L07: Hive, Spark SQL

ANLY 502: Massive Data Fundamentals

Simson Garfinkel & Ghaleb Abdulla

March 14, 2016

Outline for today's class

New Mac!

Preview of the next 3 weeks

- PS04
- Midterm
- PS05

Student Presentations

Hive

SparkSQL

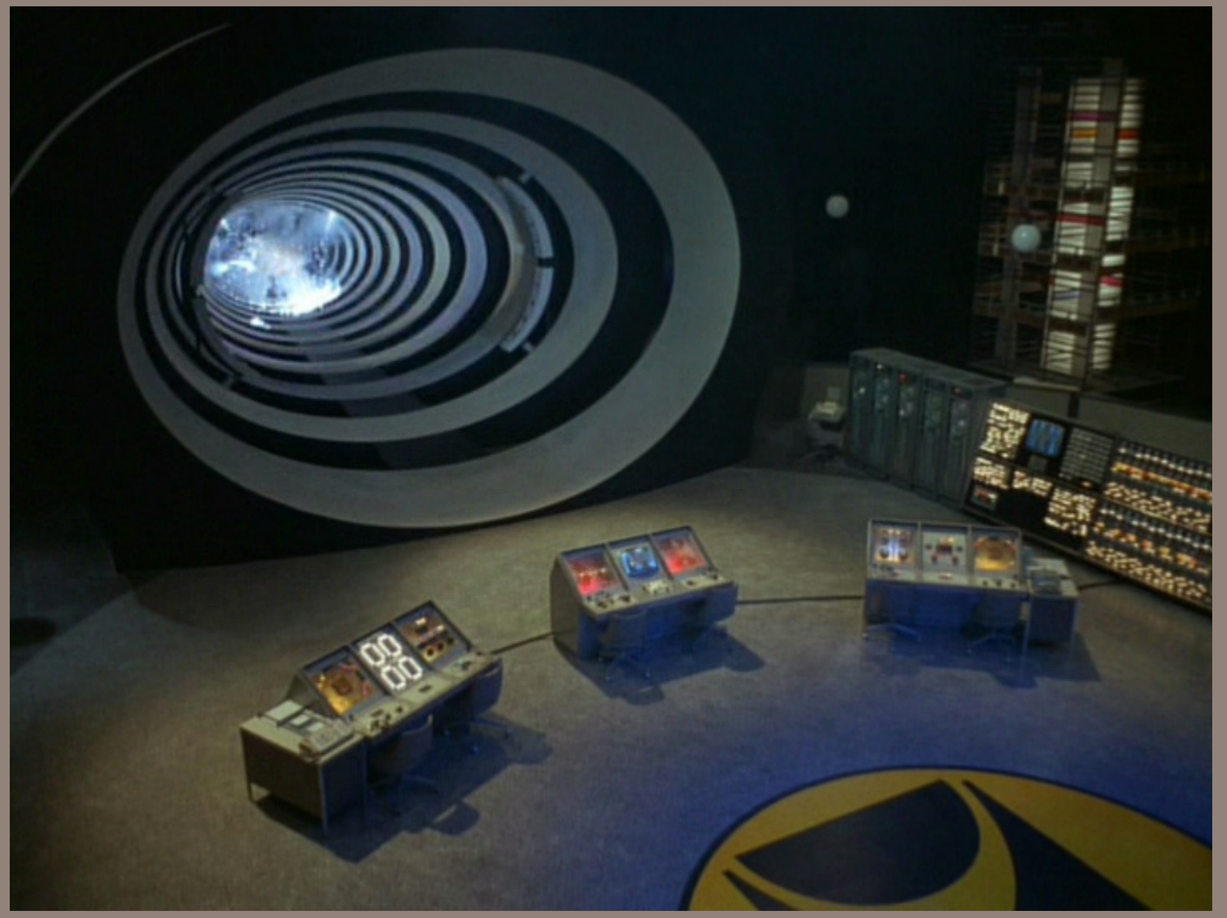
If time: Privacy and De-identification

I bought a new Airbook!

