Programming Hadoop

James G. Shanahan[1]

[1] *Church and Duncan Group and iSchool, UC Berkeley, CA*

EMAIL: James_DOT_Shanahan_AT_gmail_DOT_com

June 13, 2016

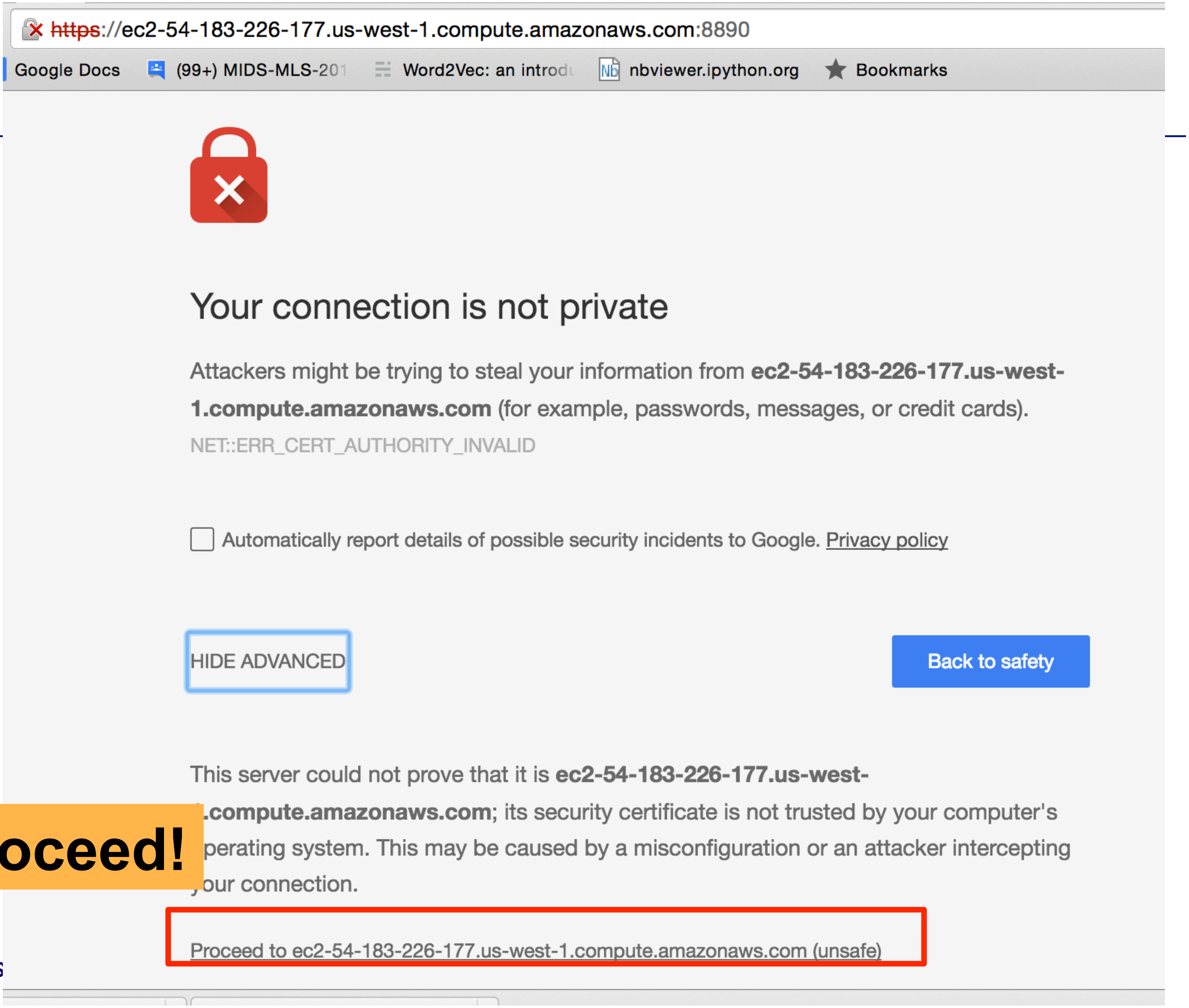
Lecture 4

To access the Jupyter notebook that is backed by a hadoop cluster

- Click here
- https://ec2-54-67-63-252.us-west-1.compute.amazonaws.com:8888

Access to Notebook Server

- https://ec2-54-183-226-177.us-west-1.compute.amazonaws.com:8890
- Enter Password ucbmids
- NOTE this cluster has 3 machines (1 host + 2 workers XLarge)



Your connection is not private

Attackers might be trying to steal your information from **ec2-54-183-226-177.us-west-1.compute.amazonaws.com** (for example, passwords, messages, or credit cards).

NET::ERR_CERT_AUTHORITY_INVALID

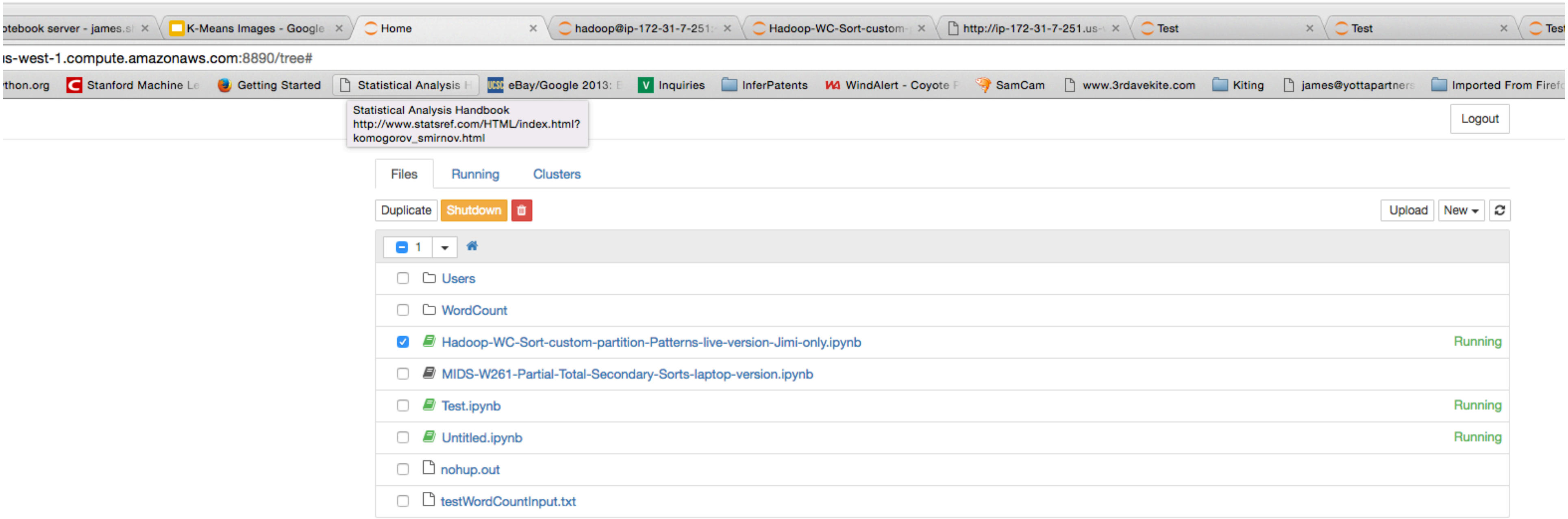
□ Automatically report details of possible security incidents to Google. Privacy policy

This server could not prove that it is **ec2-54-183-226-177.us-west-.compute.amazonaws.com**; its security certificate is not trusted by your computer's perating system. This may be caused by a misconfiguration or an attacker intercepting our connection.

Proceed to ec2-54-183-226-177.us-west-1.compute.amazonaws.com (unsafe)

Proceed!

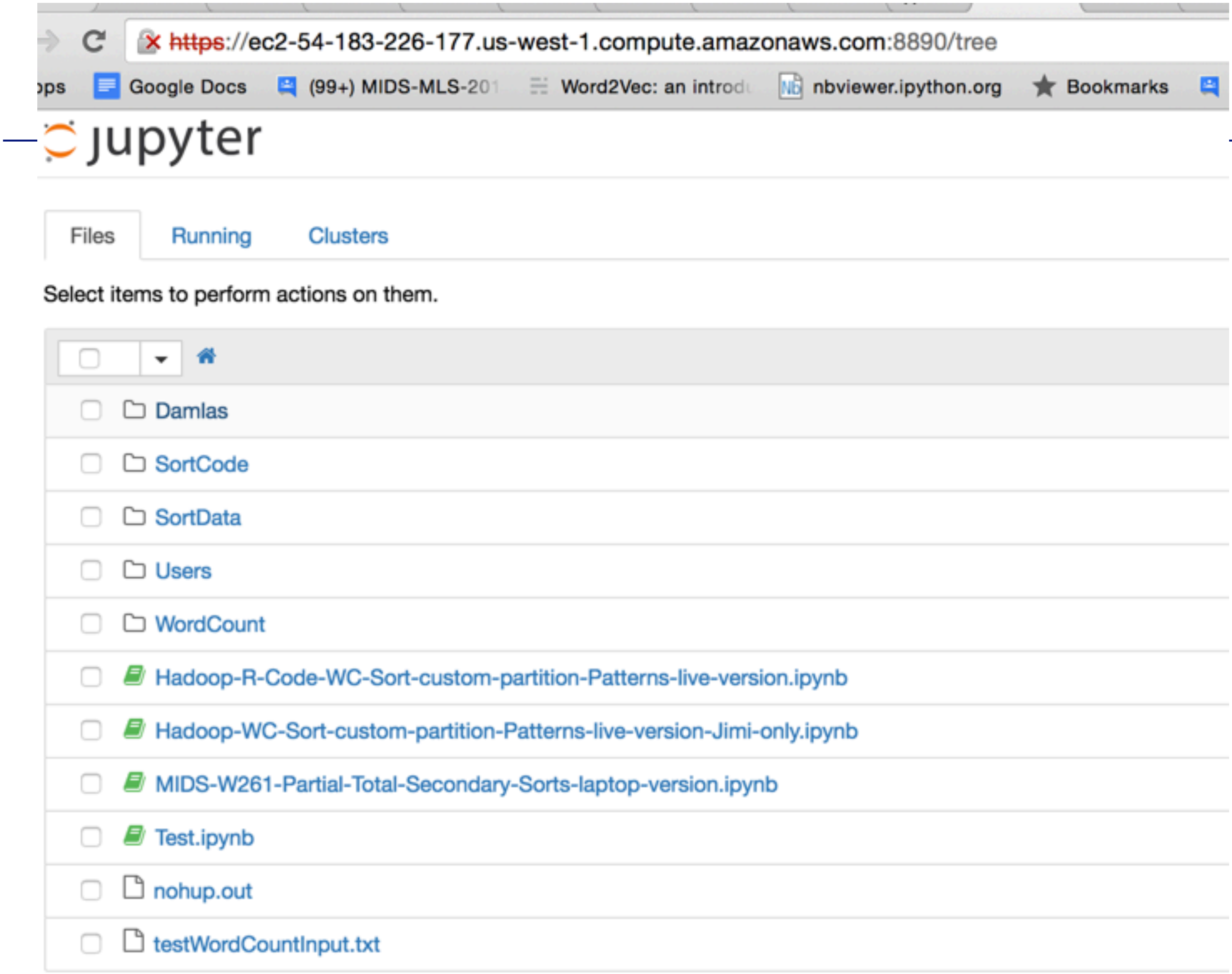
Notebooks



□	Users	
□	WordCount	
☑	Hadoop-WC-Sort-custom-partition-Patterns-live-version-Jimi-only.ipynb	Running
□	MIDS-W261-Partial-Total-Secondary-Sorts-laptop-version.ipynb	
□	Test.ipynb	Running
□	Untitled.ipynb	Running
□	nohup.out	
□	testWordCountInput.txt	

Please create your subdirectory under Users and keep your code and data there.

