNU nano 2.0.9 File: .../profile_pyspark/ipython_notebook_config.py Modified

# c.NotebookApp.base_url = '/'

# The session manager class to use.
# c.NotebookApp.session_manager_class = <class 'IPython.html.services.sessions.$

# Supply overrides for the tornado.web.Application that the IPython notebook
# uses.
# c.NotebookApp.tornado_settings = {}

# The directory to use for notebooks and kernels.
c.NotebookApp.notebook_dir = u'/usr/hdp/current/spark-client/'

# The kernel manager class to use.
# c.NotebookApp.kernel_manager_class = <class 'IPython.html.services.kernels.ke$

# The file where the cookie secret is stored.
# c.NotebookApp.cookie_secret_file = u''

# Supply SSL options for the tornado HTTPServer. See the tornado docs for
```

The current software will not create a ipython_notebook_config.py file.

You can work around the ipython_notebook_config.py issue by creating the following start_ipython_notebook.sh file

#!/bin/bash
source /opt/rh/python27/enable
IPYTHON_OPTS="notebook –port 8889 –notebook-dir=u'/usr/hdp/2.3.0.0-2557/spark/' –ip='*' –no-browser" pyspark

nigeljvm | September 1, 2015 at 4:46 pm | **Reply**

its a IPython/jupyter notebook, so the latest would look like
find the file at /root/.jupyter/jupyter_notebook_config.py
jupyter notebook –generate-config
# add the following
c.NotebookApp.ip = '*'
c.NotebookApp.open_browser = False
c.NotebookApp.port = 8889
c.NotebookApp.notebook_dir = u'/usr/hdp/current/spark-client/'

```
1. root@sandbox:~ (ssh)
  nano 2.0.9        File: /root/start_ipython_notebook.sh          Modified

#!/bin/bash
source /opt/rh/python27/enable
IPYTHON_OPTS="notebook --profile pyspark" pyspark

^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^V Next Page ^U UnCut Text^T To Spell
```

Sean Creedon | October 2, 2015 at 3:39 am | **Reply**

if you use
jupyter notebook –generate-config

you need to remove –profile option in the start_ipython_notebook.sh
to become
IPYTHON_OPTS=”notebook” pyspark

ericF | October 23, 2015 at 7:02 am | **Reply**

@nigel:
Re: # find the file at /root/.jupyter/jupyter_notebook_config.py
it’s not AT this location. Your comment was too terse to be useful. Try again?

lcr | September 19, 2015 at 8:10 am | **Reply**