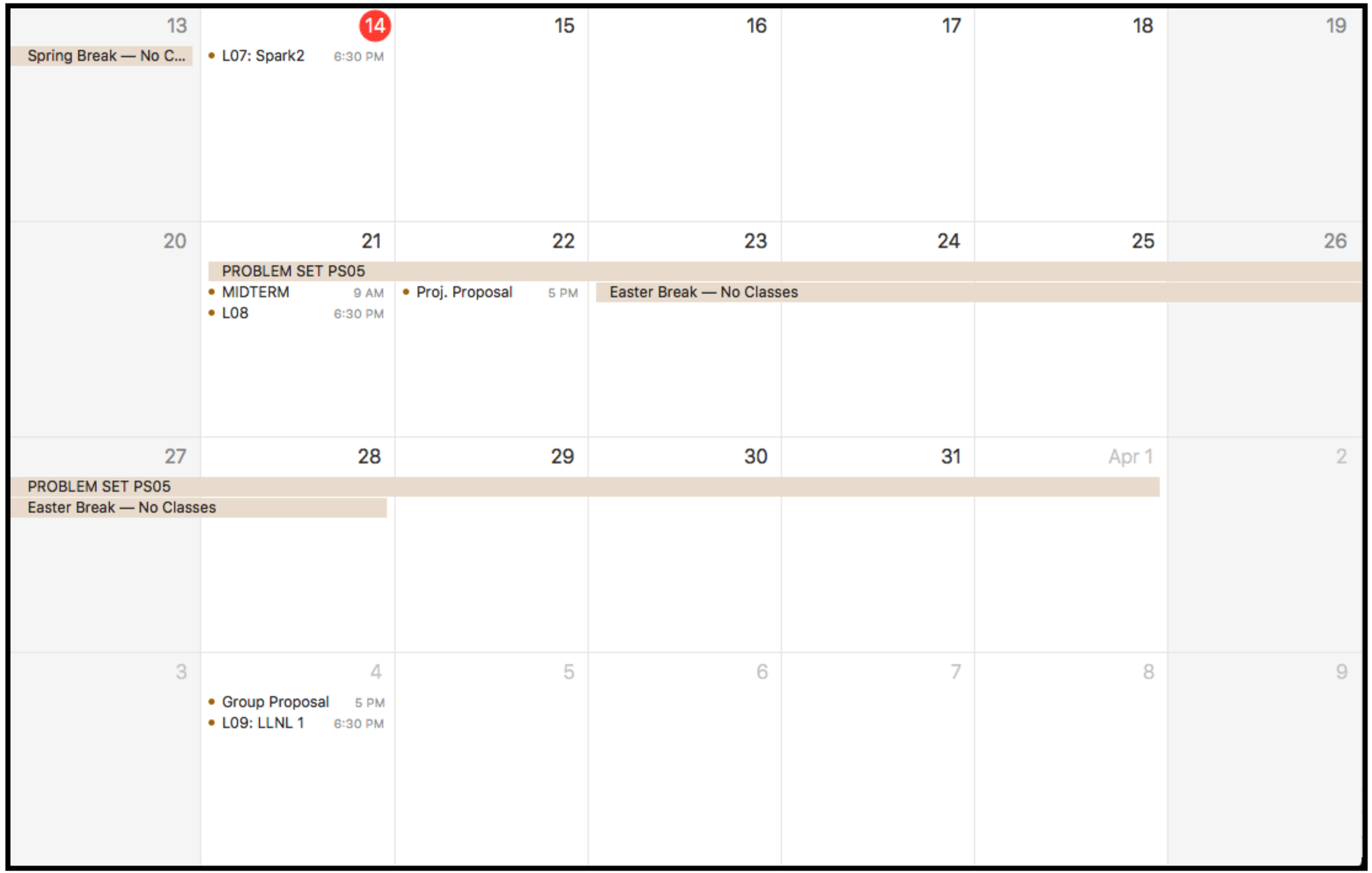
2011 Mac Airpot

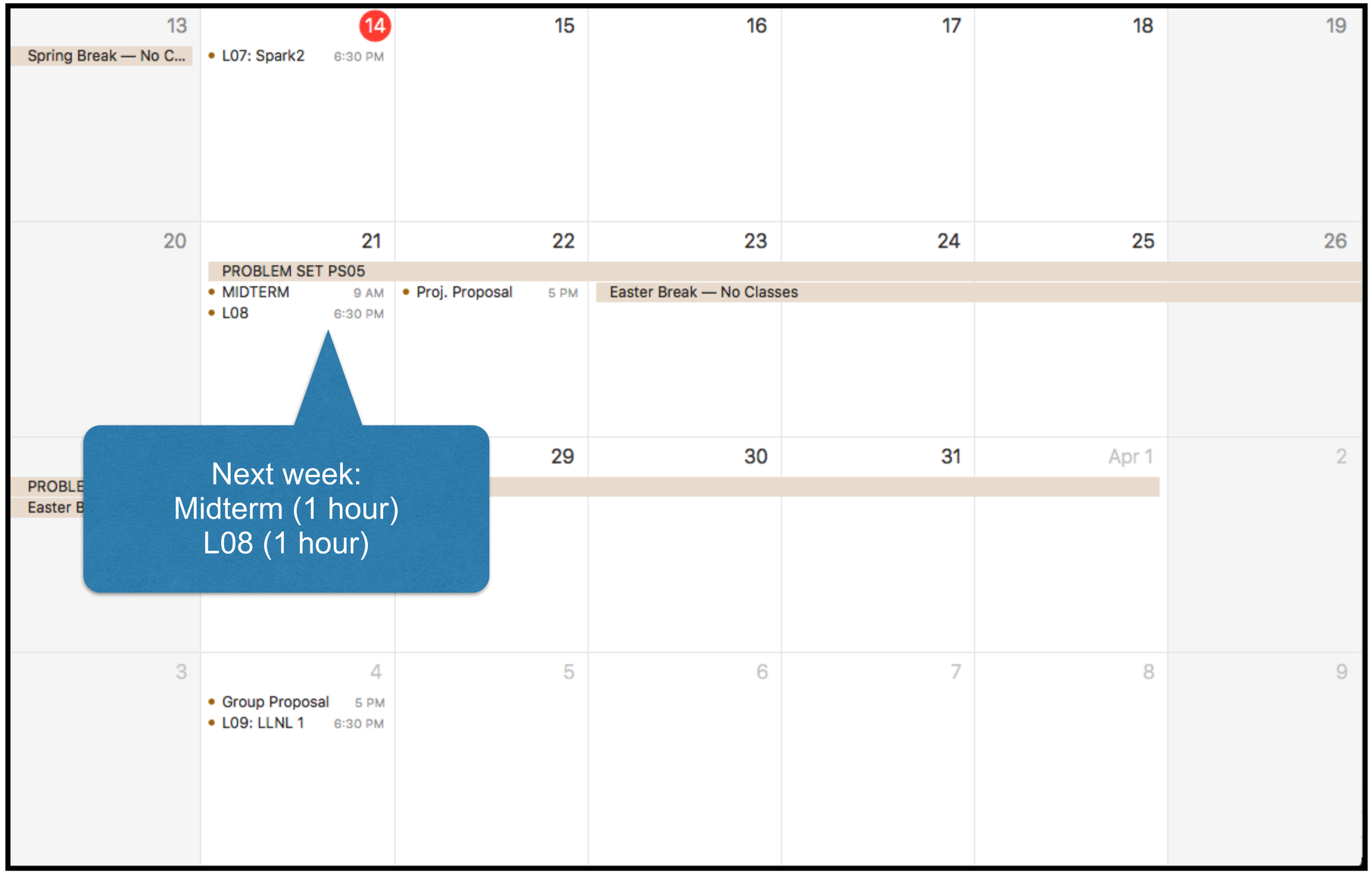
2015 Mac Airpot

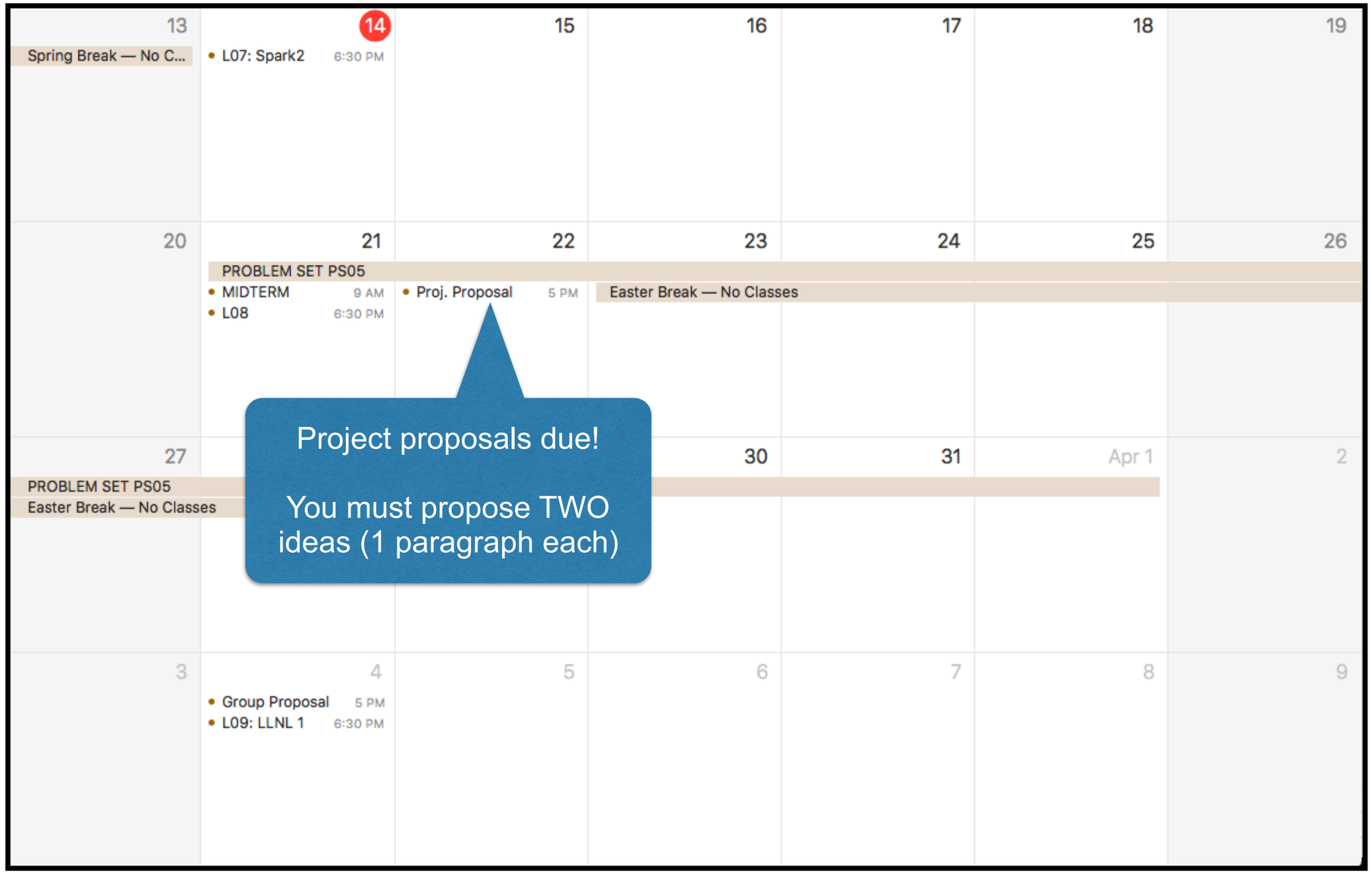
- **Faster SSD**
- **Faster RAM**
- **Faster Graphics**
- **Better Battery**
- **Slower CPU!**

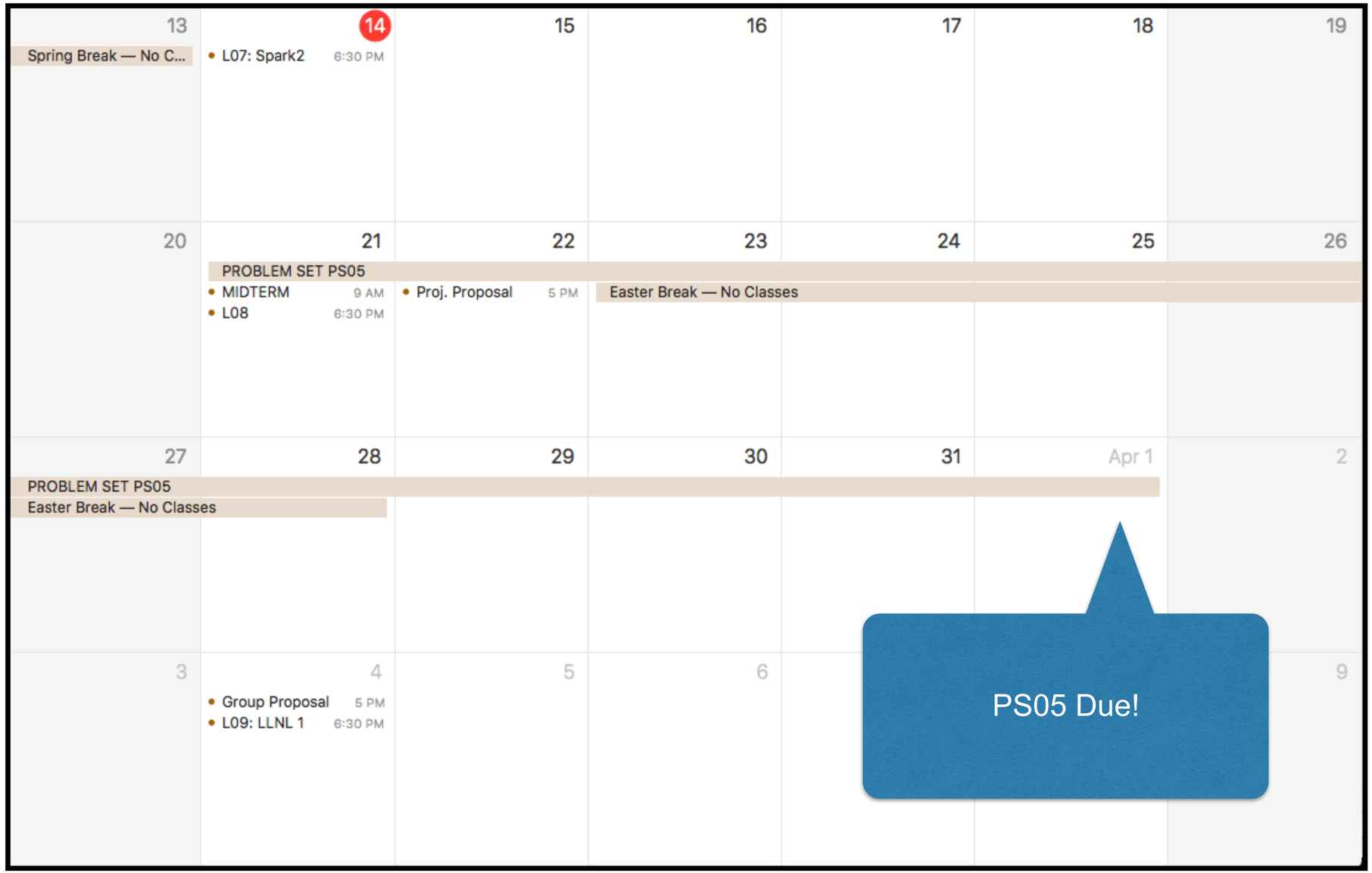


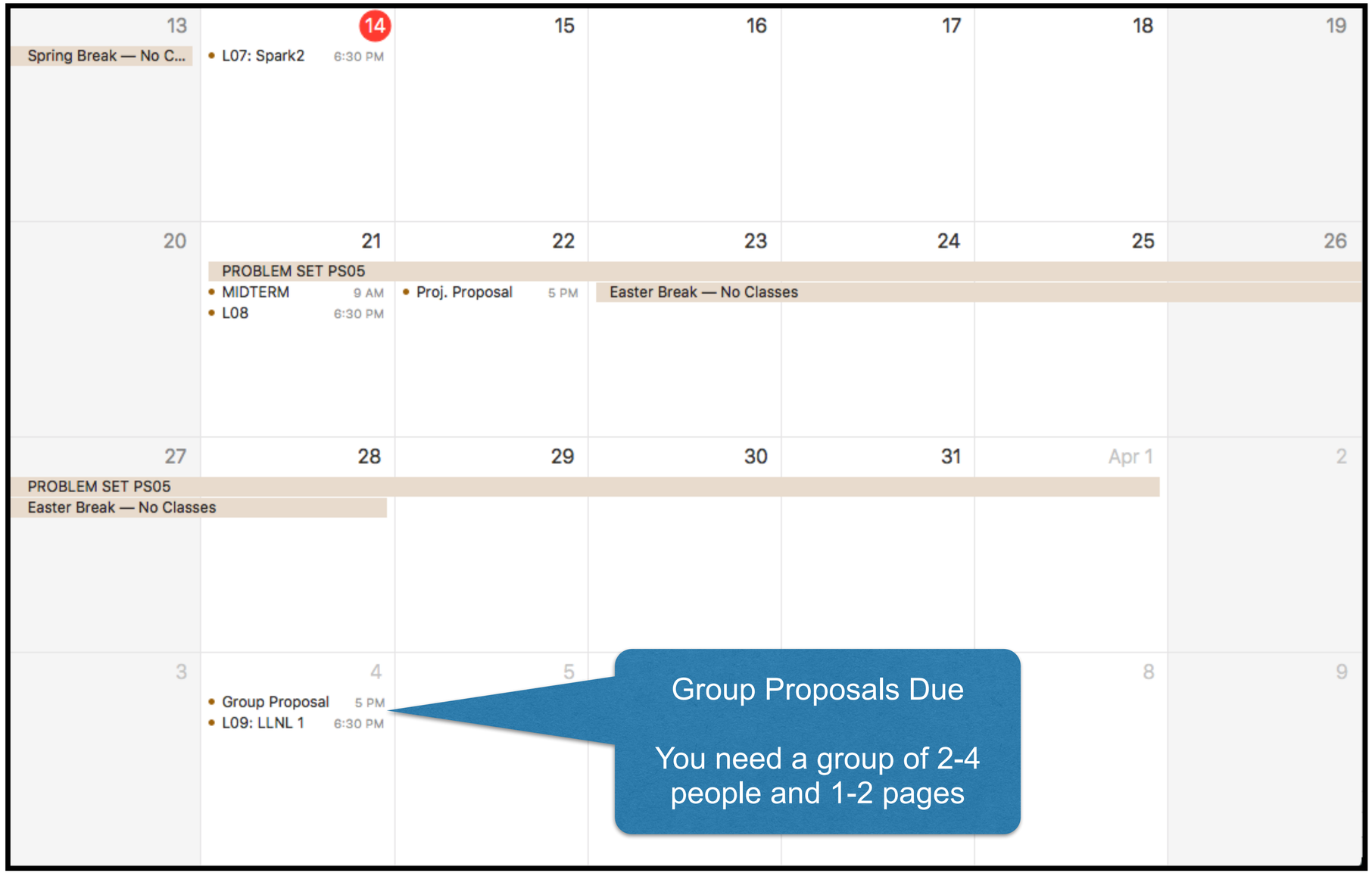
The Future

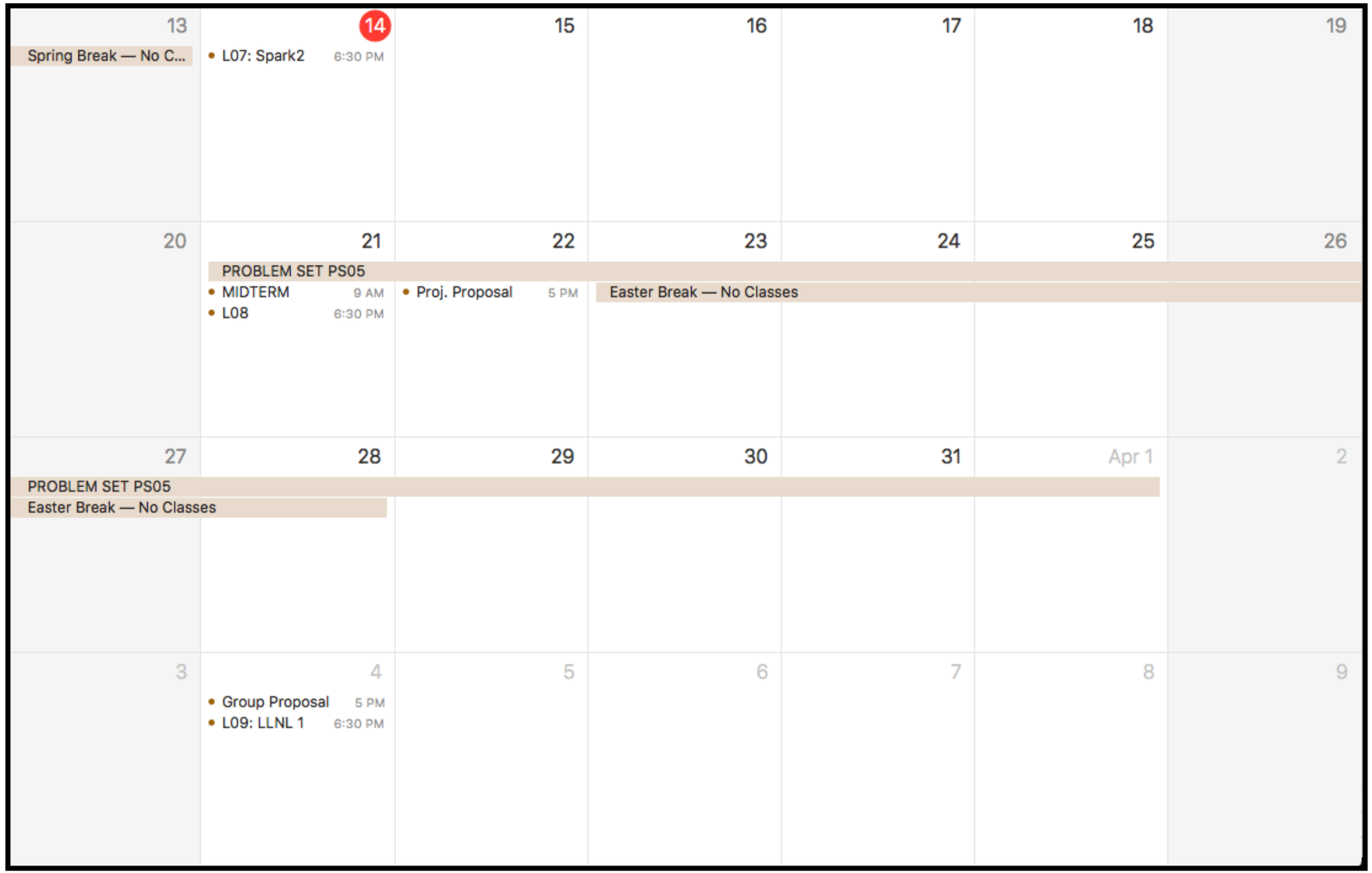

13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
Easter Break — No Classes
24
25
26
27
PROBLEM SET PS05
Easter Break — No Classes
28
29
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
6
7
8
9


13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
24
25
Easter Break — No Classes
26
PROBLE
Easter B
29
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
6
7
8
9
Next week:
Midterm (1 hour)
L08 (1 hour)


13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
Easter Break — No Classes
24
25
26
27
PROBLEM SET PS05
Easter Break — No Classes
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
6
7
8
9
Project proposals due!
You must propose TWO ideas (1 paragraph each)


13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
Easter Break — No Classes
24
25
26
27
PROBLEM SET PS05
Easter Break — No Classes
28
29
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
6
9
PS05 Due!