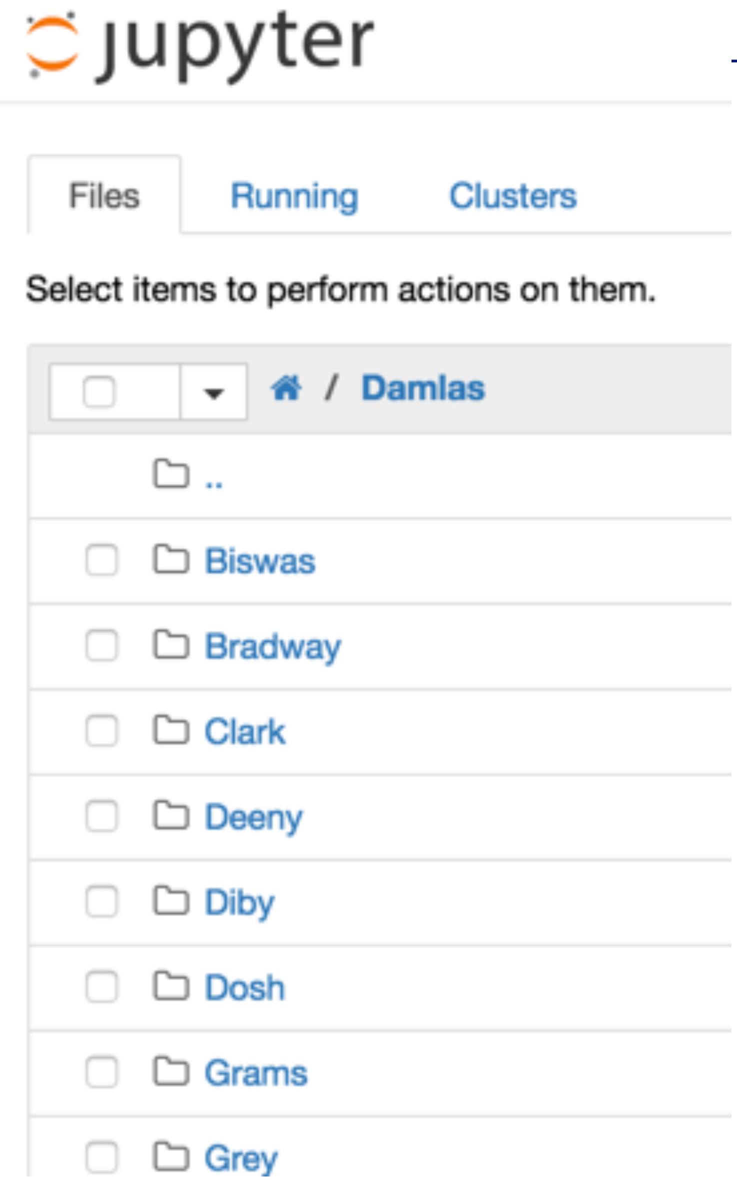
Keep backups of your notebooks locally!!



Select items to perform actions on them.

- Damlas
- SortCode
- SortData
- Users
- WordCount
- Hadoop-R-Code-WC-Sort-custom-partition-Patterns-live-version.ipynb
- Hadoop-WC-Sort-custom-partition-Patterns-live-version-Jimi-only.ipynb
- MIDS-W261-Partial-Total-Secondary-Sorts-laptop-version.ipynb
- Test.ipynb
- nohup.out
- testWordCountInput.txt

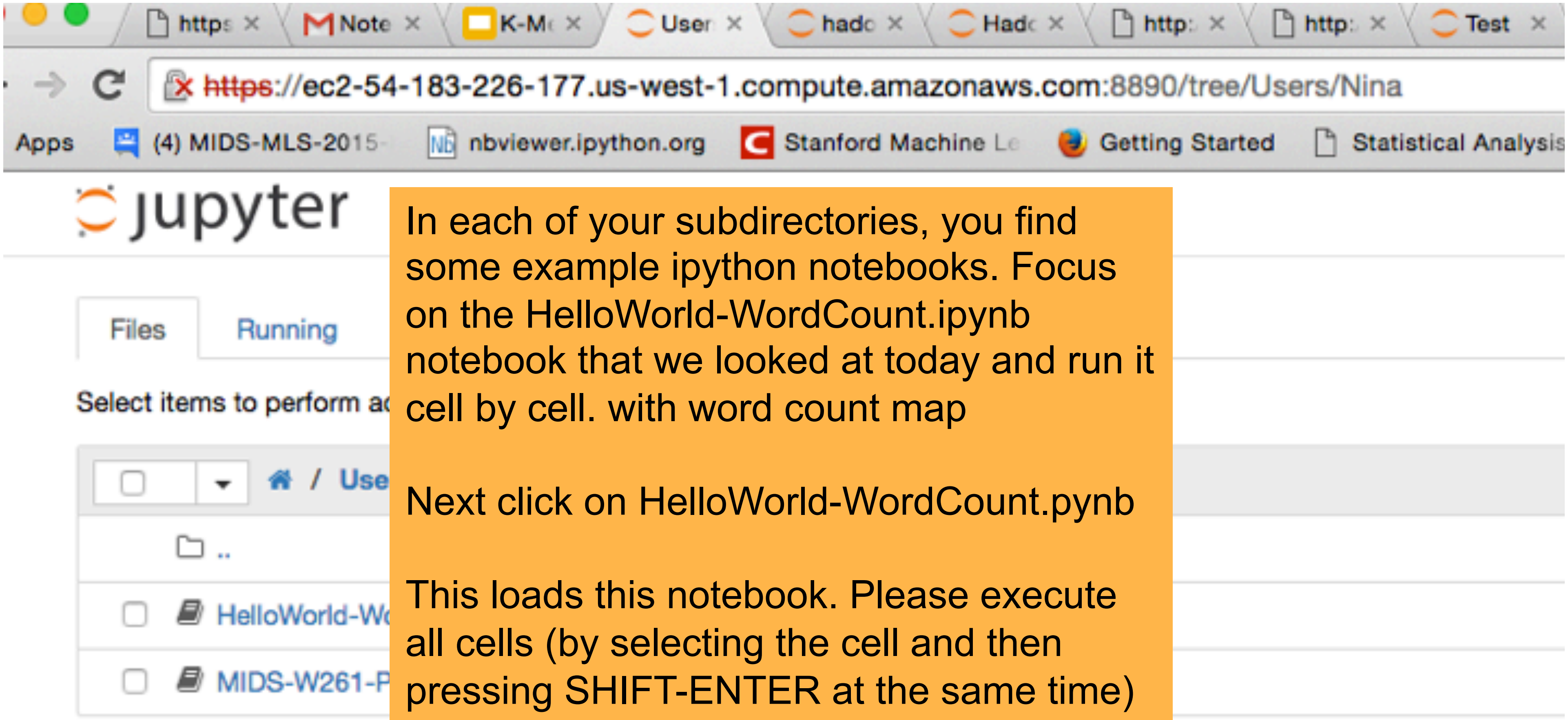
- ..



Select items to perform actions on them.

□ / Damlas

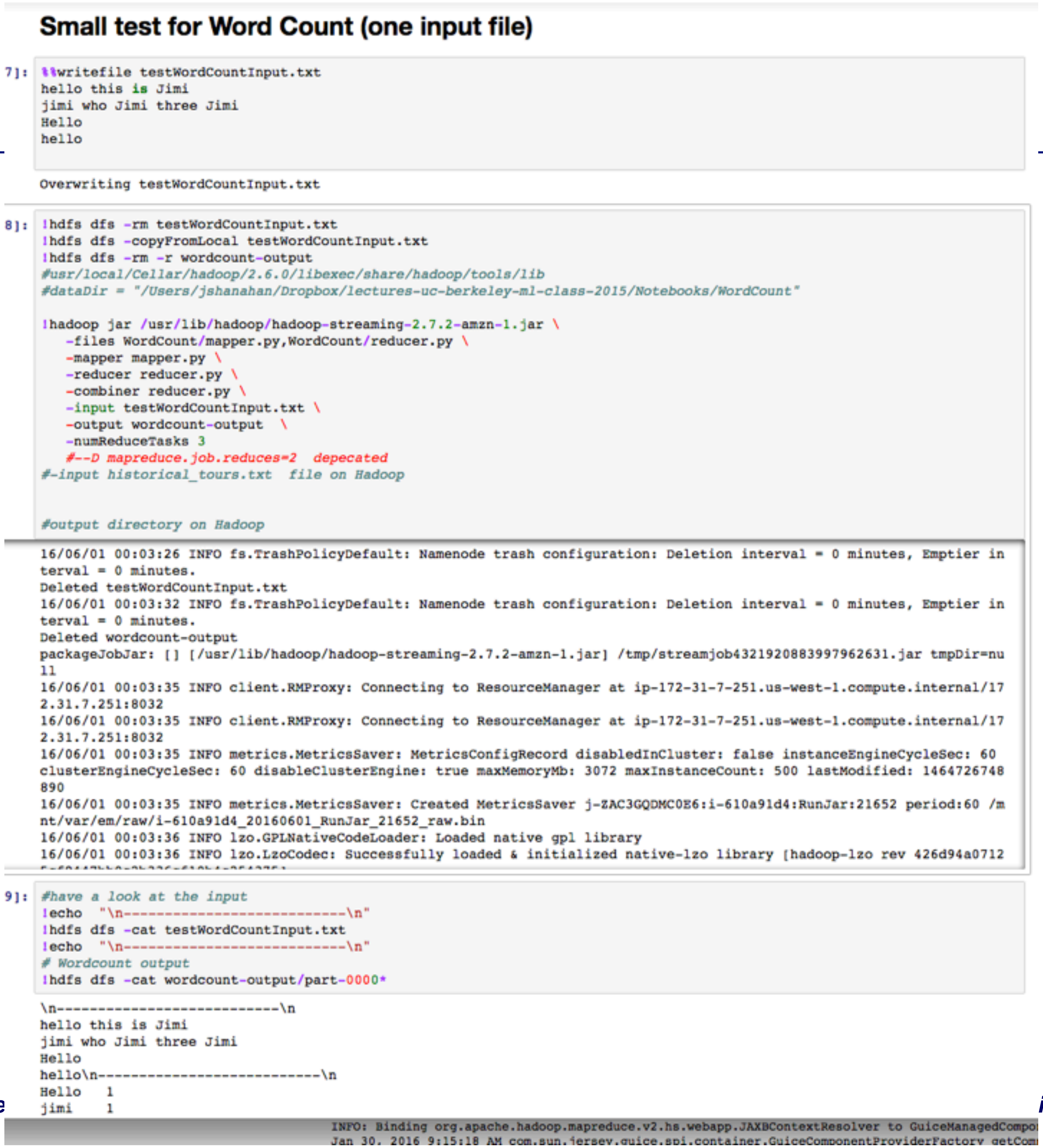
- ..
- □ Biswas
- □ Bradway
- □ Clark
- □ Deeny
- □ Diby
- □ Dosh
- □ Grams
- □ Grey



In each of your subdirectories, you find some example ipython notebooks. Focus on the HelloWorld-WordCount.ipynb notebook that we looked at today and run it cell by cell. with word count map

Next click on HelloWorld-WordCount.pynb

This loads this notebook. Please execute all cells (by selecting the cell and then pressing SHIFT-ENTER at the same time) before and the "Small test for word count" cell and then execute cell "In[17] and in[18], and In[19] to see the results of wordcount.



Small test for Word Count (one input file)

```
7]: %%writefile testWordCountInput.txt
hello this is Jimi
jimi who Jimi three Jimi
Hello
hello
```

```
Overwriting testWordCountInput.txt
```

```
8]: !hdfs dfs -rm testWordCountInput.txt
!hdfs dfs -copyFromLocal testWordCountInput.txt
!hdfs dfs -rm -r wordcount-output
#usr/local/Cellar/hadoop/2.6.0/libexec/share/hadoop/tools/lib
#dataDir = "/Users/jshanahan/Dropbox/lectures-uc-berkeley-ml-class-2015/Notebooks/WordCount"

!hadoop jar /usr/lib/hadoop/hadoop-streaming-2.7.2-amzn-1.jar \
   -files WordCount/mapper.py,WordCount/reducer.py \
   -mapper mapper.py \
   -reducer reducer.py \
   -combiner reducer.py \
   -input testWordCountInput.txt \
   -output wordcount-output  \
   -numReduceTasks 3
   #--D mapreduce.job.reduces=2  depecated
#-input historical_tours.txt  file on Hadoop


#output directory on Hadoop
```

```
16/06/01 00:03:26 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier in
terval = 0 minutes.
Deleted testWordCountInput.txt
16/06/01 00:03:32 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier in
terval = 0 minutes.
Deleted wordcount-output
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.7.2-amzn-1.jar] /tmp/streamjob4321920883997962631.jar tmpDir=nu
ll
16/06/01 00:03:35 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-251.us-west-1.compute.internal/17
2.31.7.251:8032
16/06/01 00:03:35 INFO client.RMProxy: Connecting to ResourceManager at ip-172-31-7-251.us-west-1.compute.internal/17
2.31.7.251:8032
16/06/01 00:03:35 INFO metrics.MetricsSaver: MetricsConfigRecord disabledInCluster: false instanceEngineCycleSec: 60
clusterEngineCycleSec: 60 disableClusterEngine: true maxMemoryMb: 3072 maxInstanceCount: 500 lastModified: 1464726748
890
16/06/01 00:03:35 INFO metrics.MetricsSaver: Created MetricsSaver j-ZAC3GQDMC0E6:i-610a91d4:RunJar:21652 period:60 /m
nt/var/em/raw/i-610a91d4_20160601_RunJar_21652_raw.bin
16/06/01 00:03:36 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
16/06/01 00:03:36 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 426d94a0712
```