```
1. root@sandbox:~ (ssh)
/ipython_config.py'
[ProfileCreate] Generating default config file: u'/root/.ipython/profile_pyspark
/ipython_kernel_config.py'
[ProfileCreate] Generating default config file: u'/root/.ipython/profile_pyspark
```

Instructions don't work with ipython 4.0 which is what you would install by default as of Sept 17 2015. If you want to follow the instructions you would need to install an earlier version, say
pip install "ipython[notebook]"==3.2.1

If you are trying this in an azure VM you would need to create an endpoint at port 8889, in which case it's recommended that you use a password. See here for reference: https://azure.microsoft.com/en-us/documentation/articles/virtual-machines-python-ipython-notebook/

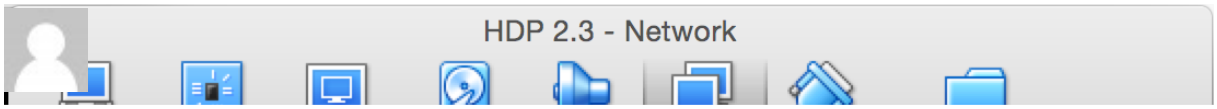
Eric | September 21, 2015 at 11:03 am | **Reply**

Instead of doing any of the "Configuring IPython" you can just run this command.

ipython notebook –port=8889 –notebook-dir=/usr/hdp/current/spark-client/ –ip=*

Pam | September 30, 2015 at 4:00 pm | **Reply**

Where can I find the "next few tutorials" that explores how to analyze and visualize data?



Werner de Jong | October 9, 2015 at 4:31 am | **Reply**

yum groupinstall “Development tools” gave an error:
not found in any dependency libraries.

After googling a bit I found a solution. Execute the following two commands before the yum groupinstall:

yum groups mark install “Development Tools”
yum groups mark convert “Development Tools”

yum groupinstall “Development Tools”

Corgi | November 22, 2015 at 8:08 am | **Reply**

HDP 2.3 - Network

Subscribe

I got for “yum groupinstall “Development tools“

Loaded plugins: fastestmirror, priorities
Setting up Group Process
Loading mirror speeds from cached hostfile
* base: mirrors.kernel.org
* epel: http://ftp.osuosl.org
* extras: mirror.sesp.northwestern.edu
* updates: mirror.supremebytes.com
No packages in any requested group available to install or update

It would be nice if these tutorials are “proofed” and corrected regularly!

Werner de Jong | October 9, 2015 at 5:32 am | **Reply**

when pip install “ipython[notebook]“ fails.

[root@sandbox ~]# pip install “ipython[notebook]“
Exception:
Traceback (most recent call last):
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/basecommand.py”, line 211, in main

status = self.run(options, args)
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/commands/install.py”, line 282, in run
wheel_cache
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/basecommand.py”, line 272, in populate_requirement_set
wheel_cache=wheel_cache
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/req/req_install.py”, line 213, in from_line
wheel_cache=wheel_cache, constraint=constraint)
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/req/req_install.py”, line 67, in __init__
req = pkg_resources.Requirement.parse(req)
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/pkg_resources/__init__.py”, line 2980, in parse
reqs = list(parse_requirements(s))
File “/opt/rh/python27/root/usr/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/pkg_resources/__init__.py”, line 2911, in parse_requirements
raise RequirementParseError(“Missing distribution spec”, line)
RequirementParseError: Missing distribution spec “ipython[notebook]“

sometimes you can run the command as upgrade:

pip install –upgrade ipython[notebook]

Abhik | October 12, 2015 at 8:40 pm | **Reply**

Try this one without the quotes , worked for me …. for the pip install ipython error mentioned in the earlier comment

pip install ipython[notebook]

abhik | October 12, 2015 at 9:23 pm | **Reply**

https://github.com/zeromq/pyzmq/issues/658 – this gives the issue wth pip install pyzmq

James Sharrett | November 25, 2015 at 9:09 am | **Reply**

removing the quotes as Abhik suggested fixed this

Andreas | October 22, 2015 at 7:20 am | **Reply**

Why not simply use

yum install -y python-pip

??

Andy | October 27, 2015 at 7:18 am | **Reply**

I have installed everything. When I run ‘ipython profile create pyspark’, it does not create ipython_notebook_config.py file. I tried the solution proposed by John Kerley-Weeks, it did not work.

I use VMware player on Window. It does not have ‘Port Forwarding’ setting up.

Could anyone help me with this problem

Ivan | October 30, 2015 at 2:44 pm | **Reply**

I had the same issue. Seems the instructions are not up to date.
I created the file manually, containing only these lines:

c = get_config()

c.NotebookApp.ip = ‘*’
c.NotebookApp.open_browser = False
c.NotebookApp.port = 8880

John | November 16, 2015 at 3:40 pm | **Reply**

The fix shown below worked for me. Apparently, the structure of the config files between ipython and jupyter changed with ipython 4.0

http://datascience.stackexchange.com/questions/6555/issue-with-ipython-jupyter-on-spark-unrecognized-alias

ABOUT US

Investor Relations
Quick Facts
Management Team
Board Of Directors
Founders
Careers
Internships

PRESS

Press Releases

In The Press

PARTNERS

Systems Integrators

Technology Partners

Resellers

Certification

Become A Partner

CONNECT

Blog

Webinars

Events

CONTACT

Contact

+1 408 675-0983

1 855 8-HORTON

Privacy Policy | Terms of Service