13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
Easter Break — No Classes
24
25
26
27
PROBLEM SET PS05
Easter Break — No Classes
28
29
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
8
9
Group Proposals Due
You need a group of 2-4 people and 1-2 pages


13
Spring Break — No C...
14
L07: Spark2 6:30 PM
15
16
17
18
19
20
21
PROBLEM SET PS05
MIDTERM 9 AM
L08 6:30 PM
22
Proj. Proposal 5 PM
23
Easter Break — No Classes
24
25
26
27
PROBLEM SET PS05
Easter Break — No Classes
28
29
30
31
Apr 1
2
3
4
Group Proposal 5 PM
L09: LLNL 1 6:30 PM
5
6
7
8
9

Any issues with PS04 — Pig and Spark

Part 1 — Word Count with Pig (3 problems)

Part 2 — Analyzing ForensicsWiki logs with Pig (2 problems)

Part 3 — Spark — WordCount and Wikipedia

Extra Credit: Maxmind join with broadcast variables and the full IP address space.

mid·term

/ˈmidˌtərm/

noun

noun: **midterm**; plural noun: **midterms**; noun: **mid-term**; plural noun: **mid-terms**

the middle of a period of office, an academic term, or a pregnancy.
"Nixon resigned **in midterm**"

- NORTH AMERICAN
  an exam in the middle of an academic term.

Translate midterm to Choose language

Use over time for: midterm

Midterm!

March 21

Midterm Study Guide — March 21, 2016

Python — classes, generators, lists, filters

What's Massive Data

- Technology trends. Scientific computing vs. commercial computing. AWS EC2, EMR

Map Reduce

- mapper, combiner, shuffle, reducer, partitioner, counters
- Impact of # of partitions
- Java vs. Hadoop Streaming vs. mrjob — understand what each does, and the advantages of each.

Storage

- Google File System, HDFS & S3

Pig & Hive

Spark

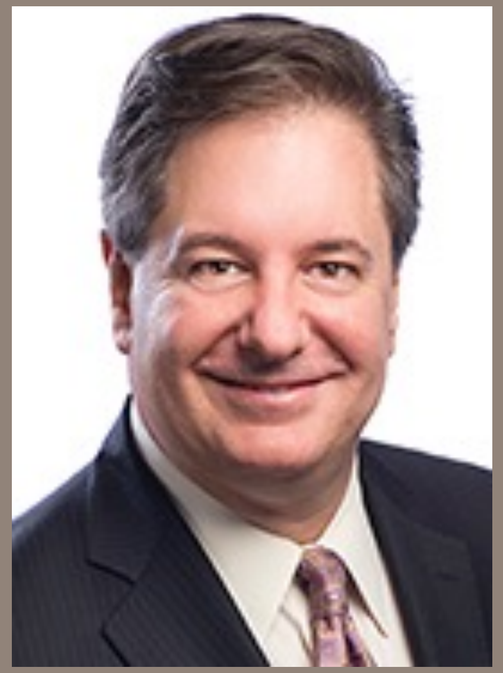
- RDDs, Key/value RDDs, Broadcast variables, Counters, Spark SQL

Format: Multiple Choice, Some debugging.

https://pixabay.com/en/learn-know-students-chalk-children-916677/

Student Presentations

Yu YU	Program	Hadoop Aggregate Package
Zhengning Li	Program	Alluxio a memory speed virtual distributed storage system
Jiayao Wang	Program	CloudBurst: highly sensitive read mapping with MapReduce
Nathan Hauke	Paper	Meta-MapReduce for scalable data mining



JAMES H. KOENIG
Of Counsel, Litigation Department
New York
T 1(212) 318-6005
F 1(212) 303-7005
jimkoenig@paulhastings.com

EXPERIENCE

Mr. Koenig has worked with an array of global clients on projects involving more than 125 countries.
Over the last five years, he has represented 35% of the companies currently in the Fortune 100.
Recent projects relate to:

- **global privacy compliance;**
- **security/cybersecurity and breach response;**
- **new data uses relating to the cloud, data analytics ("Big Data"), mobile platforms, digital marketing, social media, de-identification and the Internet of Things; and**
- **regulatory investigations and enforcement actions or class-action litigations relating to privacy and cybersecurity practices.**

Jim Koenig

Privacy and de-identification

Accessing the Hadoop Web Interface from EMR

We used the web interface with the Cloudera VM

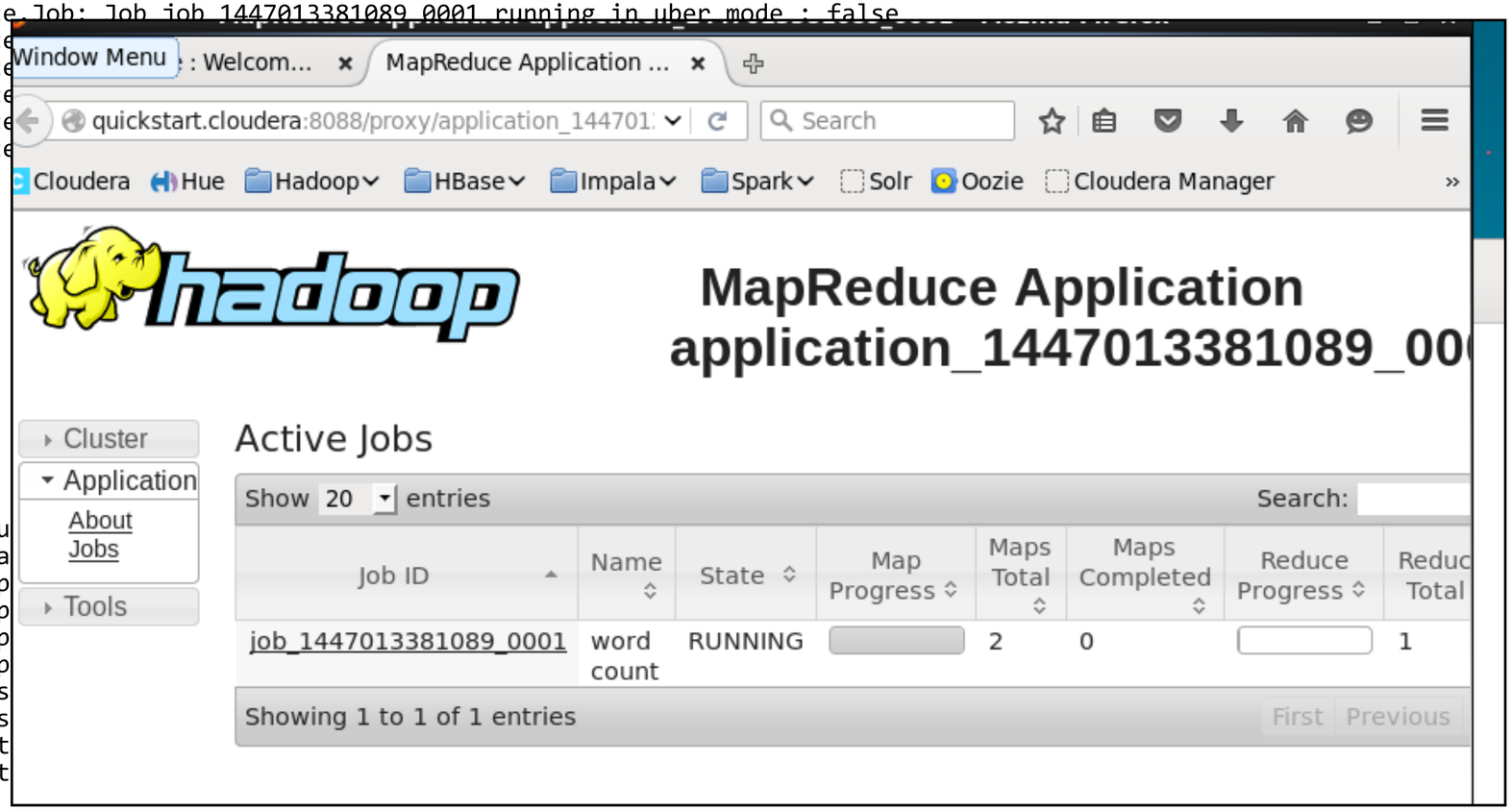
```
$ hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input/ /user/cloudera/wordcount/output/
15/11/08 13:57:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/11/08 13:57:02 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
with ToolRunner to remedy this.
15/11/08 13:57:02 INFO input.FileInputFormat: Total input paths to process : 2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: number of splits:2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1447013381089_0001
15/11/08 13:57:03 INFO impl.YarnClientImpl: Submitted application application_1447013381089_0001
15/11/08 13:57:03 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1447013381089_0001/
15/11/08 13:57:03 INFO mapreduce.Job: Running job: job_1447013381089_0001
15/11/08 13:57:13 INFO mapreduce.Job: Job job_1447013381089_0001 running in uber mode : false
15/11/08 13:57:13 INFO mapreduce.Job:  map 0% reduce 0%
15/11/08 13:57:24 INFO mapreduce.Job:  map 100% reduce 0%
15/11/08 13:57:30 INFO mapreduce.Job:  map 100% reduce 100%
15/11/08 13:57:31 INFO mapreduce.Job: Job job_1447013381089_0001 completed successfully
15/11/08 13:57:31 INFO mapreduce.Job: Counters: 49
      File System Counters
                  FILE: Number of bytes read=61
                  FILE: Number of bytes written=332053
                  FILE: Number of read operations=0
                  FILE: Number of large read operations=0
                  FILE: Number of write operations=0
                  HDFS: Number of bytes read=298
                  HDFS: Number of bytes written=26
                  HDFS: Number of read operations=9
                  HDFS: Number of large read operations=0
                  HDFS: Number of write operations=2
      Job Counters
                  Launched map tasks=2
                  Launched reduce tasks=1
                  Data-local map tasks=2
                  Total time spent by all maps in occupied slots (ms)=17757
                  Total time spent by all reduces in occupied slots (ms)=4511
                  Total time spent by all map tasks (ms)=17757
                  Total time spent by all reduce tasks (ms)=4511
                  Total vcore-seconds taken by all map tasks=17757
                  Total vcore-seconds taken by all reduce tasks=4511
                  Total megabyte-seconds taken by all map tasks=18183168
                  Total megabyte-seconds taken by all reduce tasks=4619264
```