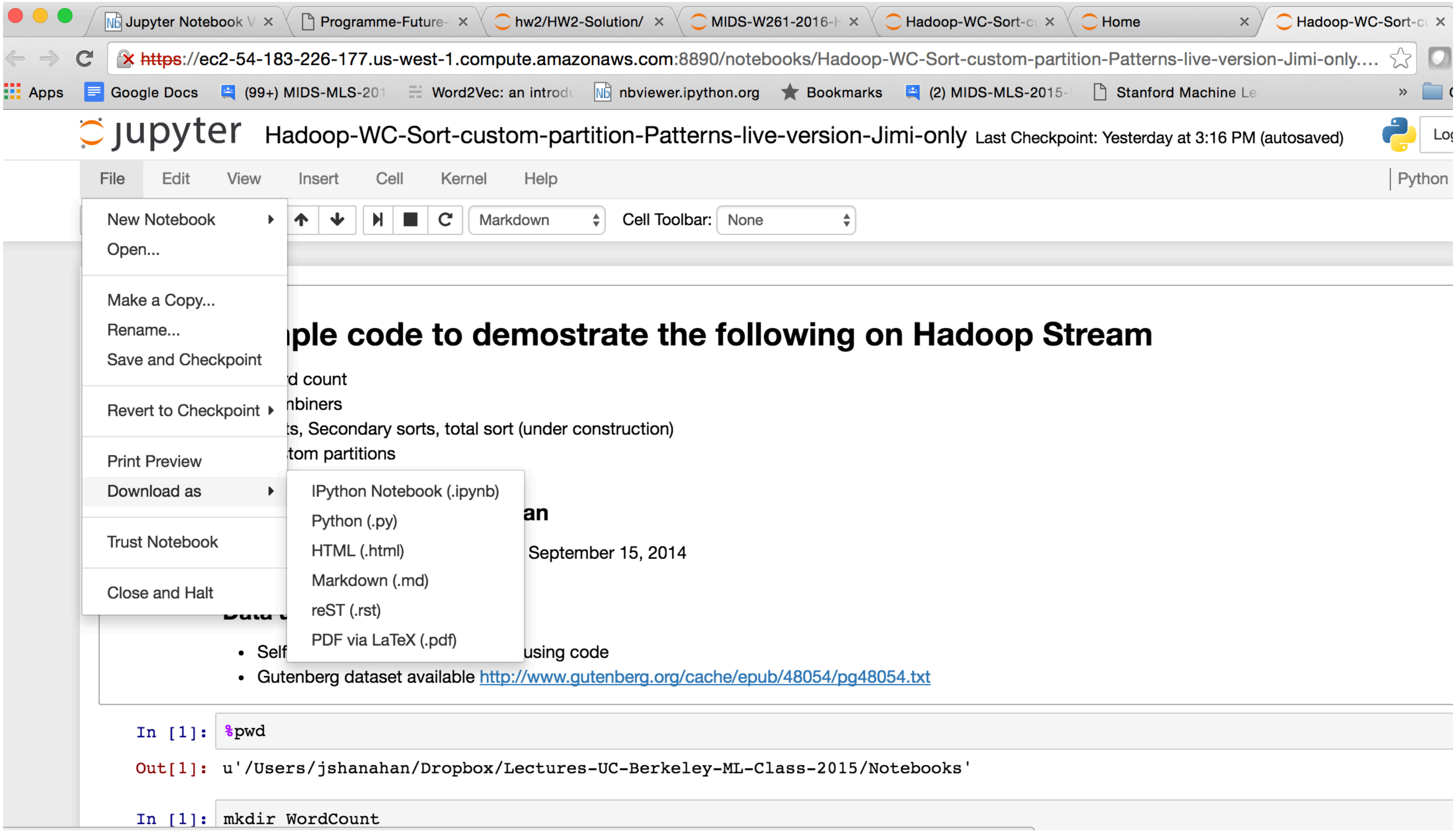
```
9]: #have a look at the input
!echo  "\n--------------------------\n"
!hdfs dfs -cat testWordCountInput.txt
!echo  "\n--------------------------\n"
# Wordcount output
!hdfs dfs -cat wordcount-output/part-0000*
```

```
\n--------------------------\n
hello this is Jimi
jimi who Jimi three Jimi
Hello
hello\n--------------------------\n
Hello    1
jimi     1
```

```
INFO: Binding org.apache.hadoop.mapreduce.v2.hs.webapp.JAXBContextResolver to GuiceManagedCompo
Jan 30, 2016 9:15:18 AM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getCom
```

- ..

Save cloud notebook to your local machine



ple code to demostrate the following on Hadoop Stream

d count

nbiners

ts, Secondary sorts, total sort (under construction)

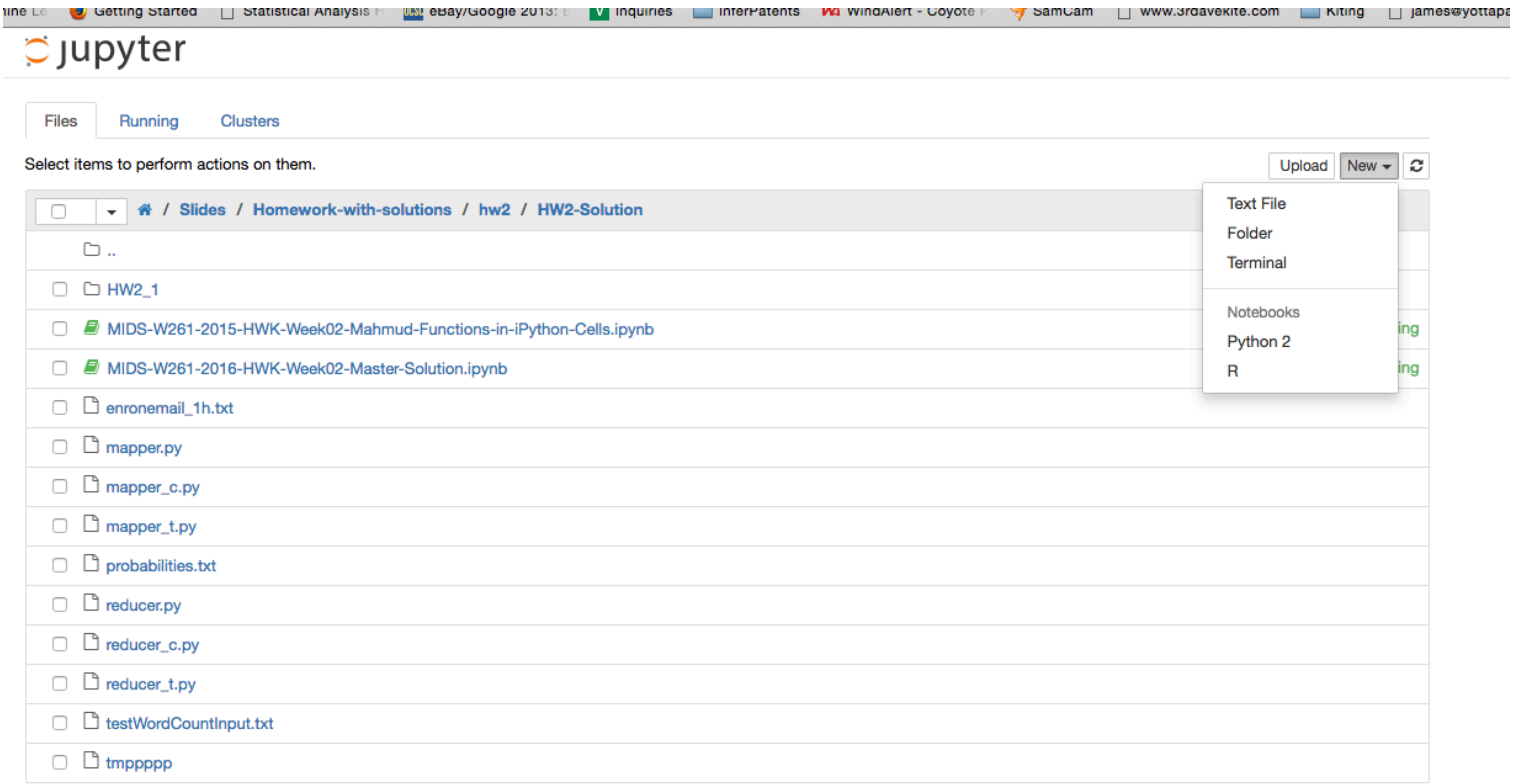
tom partitions

an

September 15, 2014

- Self using code
- Gutenberg dataset available http://www.gutenberg.org/cache/epub/48054/pg48054.txt

```
In [1]: %pwd

Out[1]: u'/Users/jshanahan/Dropbox/Lectures-UC-Berkeley-ML-Class-2015/Notebooks'

In [1]: mkdir WordCount
```



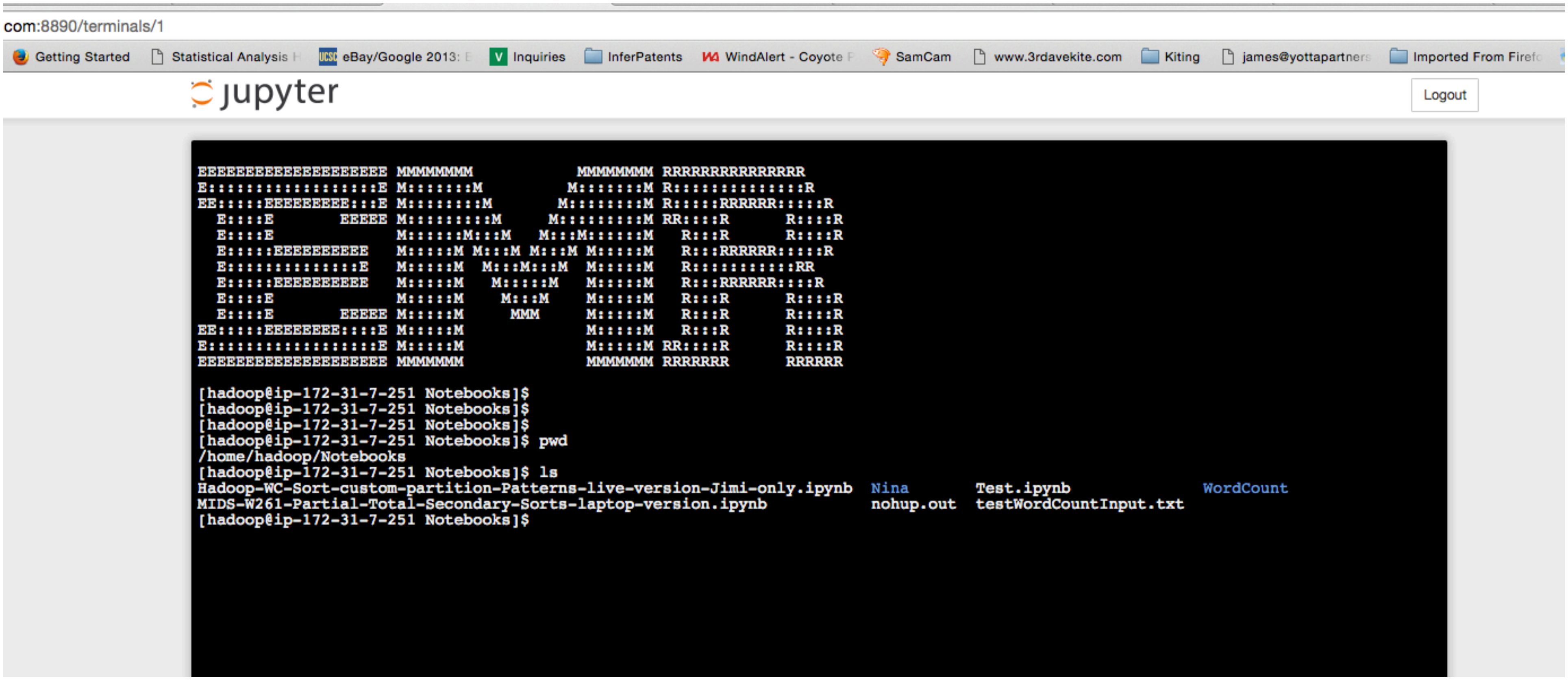
Select items to perform actions on them.