We used the web interface with the Cloudera VM

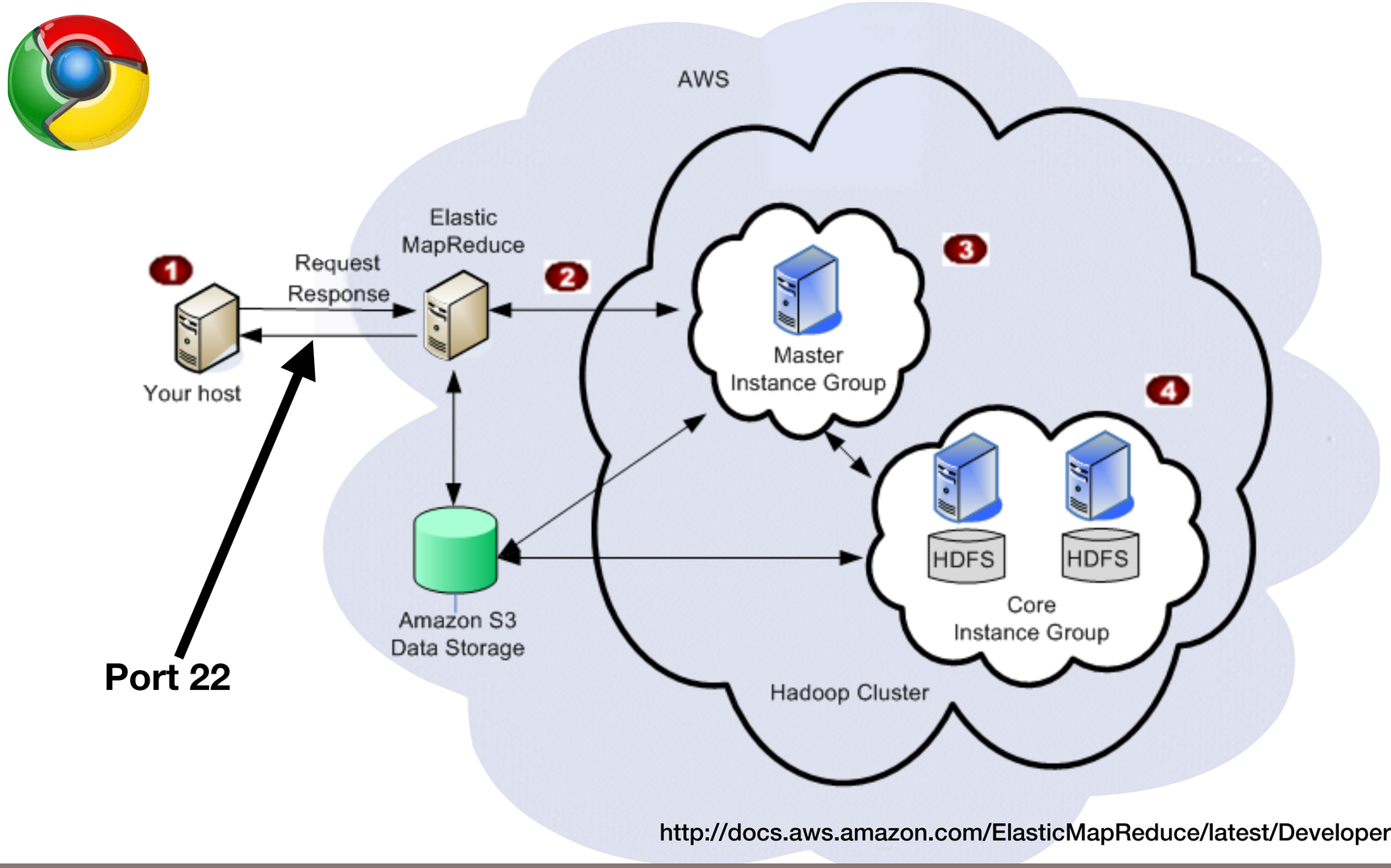
```
$ hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input/ /user/cloudera/wordcount/output/
15/11/08 13:57:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/11/08 13:57:02 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
with ToolRunner to remedy this.
15/11/08 13:57:02 INFO input.FileInputFormat: Total input paths to process : 2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: number of splits:2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1447013381089_0001
15/11/08 13:57:03 INFO impl.YarnClientImpl: Submitted application application_1447013381089_0001
15/11/08 13:57:03 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1447013381089_0001/
15/11/08 13:57:03 INFO mapreduce.Job: Running job: job_1447013381089_0001
15/11/08 13:57:13 INFO mapreduce.Job: Job job_1447013381089_0001 running in uber mode : false
15/11/08 13:57:13 INFO mapreduce.Job:  map 0% reduce 0%
15/11/08 13:57:24 INFO mapreduce.Job:  map 100% reduce 0%
15/11/08 13:57:30 INFO mapreduce.Job:  map 100% reduce 100%
15/11/08 13:57:31 INFO mapreduce.Job: Job job_1447013381089_0001 completed successfully
15/11/08 13:57:31 INFO mapreduce.Job: Counters: 49
      File System Counters
                  FILE: Number of bytes read=61
                  FILE: Number of bytes written=332053
                  FILE: Number of read operations=0
                  FILE: Number of large read operations=0
                  FILE: Number of write operations=0
                  HDFS: Number of bytes read=298
                  HDFS: Number of bytes written=26
                  HDFS: Number of read operations=9
                  HDFS: Number of large read operations=0
                  HDFS: Number of write operations=2
      Job Counters
                  Launched map tasks=2
                  Launched reduce tasks=1
                  Data-local map tasks=2
                  Total time spent by all maps in occupied slots (ms)=17757
                  Total time spent by all reduces in occupied slots (ms)=4511
                  Total time spent by all map tasks (ms)=17757
                  Total time spent by all reduce tasks (ms)=4511
                  Total vcore-seconds taken by all map tasks=17757
                  Total vcore-seconds taken by all reduce tasks=4511
                  Total megabyte-seconds taken by all map tasks=18183168
                  Total megabyte-seconds taken by all reduce tasks=4619264
```

We used the web interface with the Cloudera VM

```
$ hadoop jar wordcount.jar WordCount /user/cloudera/wordcount/input/ /user/cloudera/wordcount/output/
15/11/08 13:57:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/11/08 13:57:02 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application
with ToolRunner to remedy this.
15/11/08 13:57:02 INFO input.FileInputFormat: Total input paths to process : 2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: number of splits:2
15/11/08 13:57:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1447013381089_0001
15/11/08 13:57:03 INFO impl.YarnClientImpl: Submitted application application_1447013381089_0001
15/11/08 13:57:03 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1447013381089_0001/
15/11/08 13:57:03 INFO mapreduce.Job: Running job: job_1447013381089_0001
15/11/08 13:57:13 INFO mapreduce.Job: Job job_1447013381089_0001 running in uber mode : false
15/11/08 13:57:13 INFO mapreduce
15/11/08 13:57:24 INFO mapreduce
15/11/08 13:57:30 INFO mapreduce
15/11/08 13:57:31 INFO mapreduce
15/11/08 13:57:31 INFO mapreduce
      File System Counters
                  FILE: Number
                  FILE: Number
                  FILE: Number
                  FILE: Number
                  FILE: Number
                  HDFS: Number
                  HDFS: Number
                  HDFS: Number
                  HDFS: Number
                  HDFS: Number
      Job Counters
                  Launched map
                  Launched redu
                  Data-local ma
                  Total time sp
                  Total time sp
                  Total time sp
                  Total time sp
                  Total vcore-s
                  Total vcore-s
                  Total megabyt
                  Total megabyt
```



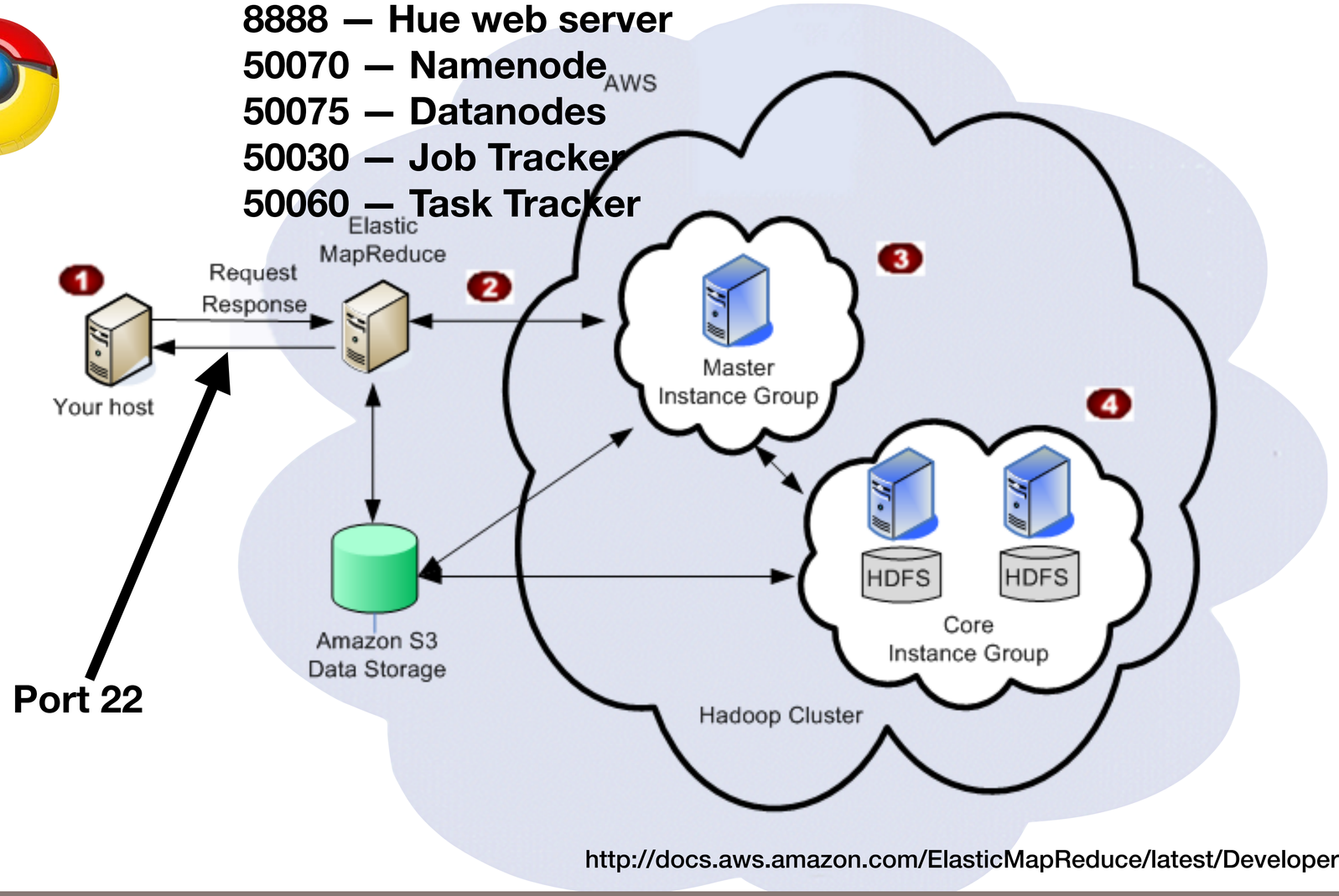
We communicate with EMR over port 22



http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-manage-resize.html

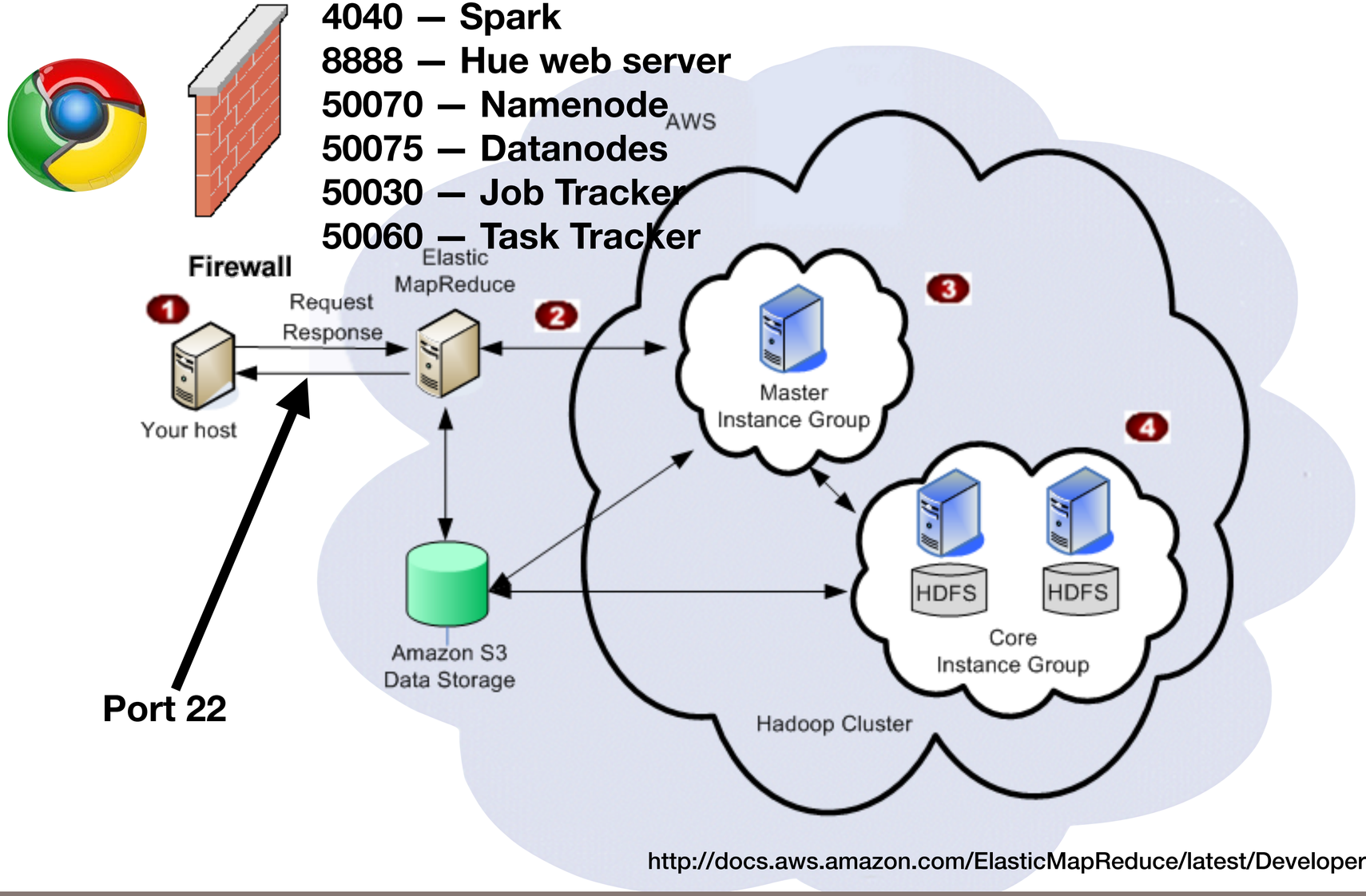
We communicate with EMR over port 22

4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker



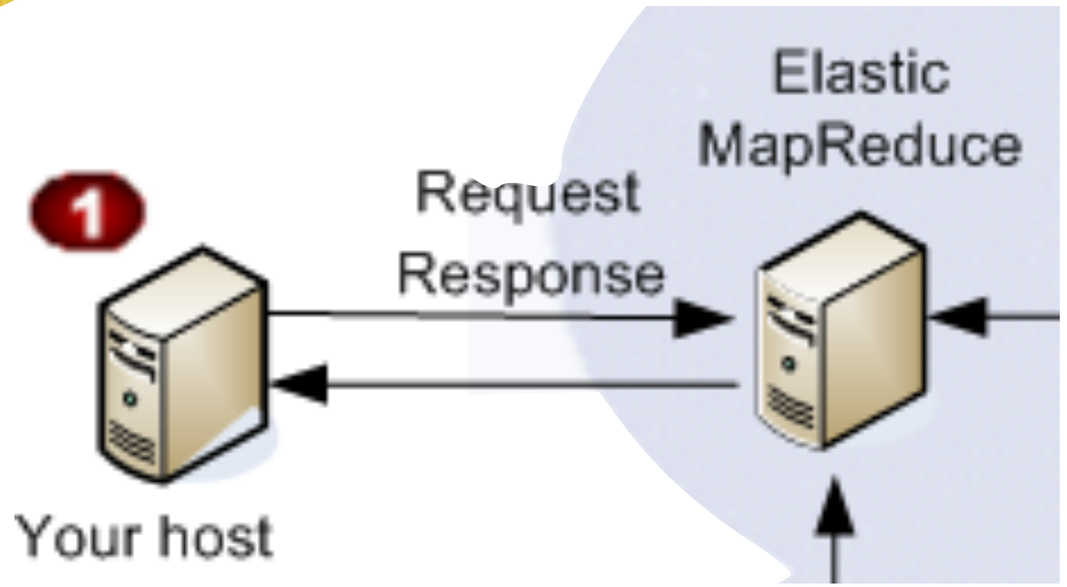
http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-manage-resize.html

We communicate with EMR over port 22

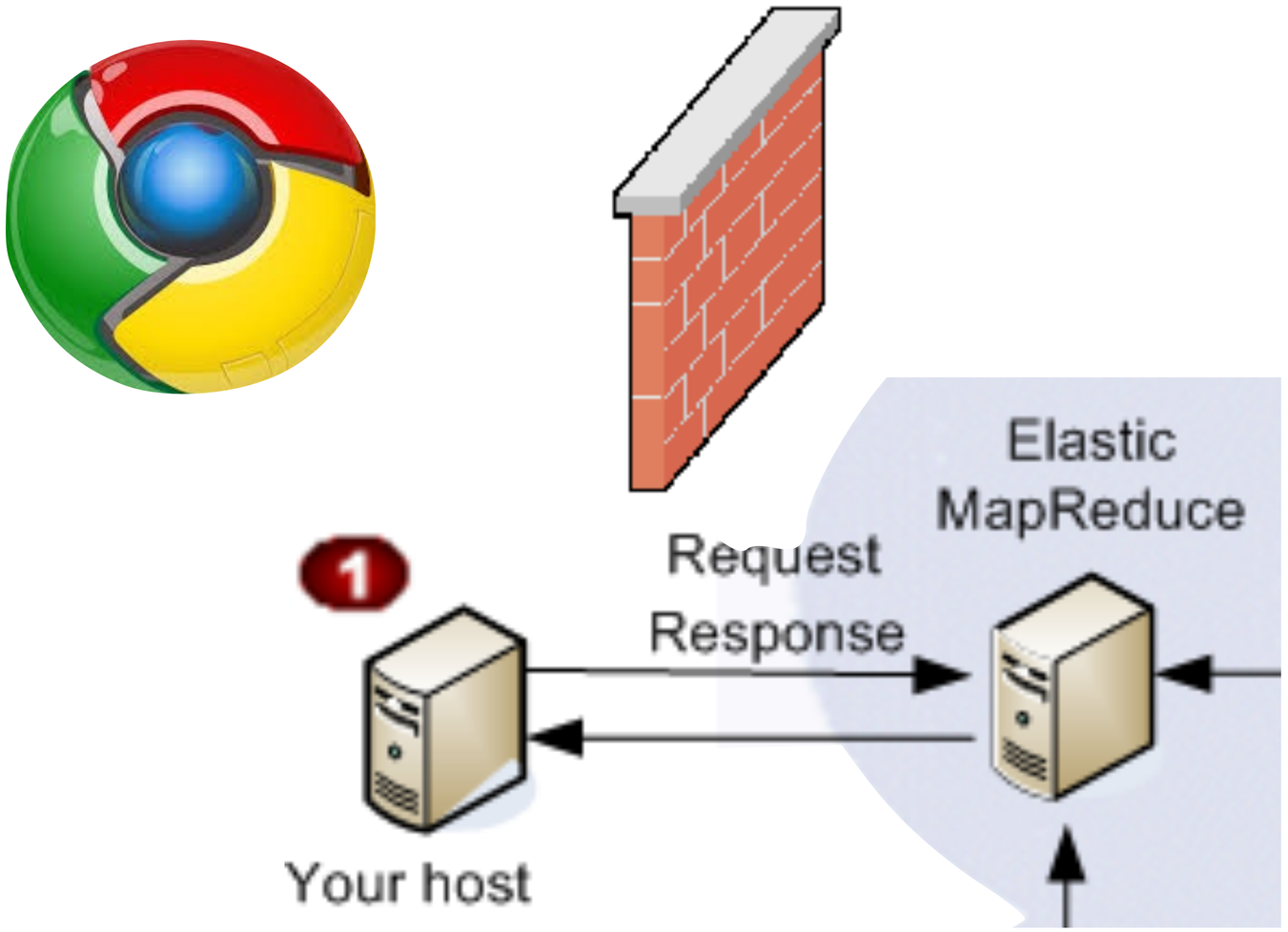


http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-manage-resize.html

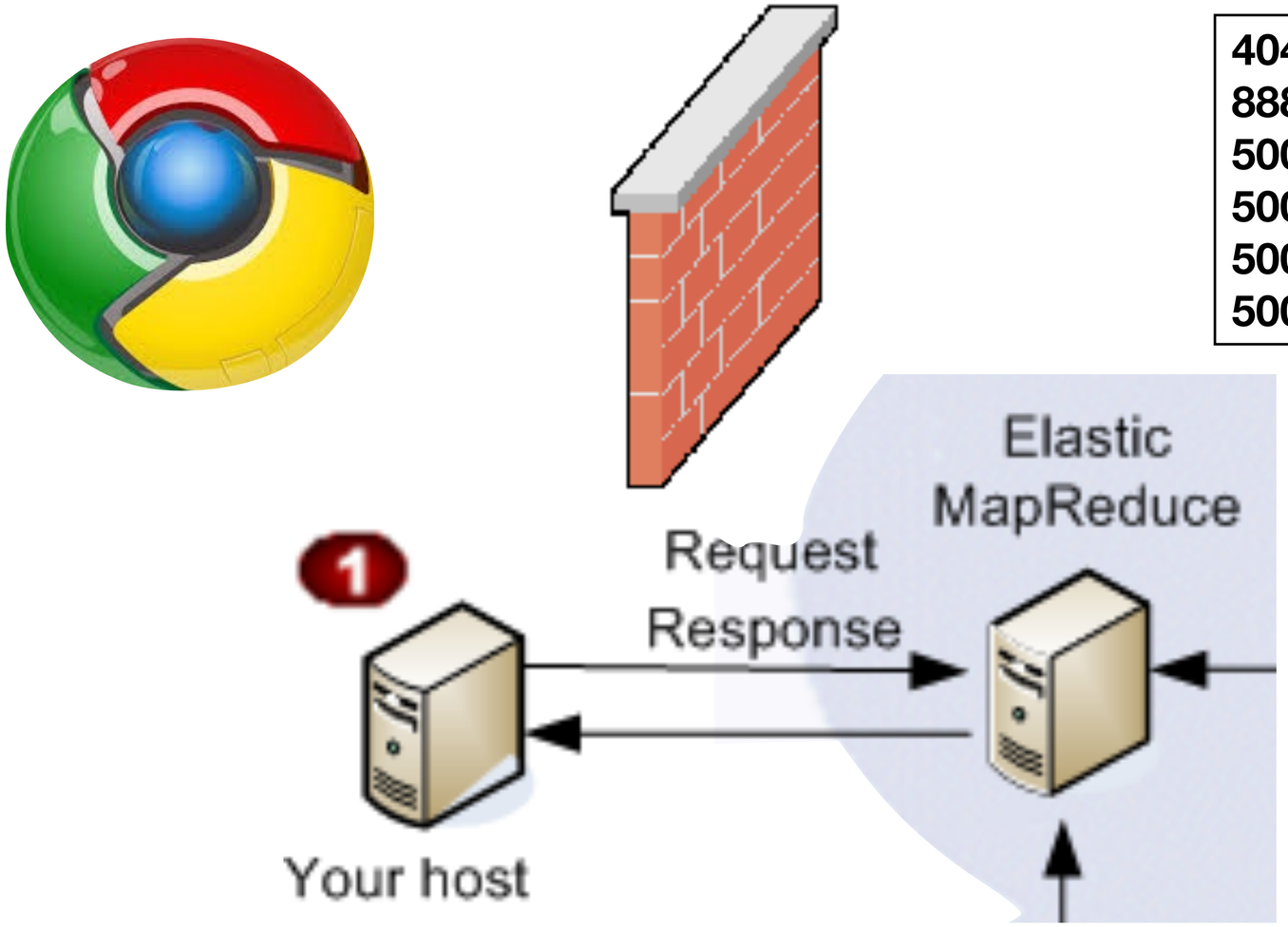
We use SSH to proxy the web connections



We use SSH to proxy the web connections

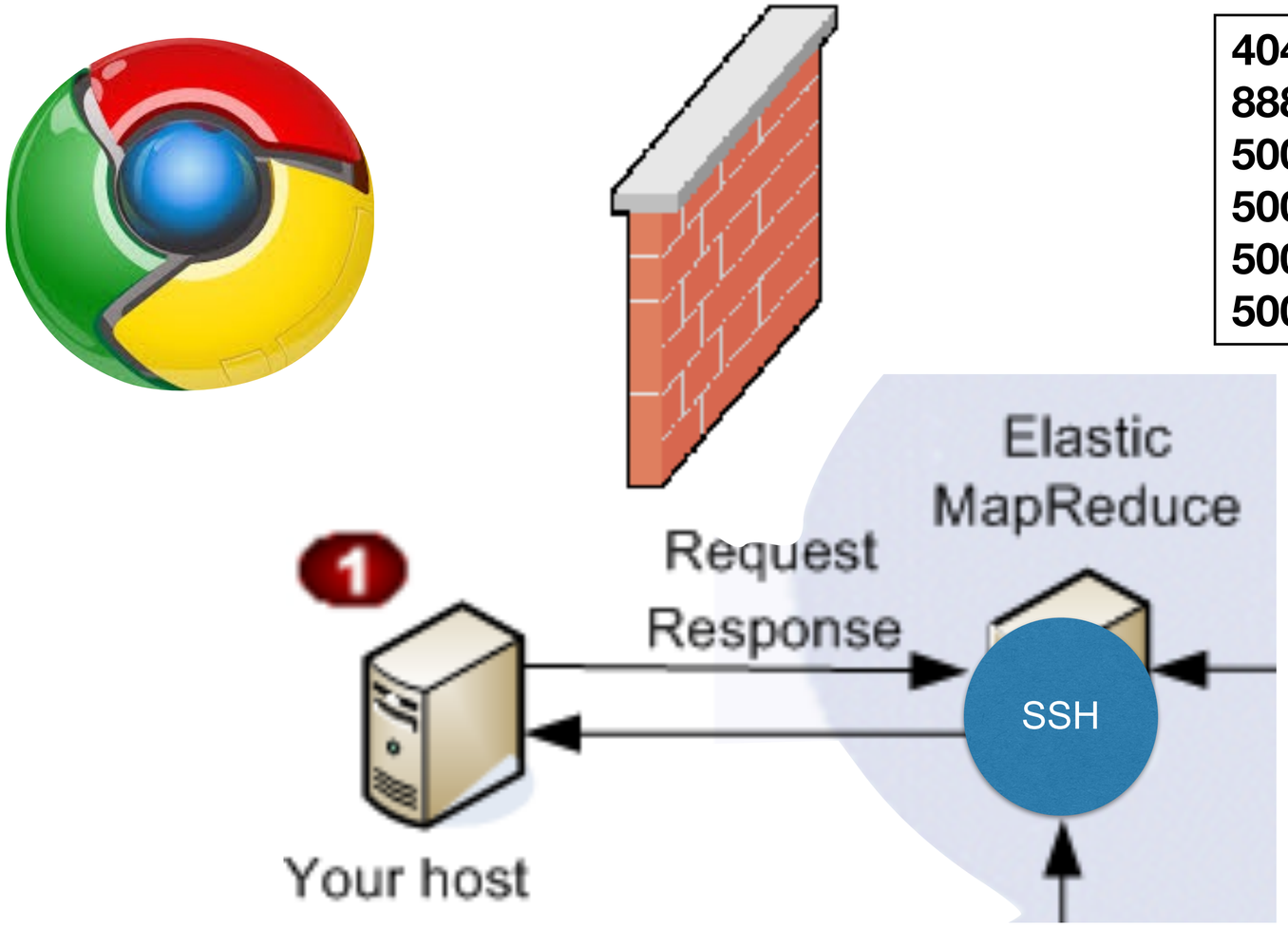


We use SSH to proxy the web connections



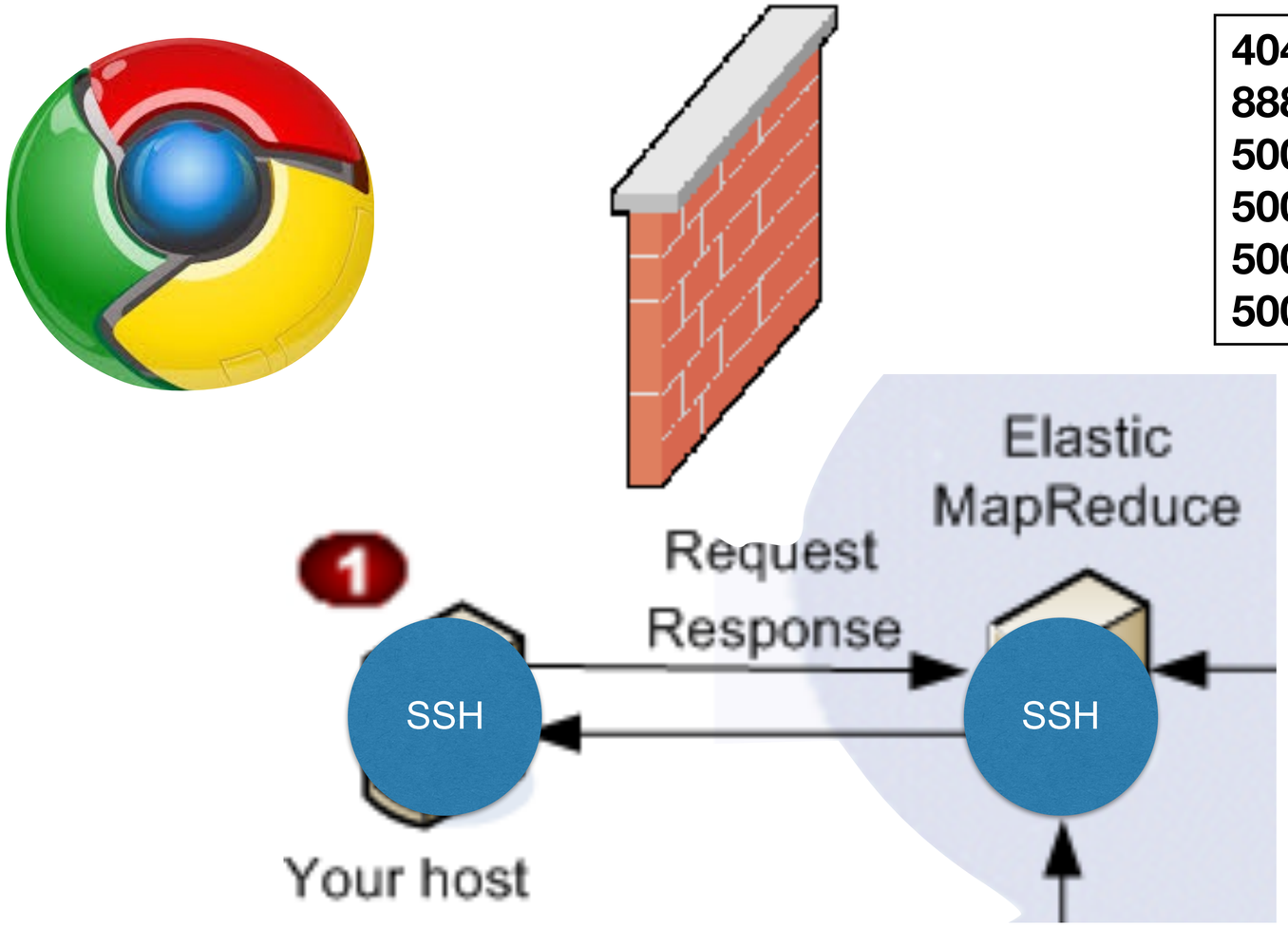
4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



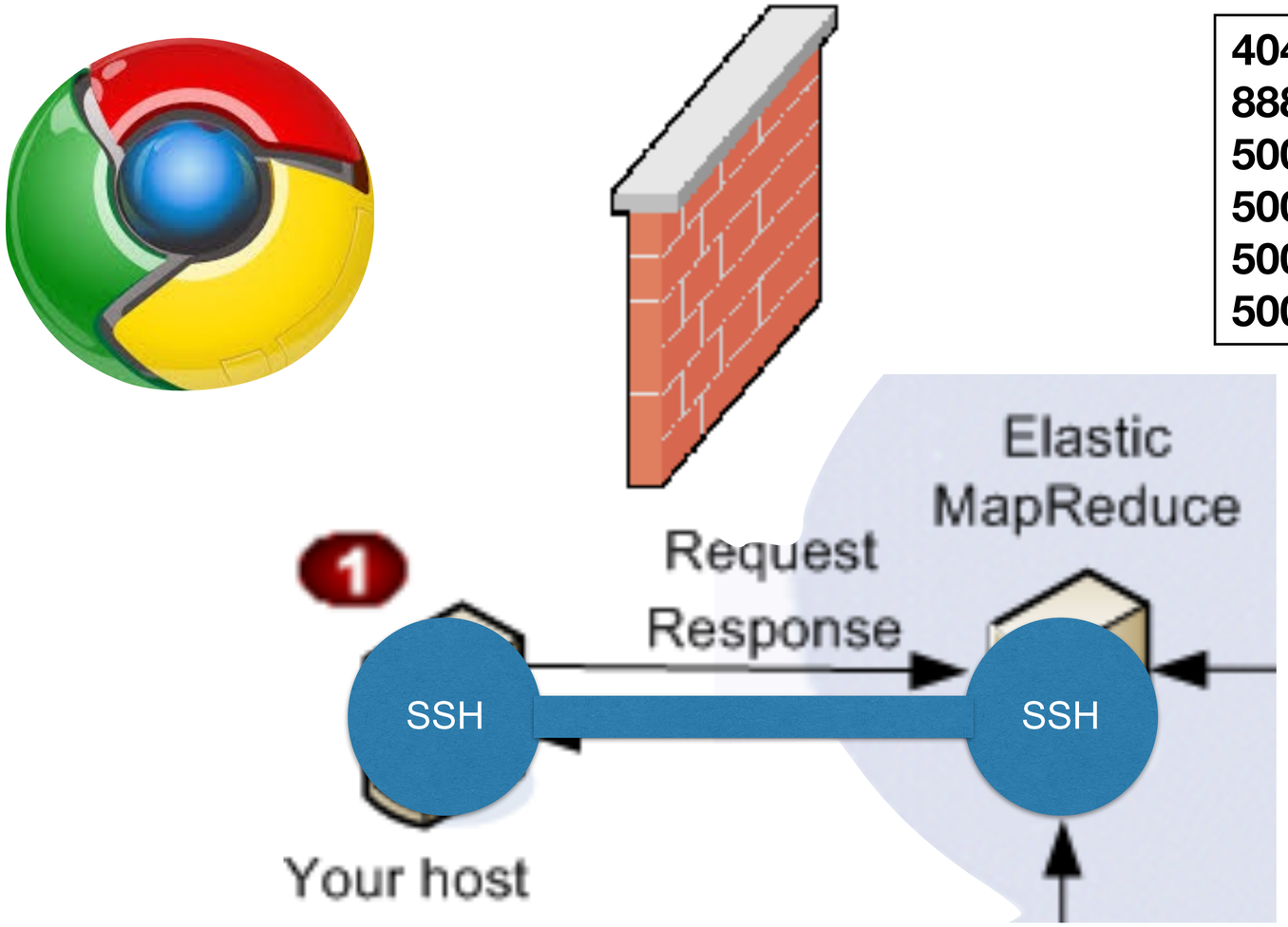
4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



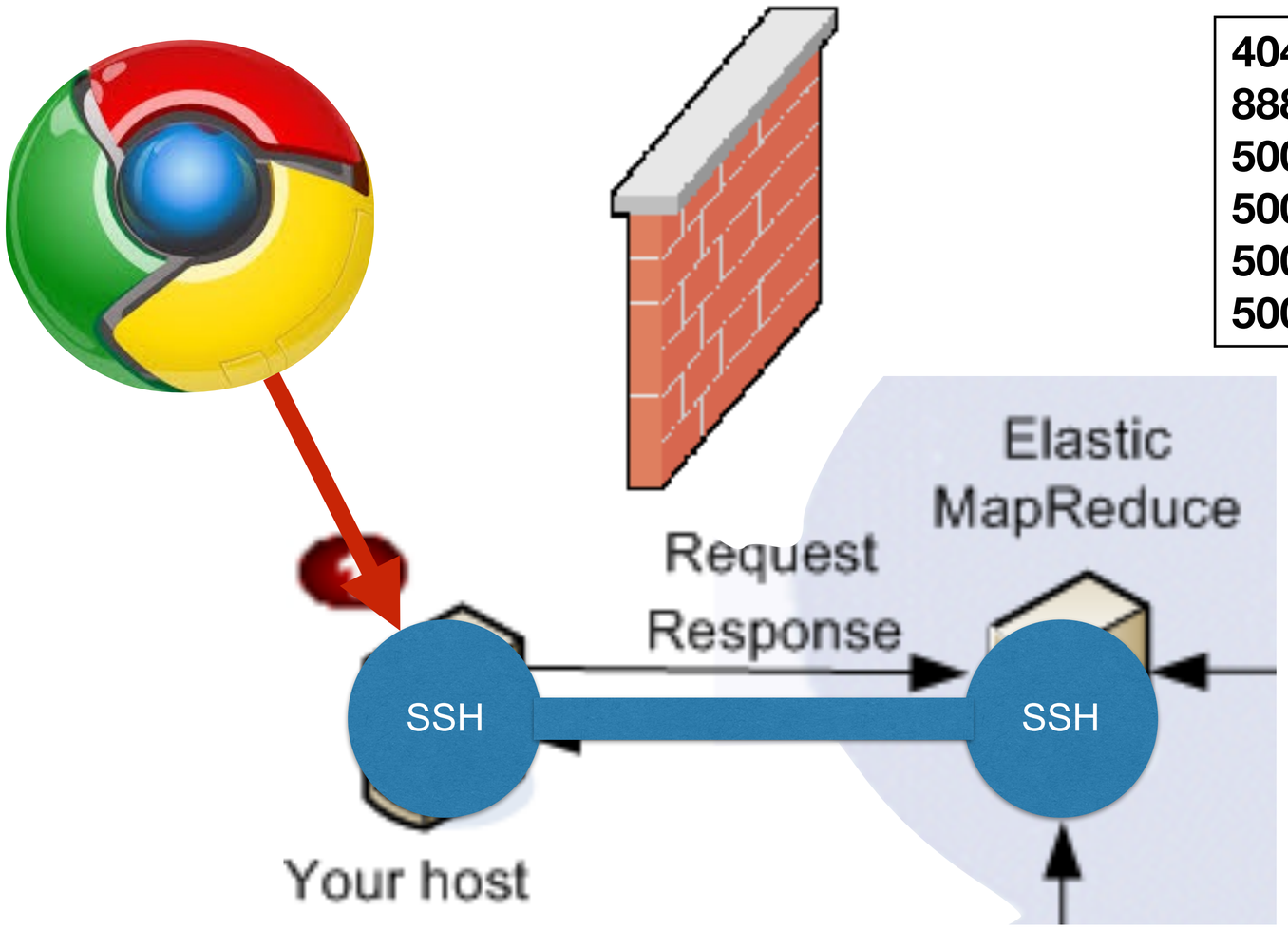
4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