/ Slides / Homework-with-solutions / hw2 / HW2-Solution

- ..
- □ HW2_1
- □ MIDS-W261-2015-HWK-Week02-Mahmud-Functions-in-iPython-Cells.ipynb
- □ MIDS-W261-2016-HWK-Week02-Master-Solution.ipynb
- □ enronemail_1h.txt
- □ mapper.py
- □ mapper_c.py
- □ mapper_t.py
- □ probabilities.txt
- □ reducer.py
- □ reducer_c.py
- □ reducer_t.py
- □ testWordCountInput.txt
- □ tmppppp

R mapper and reducer for wordcount

- **Hadoop-R-Code-WC-Sort-custom-partition-Patterns-live-version.ipynb**



```
EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-7-251 Notebooks]$
[hadoop@ip-172-31-7-251 Notebooks]$
[hadoop@ip-172-31-7-251 Notebooks]$
[hadoop@ip-172-31-7-251 Notebooks]$ pwd
/home/hadoop/Notebooks
[hadoop@ip-172-31-7-251 Notebooks]$ ls
Hadoop-WC-Sort-custom-partition-Patterns-live-version-Jimi-only.ipynb  Nina       Test.ipynb              WordCount
MIDS-W261-Partial-Total-Secondary-Sorts-laptop-version.ipynb           nohup.out  testWordCountInput.txt
[hadoop@ip-172-31-7-251 Notebooks]$
```

- **Main Steps to setup an iPython Notebook server on the cloud**

Office hours week 4: Outline

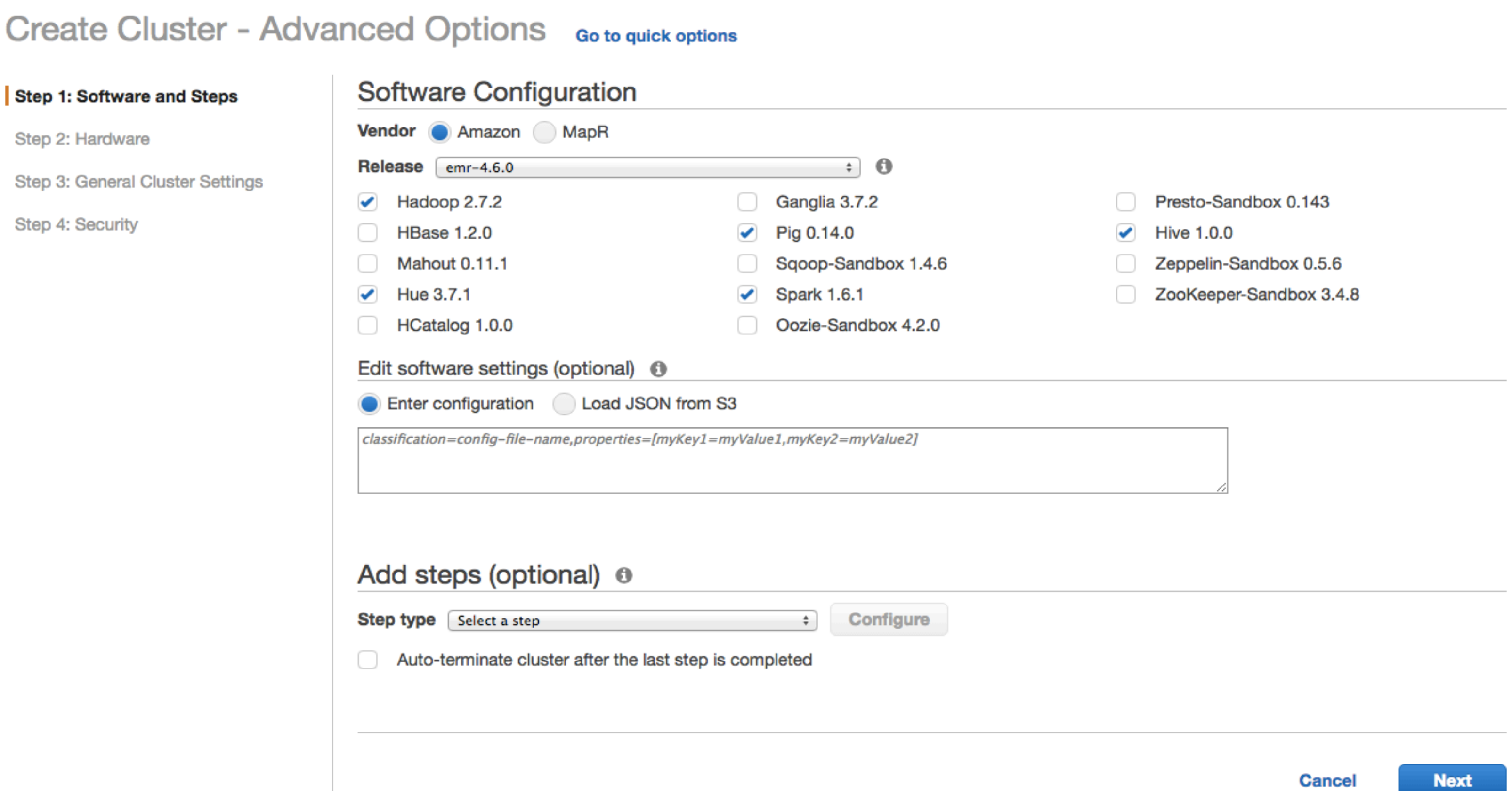
- **Hadoop on a VM**
  - https://docs.google.com/presentation/d/1qCQM-2U2C6e584uM9kqTGr675K3_a8M1mEZaiT4Wmi8/edit?usp=sharing
- **Hadoop on EMR**
- **iPython Notebook server on the cloud(see slides)**
  - http://blog.impiyush.me/2015/02/running-ipython-notebook-server-on-aws.html
  - Here are the instructions for installing Jupyter Notebook server on the headnode of a cluster. It will enable you to open notebooks on the headnode using your browser and run the notebooks just like you do on your laptop.
    - http://jupyter-notebook.readthedocs.io/en/latest/public_server.html

Main Steps to setup an iPython Notebook server on the cloud

- **Launch an EMR Cluster on Amazon**
  - Follow usual steps PLUS new security step for master node
- **Log into Master node**
  - Install ipython (via Anaconda)
  - Configure (access stuff)
- **Publish address and password**

AWS security group

Create cluster- Advanced options



Create Cluster - Advanced Options

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Vendor ● Amazon ○ MapR

Release emr-4.6.0

☑ Hadoop 2.7.2
□ HBase 1.2.0
□ Mahout 0.11.1
☑ Hue 3.7.1
□ HCatalog 1.0.0
□ Ganglia 3.7.2
☑ Pig 0.14.0
□ Sqoop-Sandbox 1.4.6
☑ Spark 1.6.1
□ Oozie-Sandbox 4.2.0
□ Presto-Sandbox 0.143
☑ Hive 1.0.0
□ Zeppelin-Sandbox 0.5.6
□ ZooKeeper-Sandbox 3.4.8

Edit software settings (optional)

● Enter configuration ○ Load JSON from S3

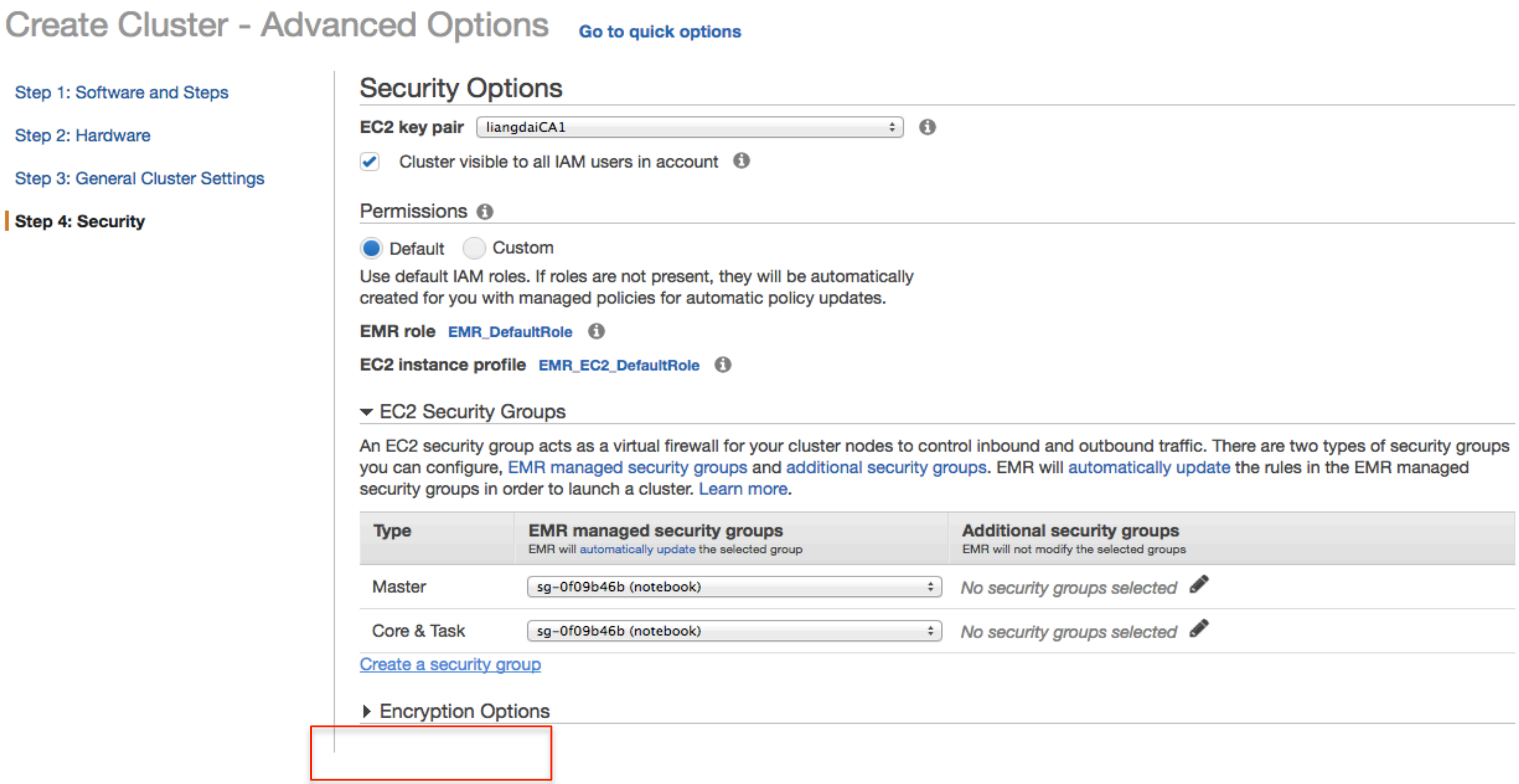
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]

Add steps (optional)

Step type Select a step

□ Auto-terminate cluster after the last step is completed

Step 4: Security



Create Cluster - Advanced Options

Go to quick options

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Security Options

EC2 key pair liangdaiCA1

☑ Cluster visible to all IAM users in account

Permissions

● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role EMR_DefaultRole

EC2 instance profile EMR_EC2_DefaultRole

EC2 Security Groups

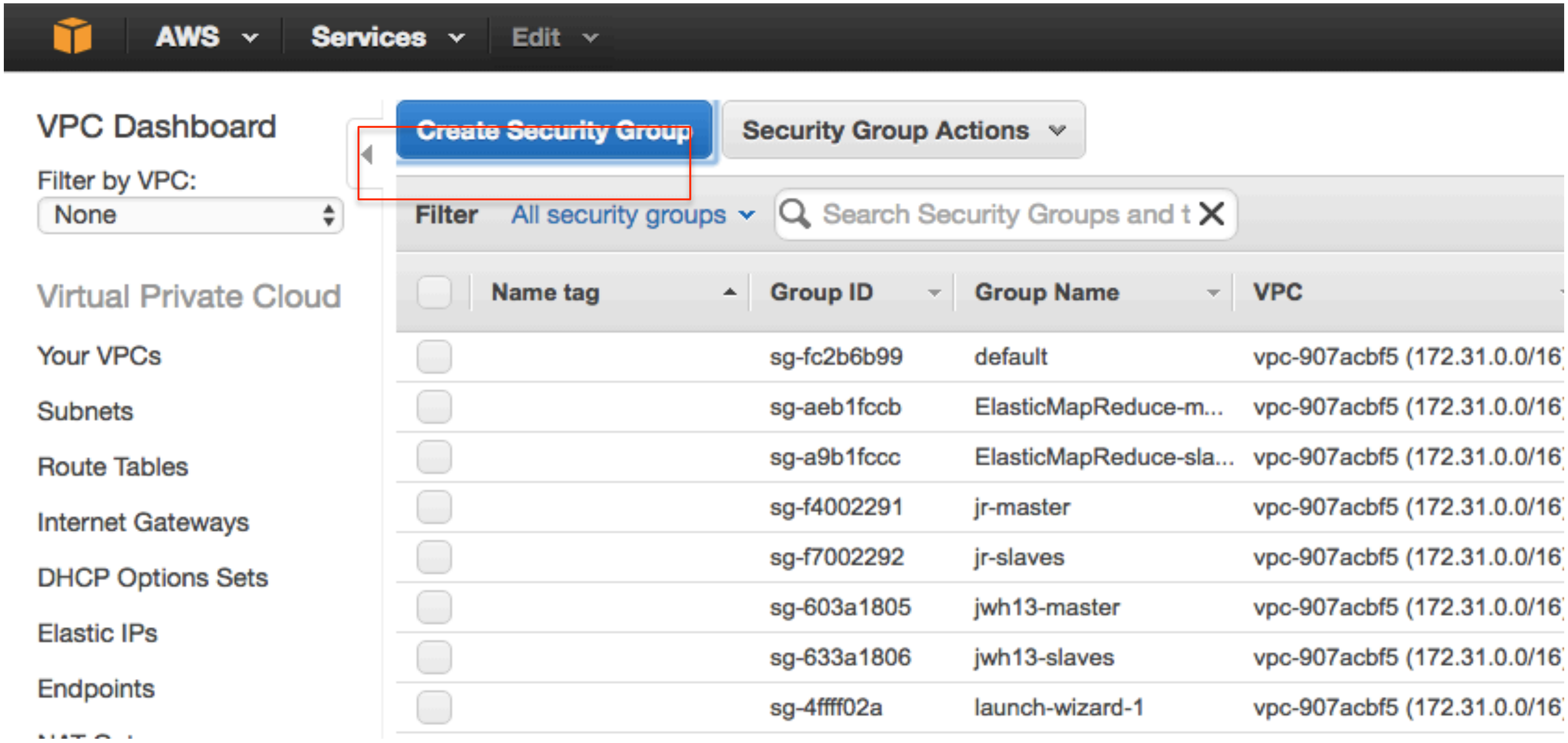
An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, EMR managed security groups and additional security groups. EMR will automatically update the rules in the EMR managed security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	sg-0f09b46b (notebook)	*No security groups selected*
Core & Task	sg-0f09b46b (notebook)	*No security groups selected*