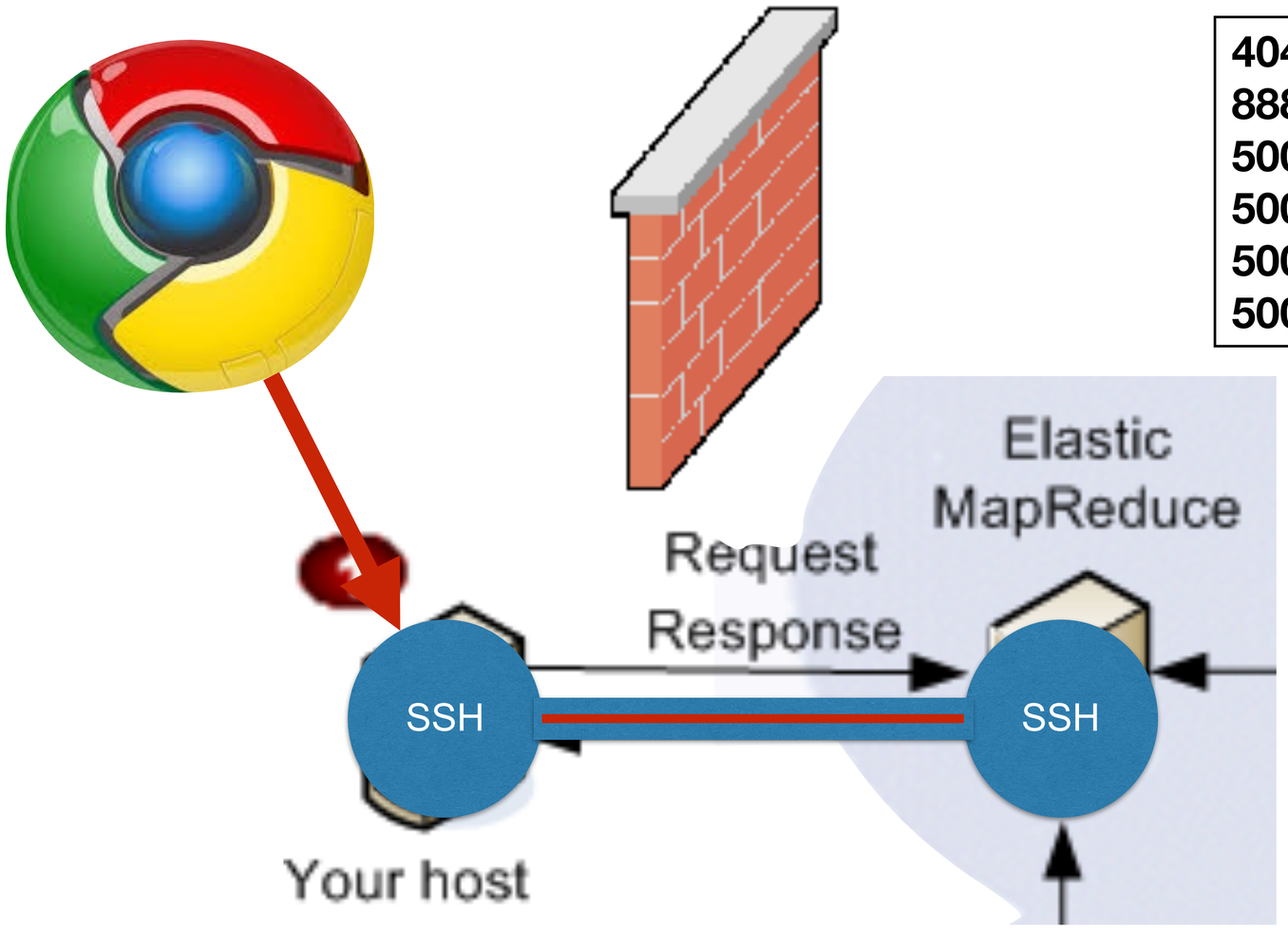
4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



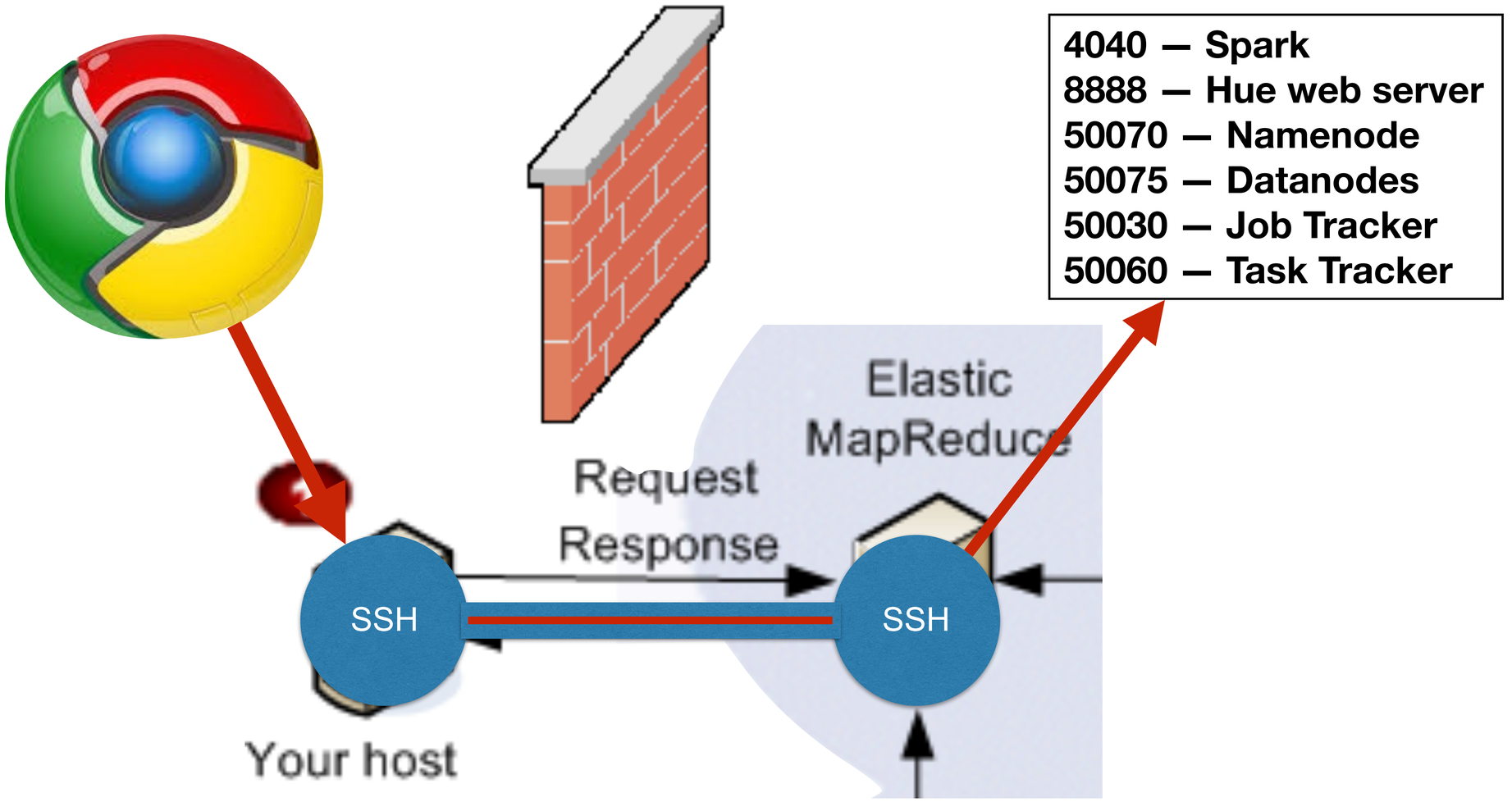
4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



4040 — Spark
8888 — Hue web server
50070 — Namenode
50075 — Datanodes
50030 — Job Tracker
50060 — Task Tracker

We use SSH to proxy the web connections



Requirements for the SSH proxy

1. Need to install "foxy proxy" in Chrome
   - Chrome add-in that sends HTTP traffic to a proxy server.
   - Rule-based; sends traffic for AWS

2. Need to configure foxy proxy
   - Download an XML file from AWS console.
   - Load the XML file into foxy proxy.

3. Need to open an SSH connection
   - SSH -ND *proxy-port hadoop@EC2-EXTERNAL-ADDRESS*
   - *(This window won't be used for anything.)*

Requirements for the SSH proxy

1. Need to install "foxy proxy" in Chrome
   - Chrome add-in that sends HTTP traffic to a proxy server.
   - Rule-based; sends traffic for AWS

2. Need to configure foxy proxy
   - Download an XML file from AWS console.
   - Load the XML file into foxy proxy.

3. Need to open an SSH connection
   - SSH -ND *proxy-port hadoop@EC2-EXTERNAL-ADDRESS*
   - *(This window won't be used for anything.)*

Requirements for the SSH proxy

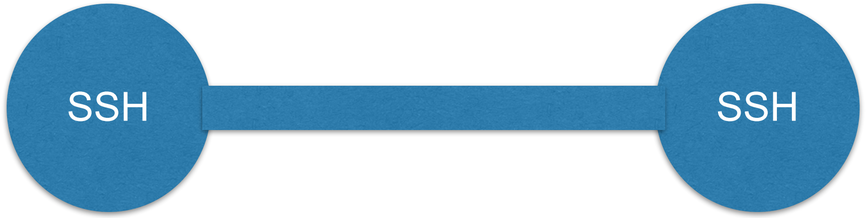
1. Need to install "foxy proxy" in Chrome
   - Chrome add-in that sends HTTP traffic to a proxy server.
   - Rule-based; sends traffic for AWS

2. Need to configure foxy proxy
   - Download an XML file from AWS console.
   - Load the XML file into foxy proxy.

3. Need to open an SSH connection
   - SSH -ND *proxy-port hadoop@EC2-EXTERNAL-ADDRESS*
   - *(This window won't be used for anything.)*

Requirements for the SSH proxy

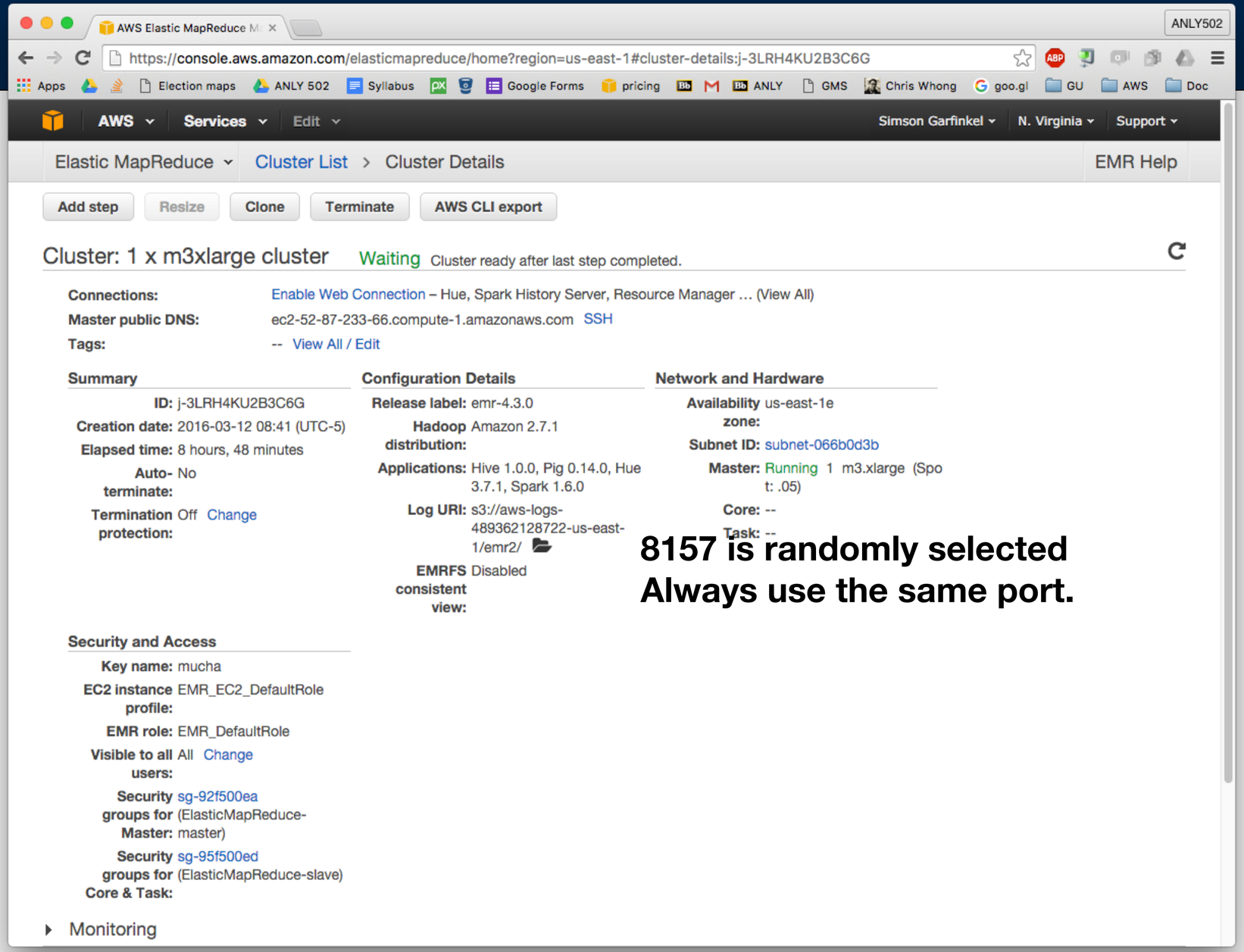
1. Need to install "foxy proxy" in Chrome
   - Chrome add-in that sends HTTP traffic to a proxy server.
   - Rule-based; sends traffic for AWS