Create a security group

Encryption Options

Create security group

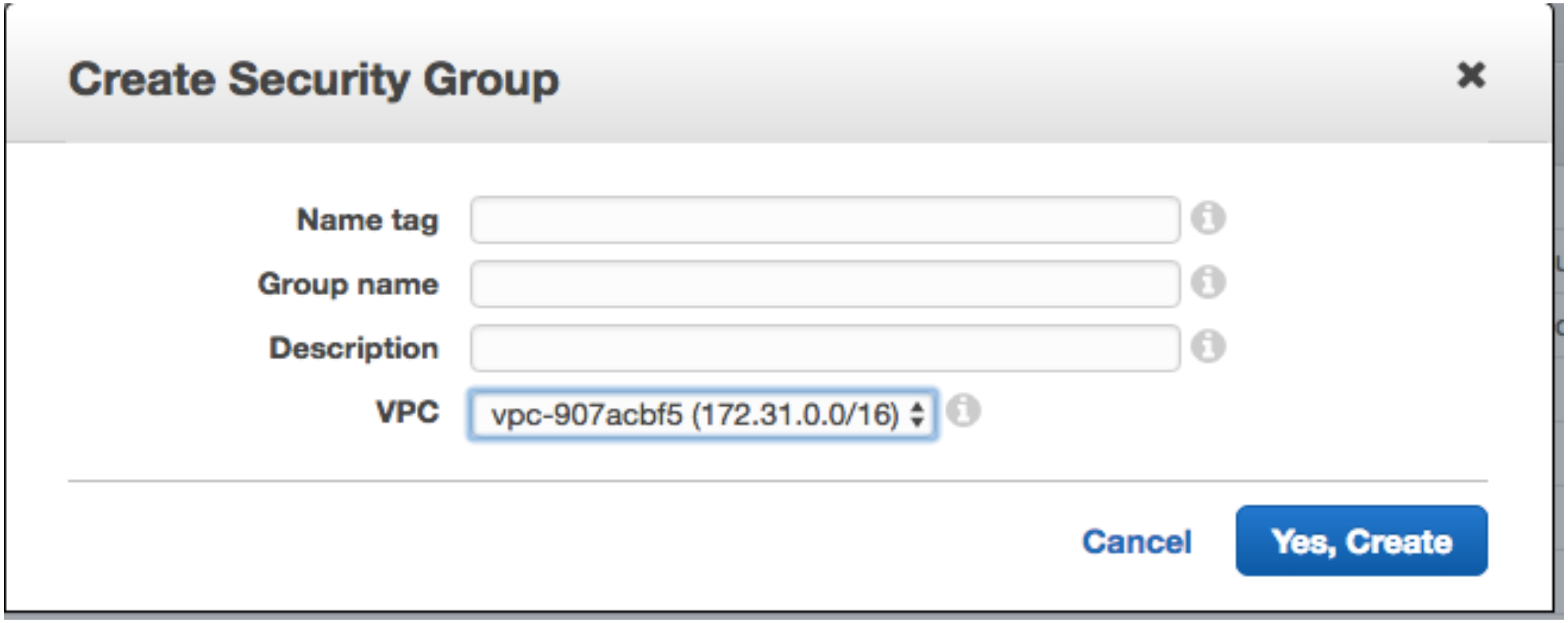


Filter by VPC: None

Filter: All security groups

□	Name tag	Group ID	Group Name	VPC
□		sg-fc2b6b99	default	vpc-907acbf5 (172.31.0.0/16
□		sg-aeb1fccb	ElasticMapReduce-m...	vpc-907acbf5 (172.31.0.0/16
□		sg-a9b1fccc	ElasticMapReduce-sla...	vpc-907acbf5 (172.31.0.0/16
□		sg-f4002291	jr-master	vpc-907acbf5 (172.31.0.0/16
□		sg-f7002292	jr-slaves	vpc-907acbf5 (172.31.0.0/16
□		sg-603a1805	jwh13-master	vpc-907acbf5 (172.31.0.0/16
□		sg-633a1806	jwh13-slaves	vpc-907acbf5 (172.31.0.0/16
□		sg-4ffff02a	launch-wizard-1	vpc-907acbf5 (172.31.0.0/16

Name the group



Create Security Group

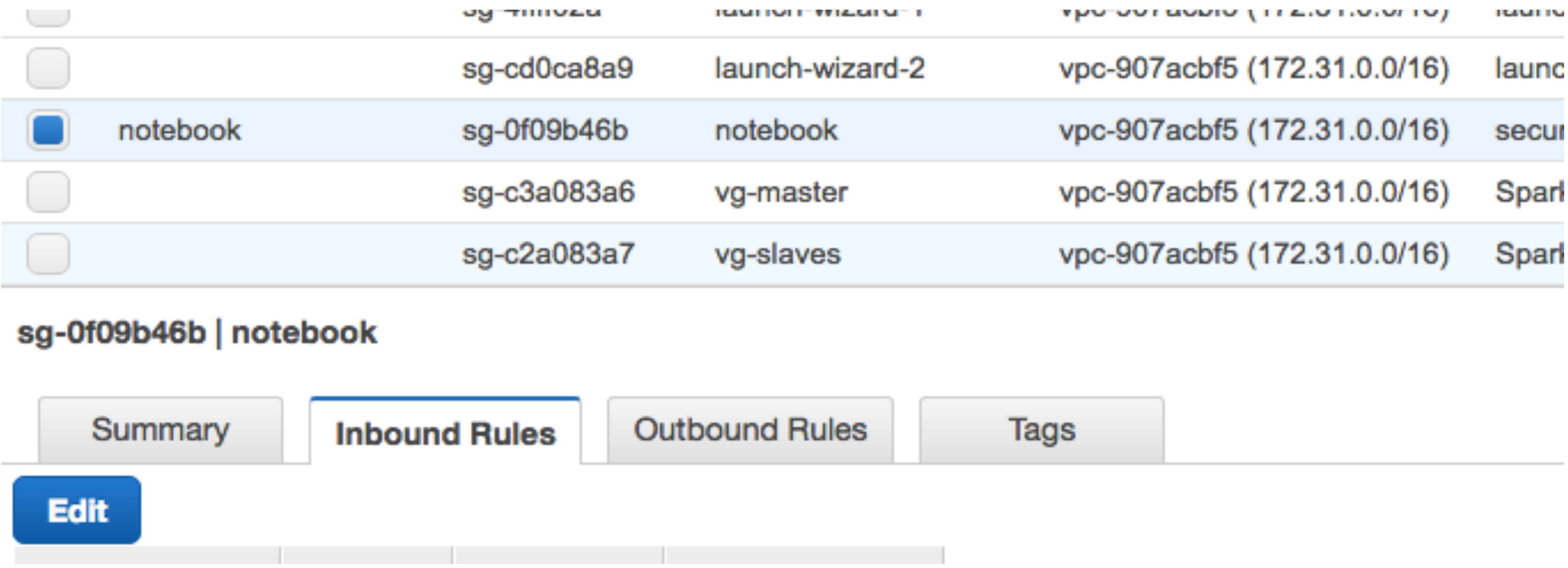
Name tag

Group name

Description

VPC vpc-907acbf5 (172.31.0.0/16)

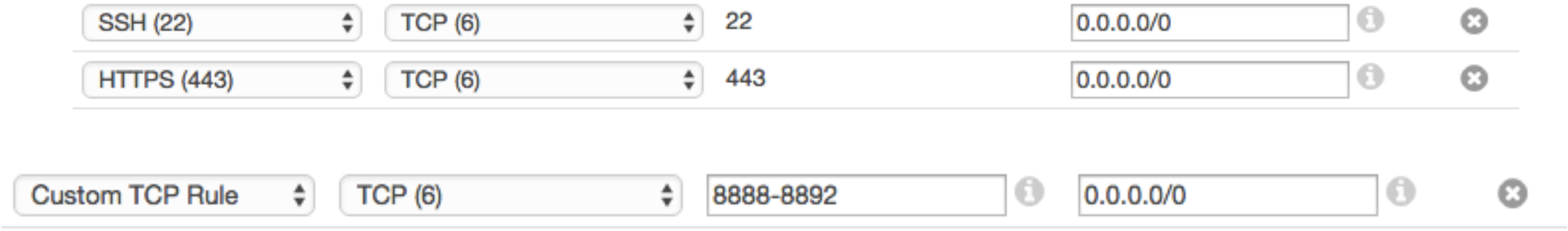
Select the new group and click the tap inbound Rules



☐		sg-cd0ca8a9	launch-wizard-2	vpc-907acbf5 (172.31.0.0/16)	launc
☑	notebook	sg-0f09b46b	notebook	vpc-907acbf5 (172.31.0.0/16)	secu
☐		sg-c3a083a6	vg-master	vpc-907acbf5 (172.31.0.0/16)	Spar
☐		sg-c2a083a7	vg-slaves	vpc-907acbf5 (172.31.0.0/16)	Spar

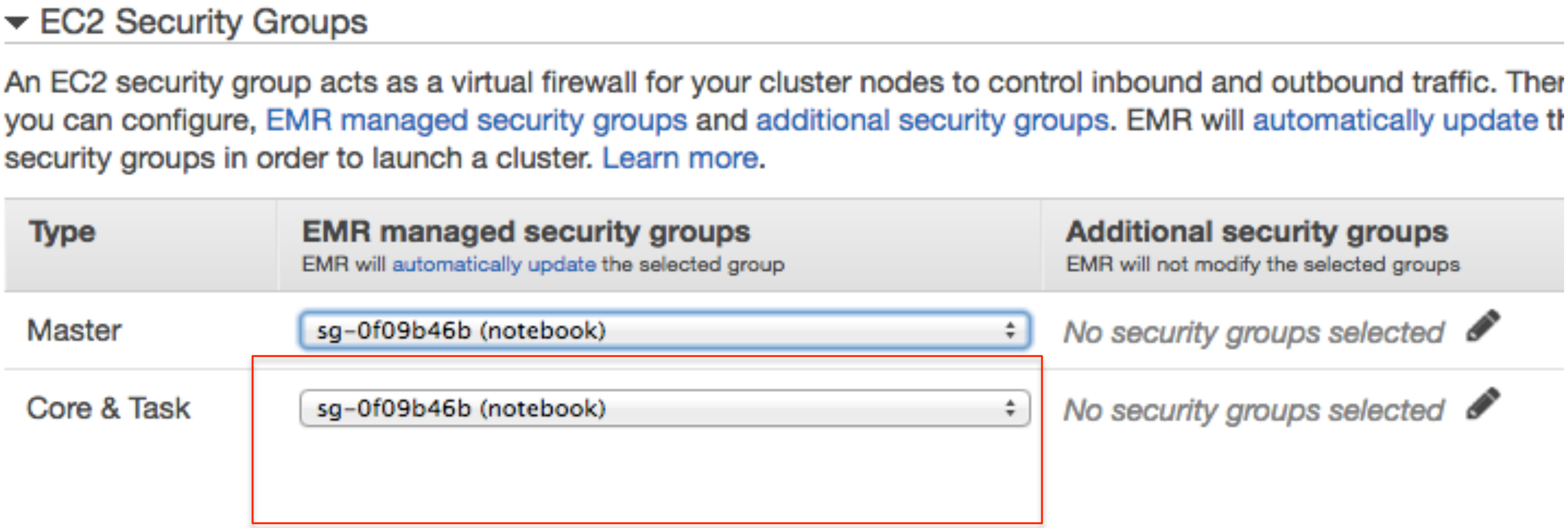
sg-0f09b46b | notebook

Edit/Add three rules



SSH (22)	TCP (6)	22	0.0.0.0/0
HTTPS (443)	TCP (6)	443	0.0.0.0/0
Custom TCP Rule	TCP (6)	8888-8892	0.0.0.0/0

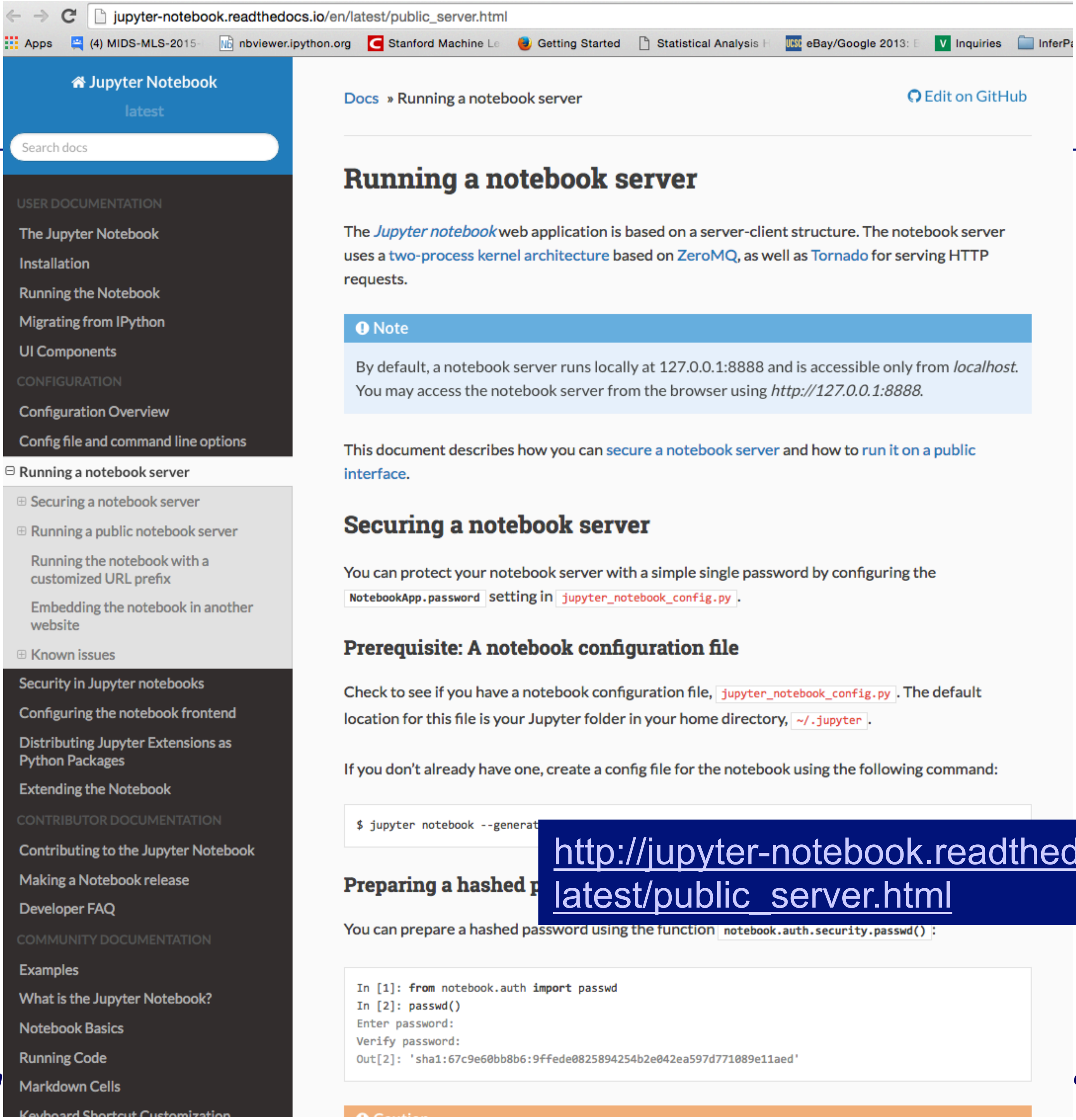
Choose the security group created at step 4 and create a cluster



▾ EC2 Security Groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. Ther you can configure, EMR managed security groups and additional security groups. EMR will automatically update tl security groups in order to launch a cluster. Learn more.

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	sg-0f09b46b (notebook)	*No security groups selected*
Core & Task	sg-0f09b46b (notebook)	*No security groups selected*


jupyter-notebook.readthedocs.io/en/latest/public_server.html
Apps
(4) MIDS-MLS-2015-
nbviewer.ipython.org
Stanford Machine Le
Getting Started
Statistical Analysis H
eBay/Google 2013: E
Inquiries
InferPa
Jupyter Notebook
latest
Search docs
USER DOCUMENTATION
The Jupyter Notebook
Installation
Running the Notebook
Migrating from IPython
UI Components
CONFIGURATION
Configuration Overview
Config file and command line options
Running a notebook server
Securing a notebook server
Running a public notebook server
Running the notebook with a customized URL prefix
Embedding the notebook in another website
Known issues
Security in Jupyter notebooks
Configuring the notebook frontend
Distributing Jupyter Extensions as Python Packages
Extending the Notebook
CONTRIBUTOR DOCUMENTATION
Contributing to the Jupyter Notebook
Making a Notebook release
Developer FAQ
COMMUNITY DOCUMENTATION
Examples
What is the Jupyter Notebook?
Notebook Basics
Running Code
Markdown Cells
Keyboard Shortcut Customization
Docs » Running a notebook server
Edit on GitHub

Running a notebook server

The *Jupyter notebook* web application is based on a server-client structure. The notebook server uses a two-process kernel architecture based on ZeroMQ, as well as Tornado for serving HTTP requests.

Note

By default, a notebook server runs locally at 127.0.0.1:8888 and is accessible only from *localhost*. You may access the notebook server from the browser using *http://127.0.0.1:8888*.

This document describes how you can secure a notebook server and how to run it on a public interface.

Securing a notebook server

You can protect your notebook server with a simple single password by configuring the `NotebookApp.password` setting in `jupyter_notebook_config.py`.

Prerequisite: A notebook configuration file

Check to see if you have a notebook configuration file, `jupyter_notebook_config.py`. The default location for this file is your Jupyter folder in your home directory, `~/.jupyter`.

If you don't already have one, create a config file for the notebook using the following command:

```
$ jupyter notebook --generat
```

http://jupyter-notebook.readthedocs.io/en/latest/public_server.html

Preparing a hashed p

You can prepare a hashed password using the function `notebook.auth.security.passwd()`:

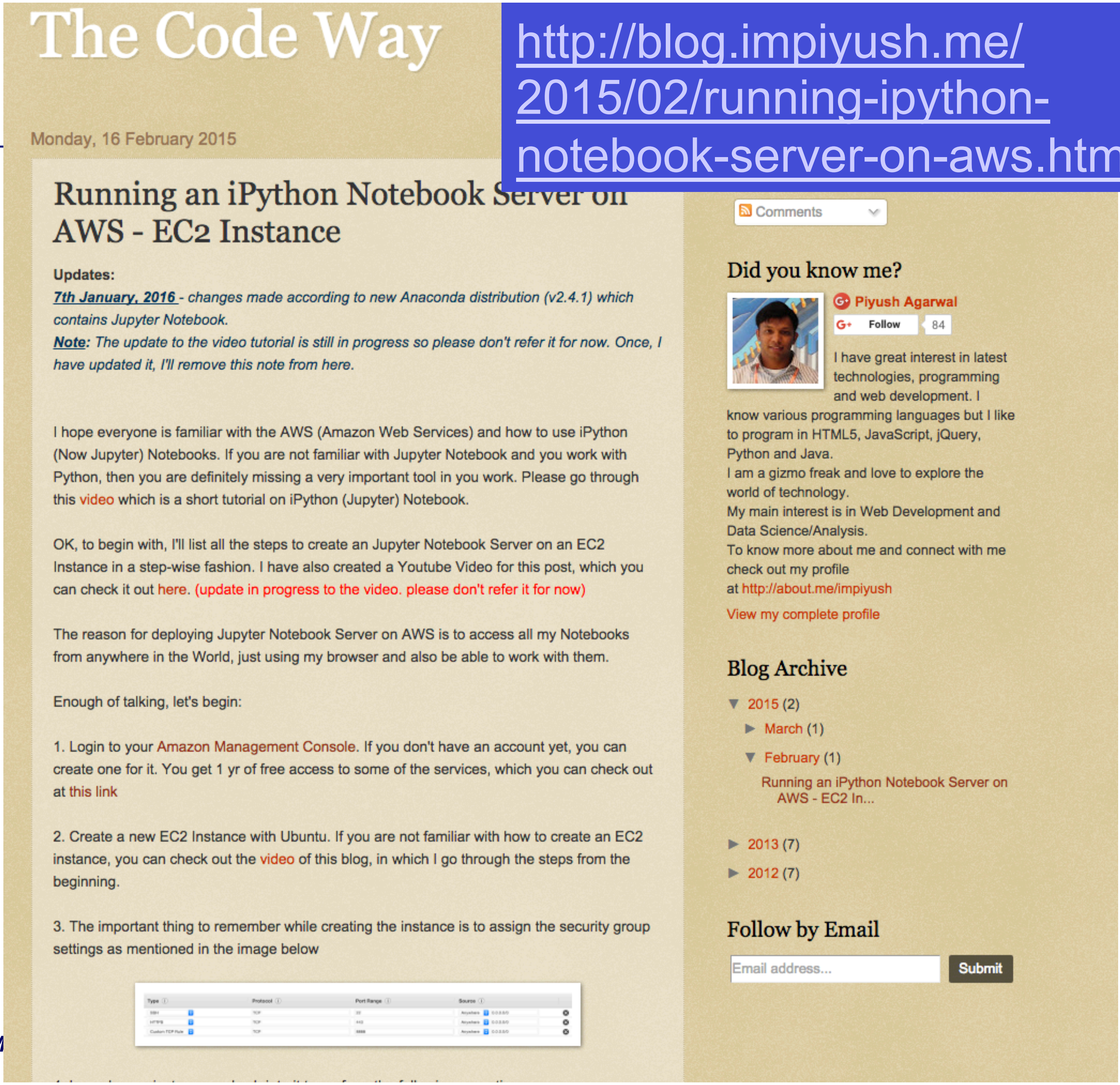
```
In [1]: from notebook.auth import passwd
In [2]: passwd()
Enter password:
Verify password:
Out[2]: 'sha1:67c9e60bb8b6:9ffede0825894254b2e042ea597d771089e11aed'
```

iPython Notebook server on the cloud

- **I believe it is free for one year to use ipython notebook server (One Micro Tier EC2 Instance).**
  - http://blog.impiyush.me/2015/02/running-ipython-notebook-server-on-aws.html

- **Another page that was also useful is:**
  - https://gist.github.com/iamatypeofwalrus/5183133

http://blog.impiyush.me/2015/02/running-ipython-notebook-server-on-aws.html



The Code Way

Monday, 16 February 2015

Running an iPython Notebook Server on AWS - EC2 Instance

Updates:

7th January, 2016 - changes made according to new Anaconda distribution (v2.4.1) which contains Jupyter Notebook.

Note: The update to the video tutorial is still in progress so please don't refer it for now. Once, I have updated it, I'll remove this note from here.

I hope everyone is familiar with the AWS (Amazon Web Services) and how to use iPython (Now Jupyter) Notebooks. If you are not familiar with Jupyter Notebook and you work with Python, then you are definitely missing a very important tool in you work. Please go through this video which is a short tutorial on iPython (Jupyter) Notebook.

OK, to begin with, I'll list all the steps to create an Jupyter Notebook Server on an EC2 Instance in a step-wise fashion. I have also created a Youtube Video for this post, which you can check it out here. (update in progress to the video. please don't refer it for now)

The reason for deploying Jupyter Notebook Server on AWS is to access all my Notebooks from anywhere in the World, just using my browser and also be able to work with them.

Enough of talking, let's begin:

1. Login to your Amazon Management Console. If you don't have an account yet, you can create one for it. You get 1 yr of free access to some of the services, which you can check out at this link

2. Create a new EC2 Instance with Ubuntu. If you are not familiar with how to create an EC2 instance, you can check out the video of this blog, in which I go through the steps from the beginning.

3. The important thing to remember while creating the instance is to assign the security group settings as mentioned in the image below

Type	Protocol	Port Range	Source
SSH	TCP	22	Anywhere 0.0.0.0/0
HTTPS	TCP	443	Anywhere 0.0.0.0/0
Custom TCP Rule	TCP	8888	Anywhere 0.0.0.0/0

Did you know me?

Piyush Agarwal

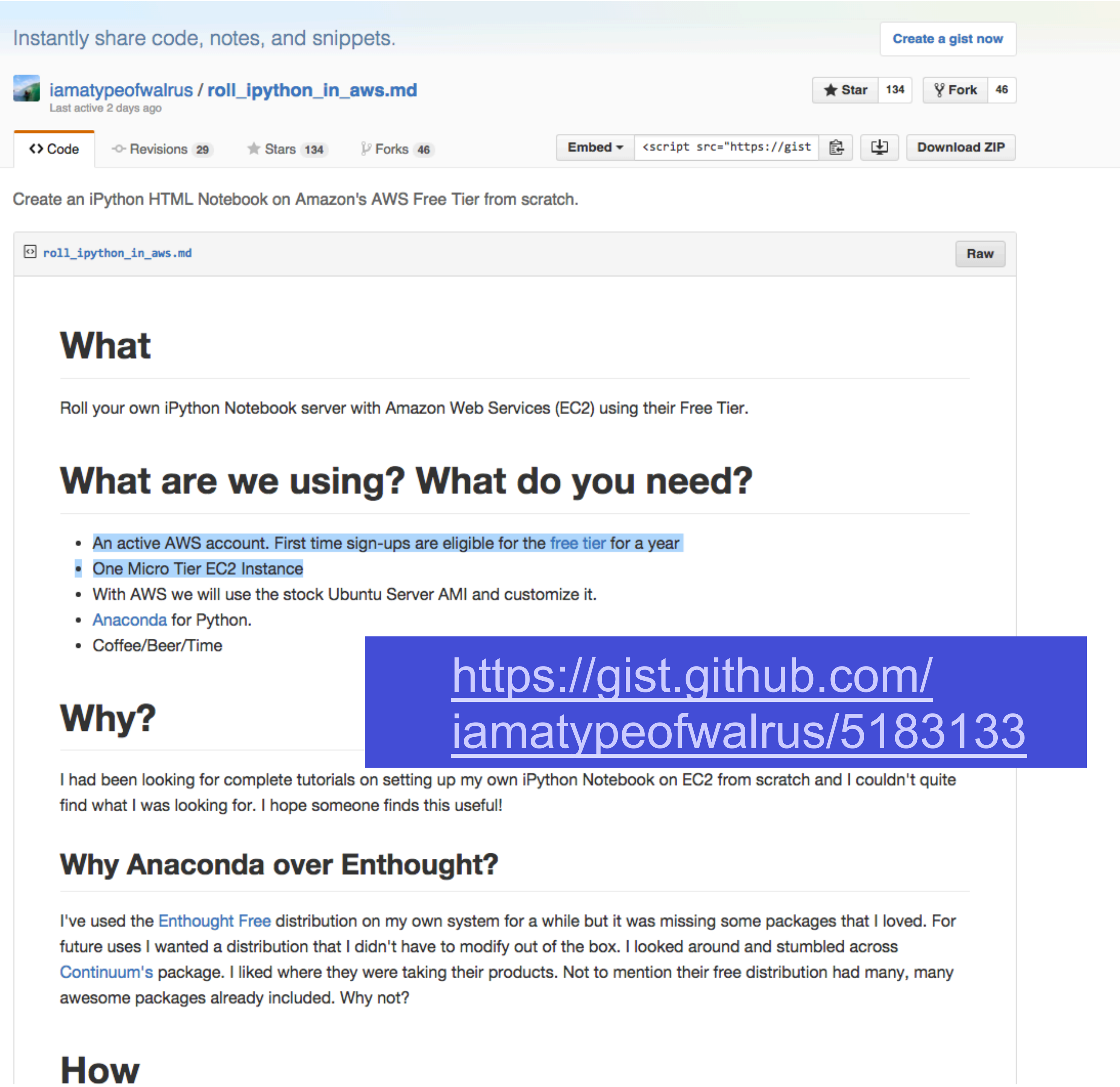
I have great interest in latest technologies, programming and web development. I know various programming languages but I like to program in HTML5, JavaScript, jQuery, Python and Java.
I am a gizmo freak and love to explore the world of technology.
My main interest is in Web Development and Data Science/Analysis.
To know more about me and connect with me check out my profile at http://about.me/impiyush

View my complete profile

Blog Archive

- ▼ 2015 (2)
  - ► March (1)
  - ▼ February (1)
    - Running an iPython Notebook Server on AWS - EC2 In...
- ► 2013 (7)
- ► 2012 (7)

Follow by Email



Create an iPython HTML Notebook on Amazon's AWS Free Tier from scratch.

What

Roll your own iPython Notebook server with Amazon Web Services (EC2) using their Free Tier.

What are we using? What do you need?

- An active AWS account. First time sign-ups are eligible for the free tier for a year
- One Micro Tier EC2 Instance
- With AWS we will use the stock Ubuntu Server AMI and customize it.
- Anaconda for Python.
- Coffee/Beer/Time

Why?

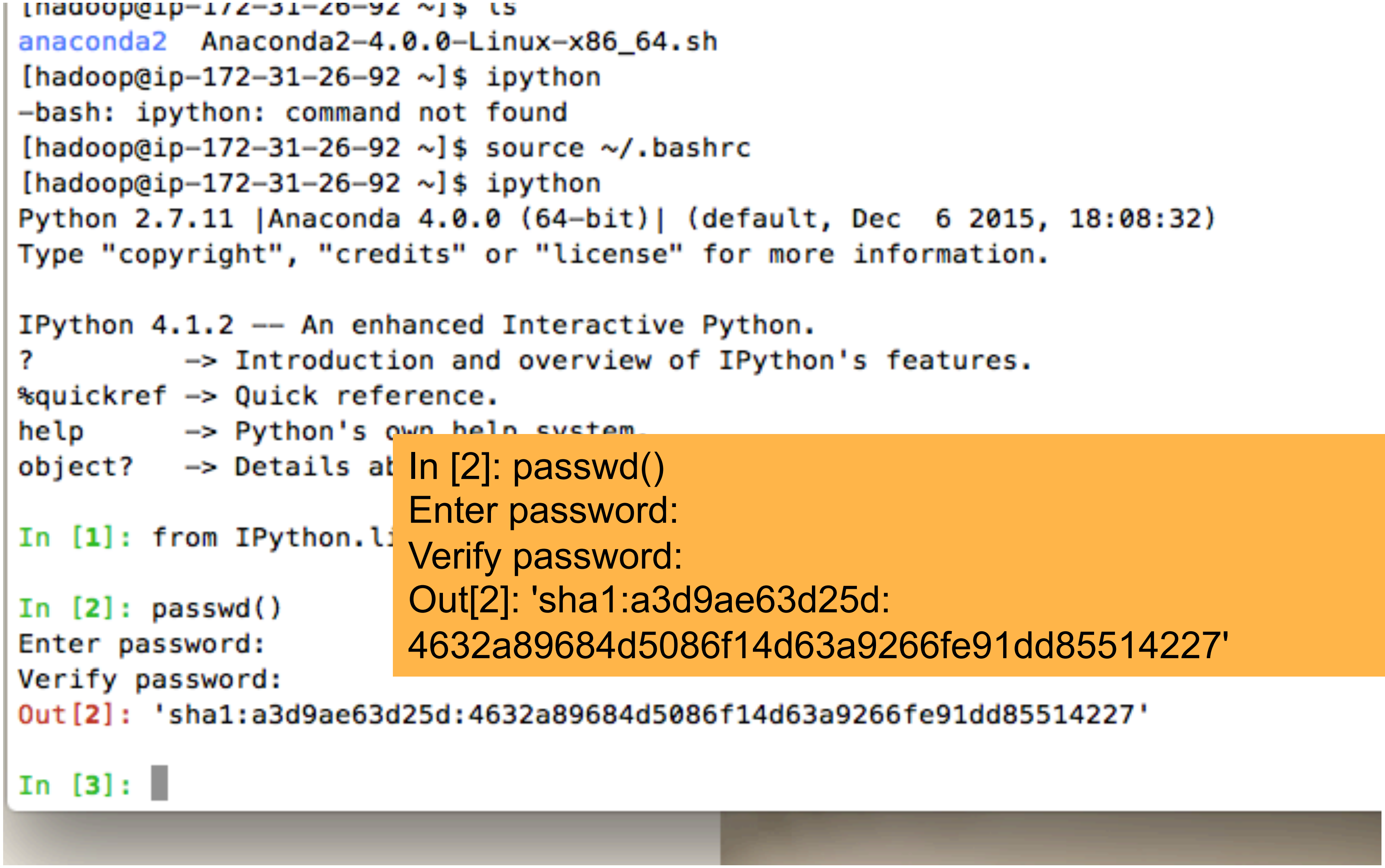
I had been looking for complete tutorials on setting up my own iPython Notebook on EC2 from scratch and I couldn't quite find what I was looking for. I hope someone finds this useful!

Why Anaconda over Enthought?

I've used the Enthought Free distribution on my own system for a while but it was missing some packages that I loved. For future uses I wanted a distribution that I didn't have to modify out of the box. I looked around and stumbled across Continuum's package. I liked where they were taking their products. Not to mention their free distribution had many, many awesome packages already included. Why not?

How

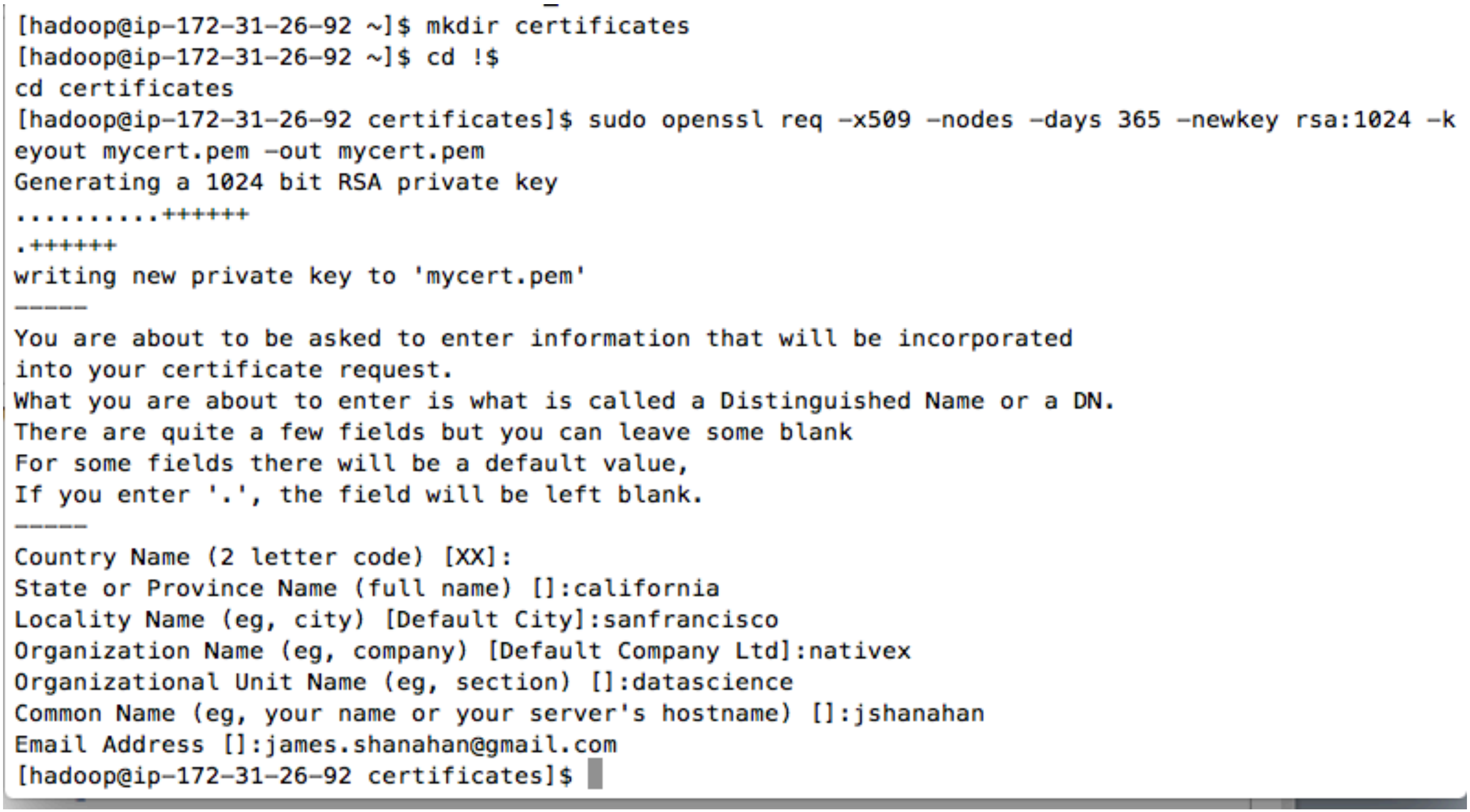
https://gist.github.com/iamatypeofwalrus/5183133



```
[hadoop@ip-172-31-26-92 ~]$ ls
anaconda2  Anaconda2-4.0.0-Linux-x86_64.sh
[hadoop@ip-172-31-26-92 ~]$ ipython
-bash: ipython: command not found
[hadoop@ip-172-31-26-92 ~]$ source ~/.bashrc
[hadoop@ip-172-31-26-92 ~]$ ipython
Python 2.7.11 |Anaconda 4.0.0 (64-bit)| (default, Dec  6 2015, 18:08:32)
Type "copyright", "credits" or "license" for more information.

IPython 4.1.2 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details ab

In [1]: from IPython.li

In [2]: passwd()
Enter password:
Verify password:
Out[2]: 'sha1:a3d9ae63d25d:4632a89684d5086f14d63a9266fe91dd85514227'

In [3]:
```

In [2]: passwd()
Enter password:
Verify password:
Out[2]: 'sha1:a3d9ae63d25d:
4632a89684d5086f14d63a9266fe91dd85514227'

sha1:a3d9ae63d25d:4632a89684d5086f14d63a9266fe91dd85514227



```
[hadoop@ip-172-31-26-92 ~]$ mkdir certificates
[hadoop@ip-172-31-26-92 ~]$ cd !$
cd certificates
[hadoop@ip-172-31-26-92 certificates]$ sudo openssl req -x509 -nodes -days 365 -newkey rsa:1024 -k
eyout mycert.pem -out mycert.pem
Generating a 1024 bit RSA private key
..........++++++
.++++++
writing new private key to 'mycert.pem'
-----
You are about to be asked to enter information that will be incorporated
into your certificate request.
What you are about to enter is what is called a Distinguished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [XX]:
State or Province Name (full name) []:california
Locality Name (eg, city) [Default City]:sanfrancisco
Organization Name (eg, company) [Default Company Ltd]:nativex
Organizational Unit Name (eg, section) []:datascience
Common Name (eg, your name or your server's hostname) []:jshanahan
Email Address []:james.shanahan@gmail.com
[hadoop@ip-172-31-26-92 certificates]$ 
```

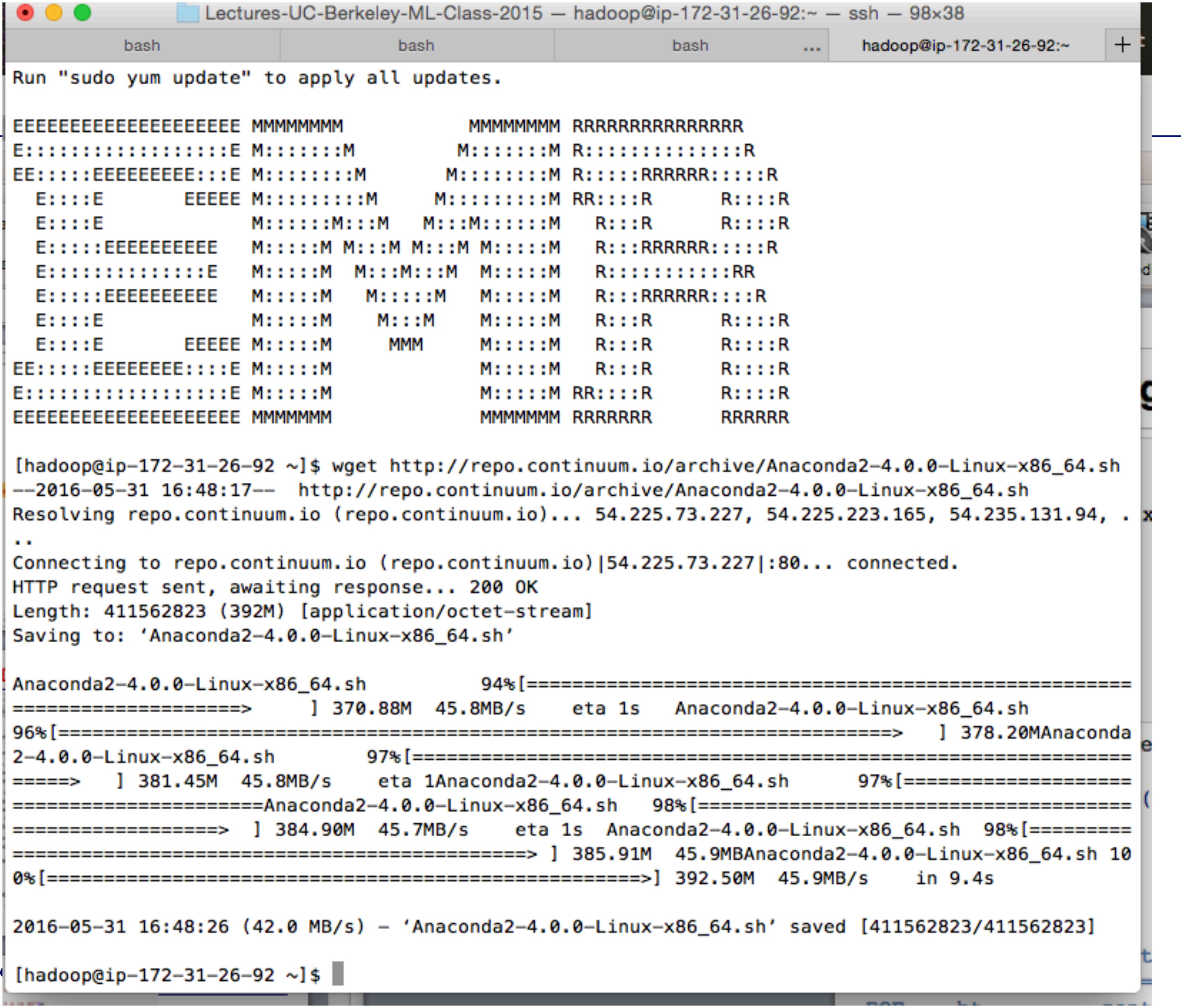
Install ipython and notebooks on Linux : on cluster namenode

Login in remotely top the name node of my cluster

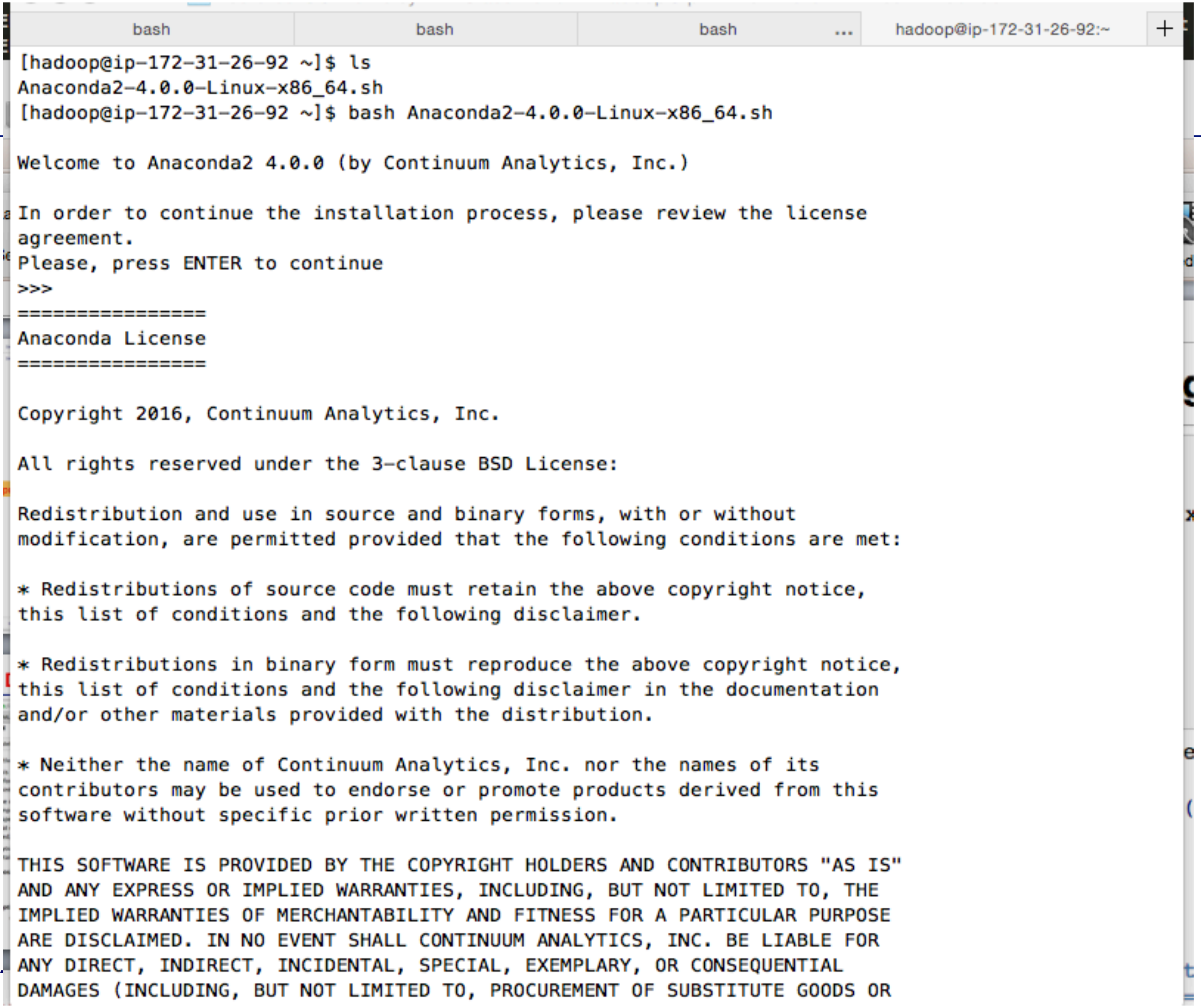
ssh -i ~/jimi-261-2016-Spring.pem hadoop@ec2-54-67-63-252.us-west-1.compute.amazonaws.com

- **Install ipython and notebooks via the command line using these two commands**
- ***wget http://repo.continuum.io/archive/Anaconda2-4.0.0-Linux-x86_64.sh***
- ***bash Anaconda2-4.0.0-Linux-x86_64.sh***

Or via your web browser PLUS command line


Lectures-UC-Berkeley-ML-Class-2015 — hadoop@ip-172-31-26-92:~ — ssh — 98×38
bash
bash
bash
hadoop@ip-172-31-26-92:~

```
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-26-92 ~]$ wget http://repo.continuum.io/archive/Anaconda2-4.0.0-Linux-x86_64.sh
--2016-05-31 16:48:17--  http://repo.continuum.io/archive/Anaconda2-4.0.0-Linux-x86_64.sh
Resolving repo.continuum.io (repo.continuum.io)... 54.225.73.227, 54.225.223.165, 54.235.131.94, .
..
Connecting to repo.continuum.io (repo.continuum.io)|54.225.73.227|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 411562823 (392M) [application/octet-stream]
Saving to: ‘Anaconda2-4.0.0-Linux-x86_64.sh’

Anaconda2-4.0.0-Linux-x86_64.sh          94%[==================================================
=====================>     ] 370.88M  45.8MB/s    eta 1s    Anaconda2-4.0.0-Linux-x86_64.sh
96%[=====================================================================>   ] 378.20MAnaconda
2-4.0.0-Linux-x86_64.sh          97%[==================================================================
=====>    ] 381.45M  45.8MB/s    eta 1Anaconda2-4.0.0-Linux-x86_64.sh      97%[===================
======================Anaconda2-4.0.0-Linux-x86_64.sh    98%[======================================
==================>   ] 384.90M  45.7MB/s    eta 1s   Anaconda2-4.0.0-Linux-x86_64.sh  98%[=========
=============================================> ] 385.91M  45.9MBAnaconda2-4.0.0-Linux-x86_64.sh 10
0%[=====================================================>] 392.50M  45.9MB/s    in 9.4s

2016-05-31 16:48:26 (42.0 MB/s) - ‘Anaconda2-4.0.0-Linux-x86_64.sh’ saved [411562823/411562823]

[hadoop@ip-172-31-26-92 ~]$ 
```



```
[hadoop@ip-172-31-26-92 ~]$ ls
Anaconda2-4.0.0-Linux-x86_64.sh
[hadoop@ip-172-31-26-92 ~]$ bash Anaconda2-4.0.0-Linux-x86_64.sh

Welcome to Anaconda2 4.0.0 (by Continuum Analytics, Inc.)

In order to continue the installation process, please review the license
agreement.
Please, press ENTER to continue
>>>
================
Anaconda License
================

Copyright 2016, Continuum Analytics, Inc.

All rights reserved under the 3-clause BSD License:

Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice,
this list of conditions and the following disclaimer.

* Redistributions in binary form must reproduce the above copyright notice,
this list of conditions and the following disclaimer in the documentation
and/or other materials provided with the distribution.

* Neither the name of Continuum Analytics, Inc. nor the names of its
contributors may be used to endorse or promote products derived from this
software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"
AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE
ARE DISCLAIMED. IN NO EVENT SHALL CONTINUUM ANALYTICS, INC. BE LIABLE FOR
ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL
DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
# Source global definitions
if [ -f /etc/bashrc ]; then
  . /etc/bashrc
fi

# set the default region for the AWS CLI
export AWS_DEFAULT_REGION=$(curl --retry 5 --silent --connect-timeout 2
http://169.254.169.254/latest/dynamic/instance-identity/document | grep
region | awk -F\" '{print $4}')
export JAVA_HOME=/etc/alternatives/jre
# added by Anaconda2 4.0.0 installer
export PATH="/home/hadoop/anaconda2/bin:$PATH"
```