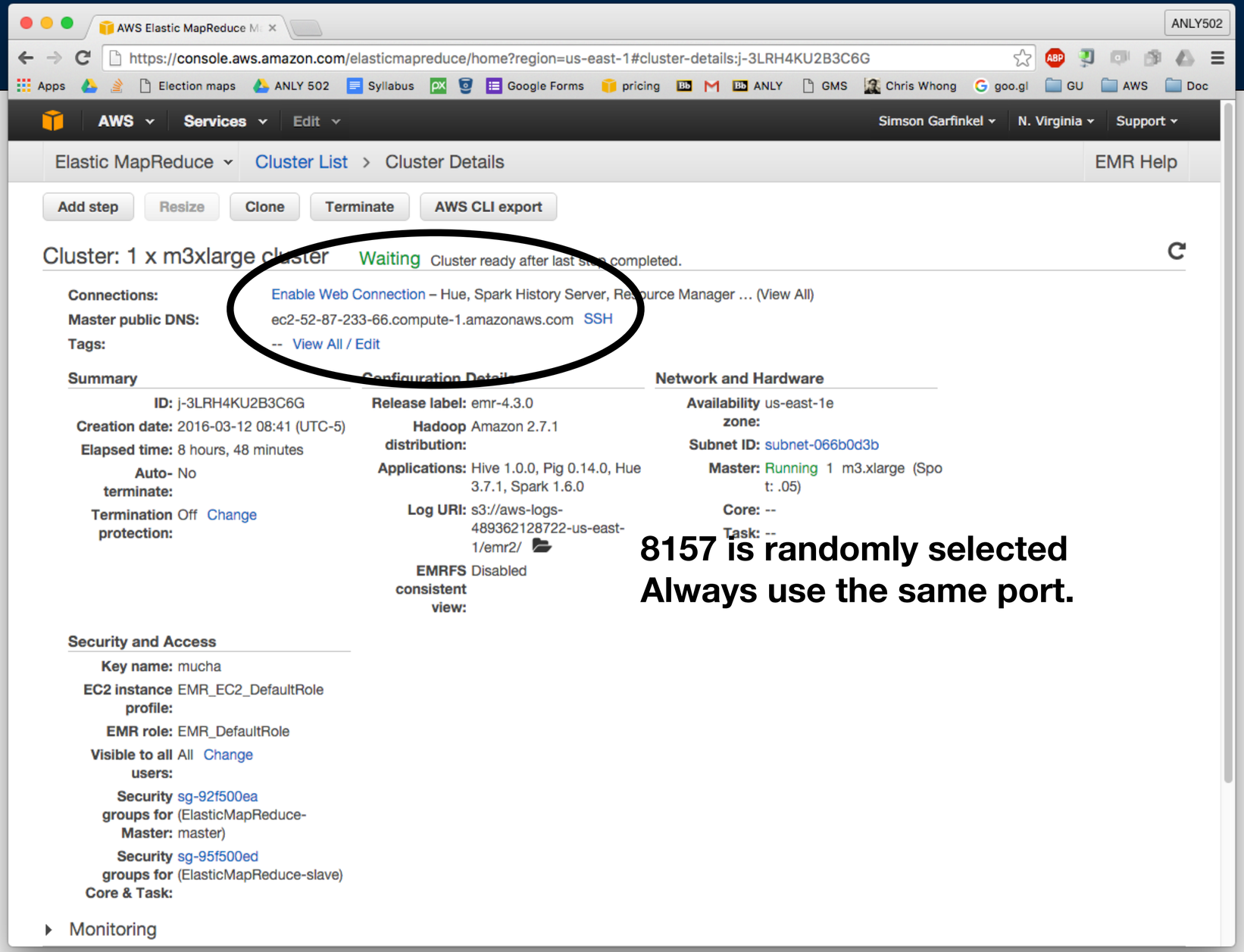
2. Need to configure foxy proxy
   - Download an XML file from AWS console.
   - Load the XML file into foxy proxy.

3. Need to open an SSH connection
   - SSH -ND *proxy-port hadoop@EC2-EXTERNAL-ADDRESS*
   - *(This window won't be used for anything.)*

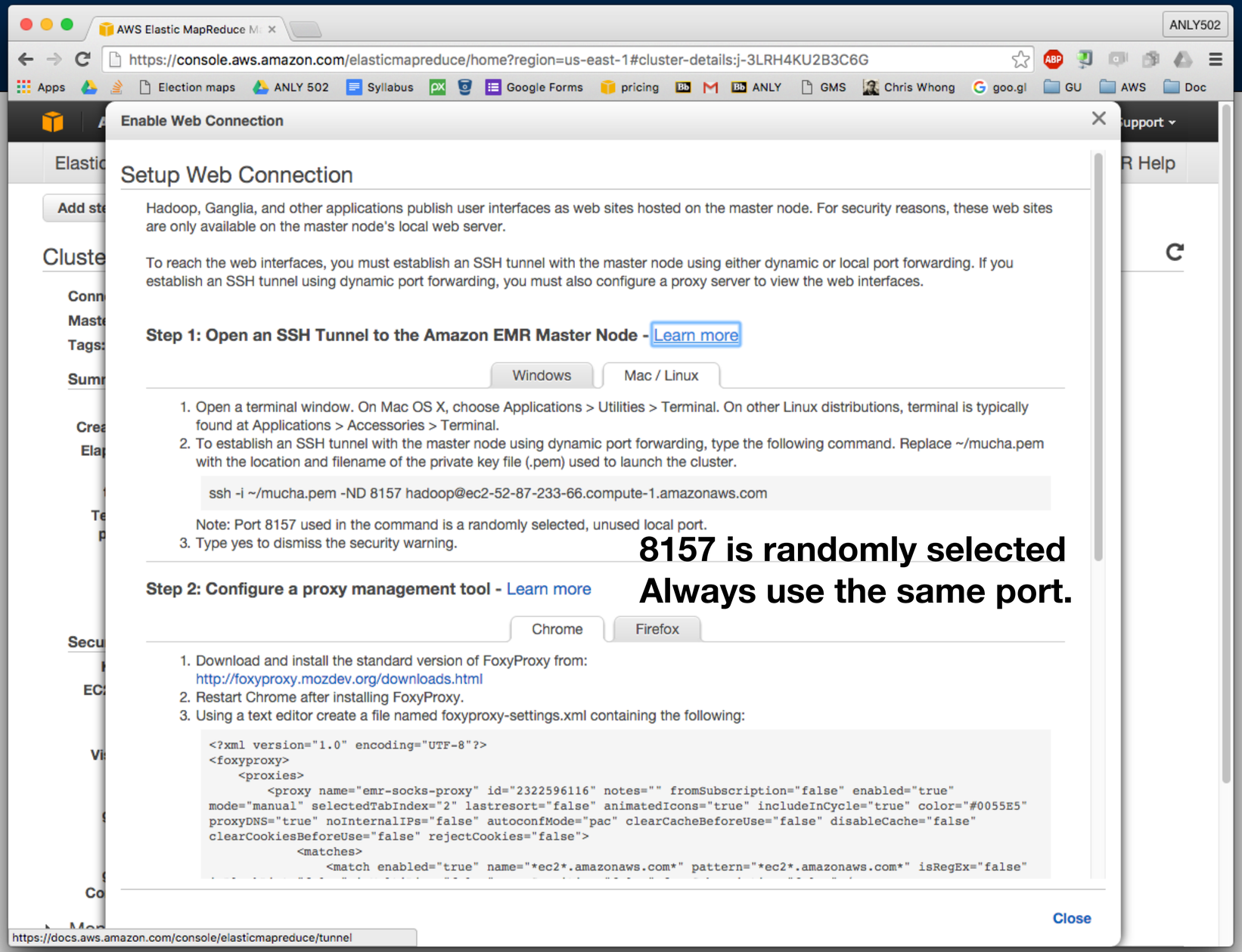
Demo:


AWS Elastic MapReduce
ANLY502
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LRH4KU2B3C6G
AWS
Services
Edit
Simson Garfinkel
N. Virginia
Support
Elastic MapReduce
Cluster List
Cluster Details
EMR Help
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: 1 x m3xlarge cluster
Waiting
Cluster ready after last step completed.
Connections:
Enable Web Connection – Hue, Spark History Server, Resource Manager … (View All)
Master public DNS:
ec2-52-87-233-66.compute-1.amazonaws.com SSH
Tags:
-- View All / Edit
Summary
ID: j-3LRH4KU2B3C6G
Creation date: 2016-03-12 08:41 (UTC-5)
Elapsed time: 8 hours, 48 minutes
Auto-terminate: No
Termination protection: Off Change
Configuration Details
Release label: emr-4.3.0
Hadoop distribution: Amazon 2.7.1
Applications: Hive 1.0.0, Pig 0.14.0, Hue 3.7.1, Spark 1.6.0
Log URI: s3://aws-logs-489362128722-us-east-1/emr2/
EMRFS consistent view: Disabled
Network and Hardware
Availability zone: us-east-1e
Subnet ID: subnet-066b0d3b
Master: Running 1 m3.xlarge (Spot: .05)
Core: --
Task: --
8157 is randomly selected
Always use the same port.
Security and Access
Key name: mucha
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-92f500ea (ElasticMapReduce-master)
Security groups for Core & Task: sg-95f500ed (ElasticMapReduce-slave)
Monitoring

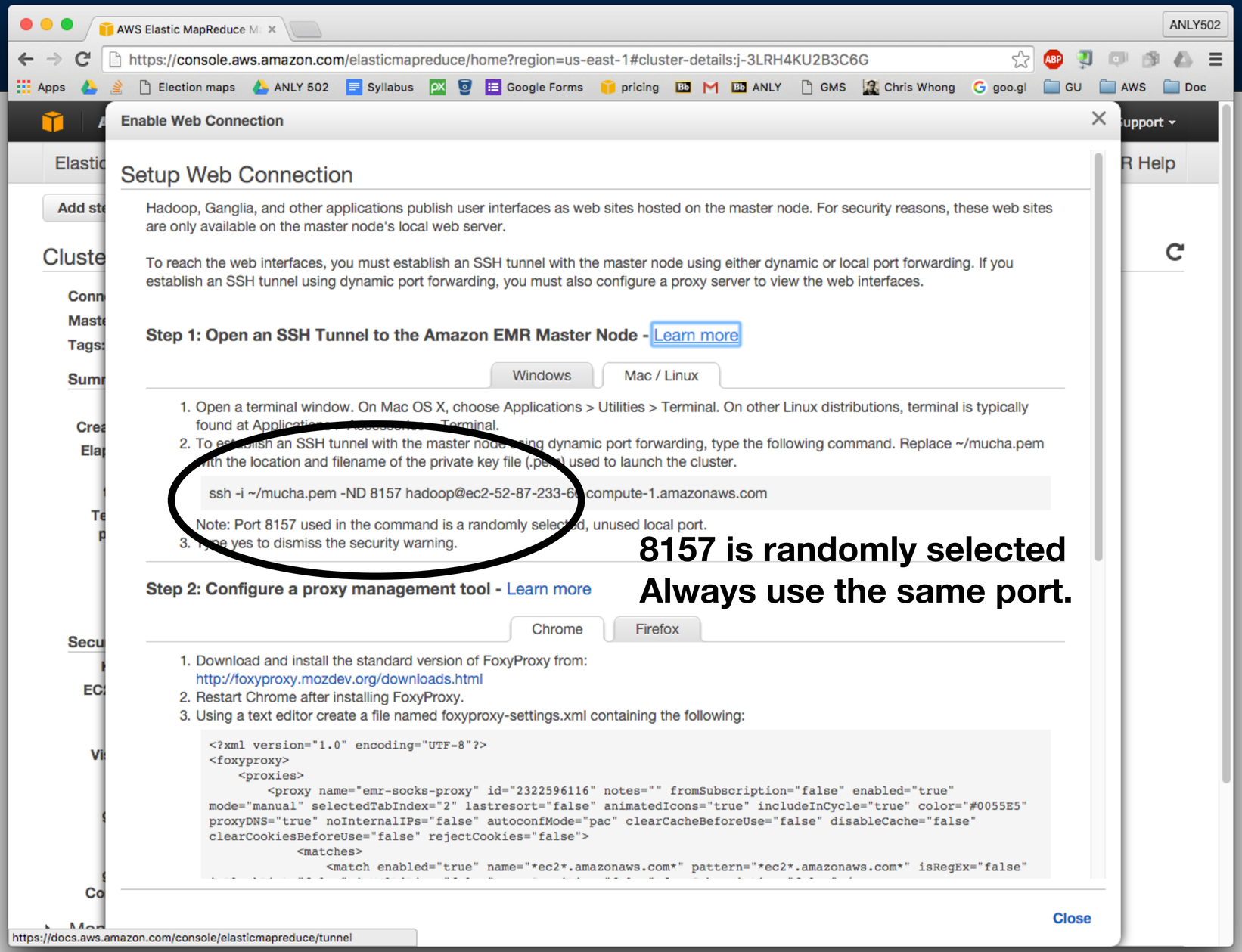
Demo:


AWS Elastic MapReduce
ANLY502
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LRH4KU2B3C6G
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
AWS
Services
Edit
Simson Garfinkel
N. Virginia
Support
Elastic MapReduce
Cluster List
Cluster Details
EMR Help
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: 1 x m3xlarge cluster
Waiting
Cluster ready after last step completed.
Connections:
Enable Web Connection – Hue, Spark History Server, Resource Manager … (View All)
Master public DNS:
ec2-52-87-233-66.compute-1.amazonaws.com SSH
Tags:
-- View All / Edit
Summary
ID: j-3LRH4KU2B3C6G
Creation date: 2016-03-12 08:41 (UTC-5)
Elapsed time: 8 hours, 48 minutes
Auto-terminate: No
Termination protection: Off Change
Configuration Details
Release label: emr-4.3.0
Hadoop distribution: Amazon 2.7.1
Applications: Hive 1.0.0, Pig 0.14.0, Hue 3.7.1, Spark 1.6.0
Log URI: s3://aws-logs-489362128722-us-east-1/emr2/
EMRFS consistent view: Disabled
Network and Hardware
Availability zone: us-east-1e
Subnet ID: subnet-066b0d3b
Master: Running 1 m3.xlarge (Spot: .05)
Core: --
Task: --
Security and Access
Key name: mucha
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-92f500ea (ElasticMapReduce-master)
Security groups for Core & Task: sg-95f500ed (ElasticMapReduce-slave)
Monitoring
8157 is randomly selected
Always use the same port.

Demo:

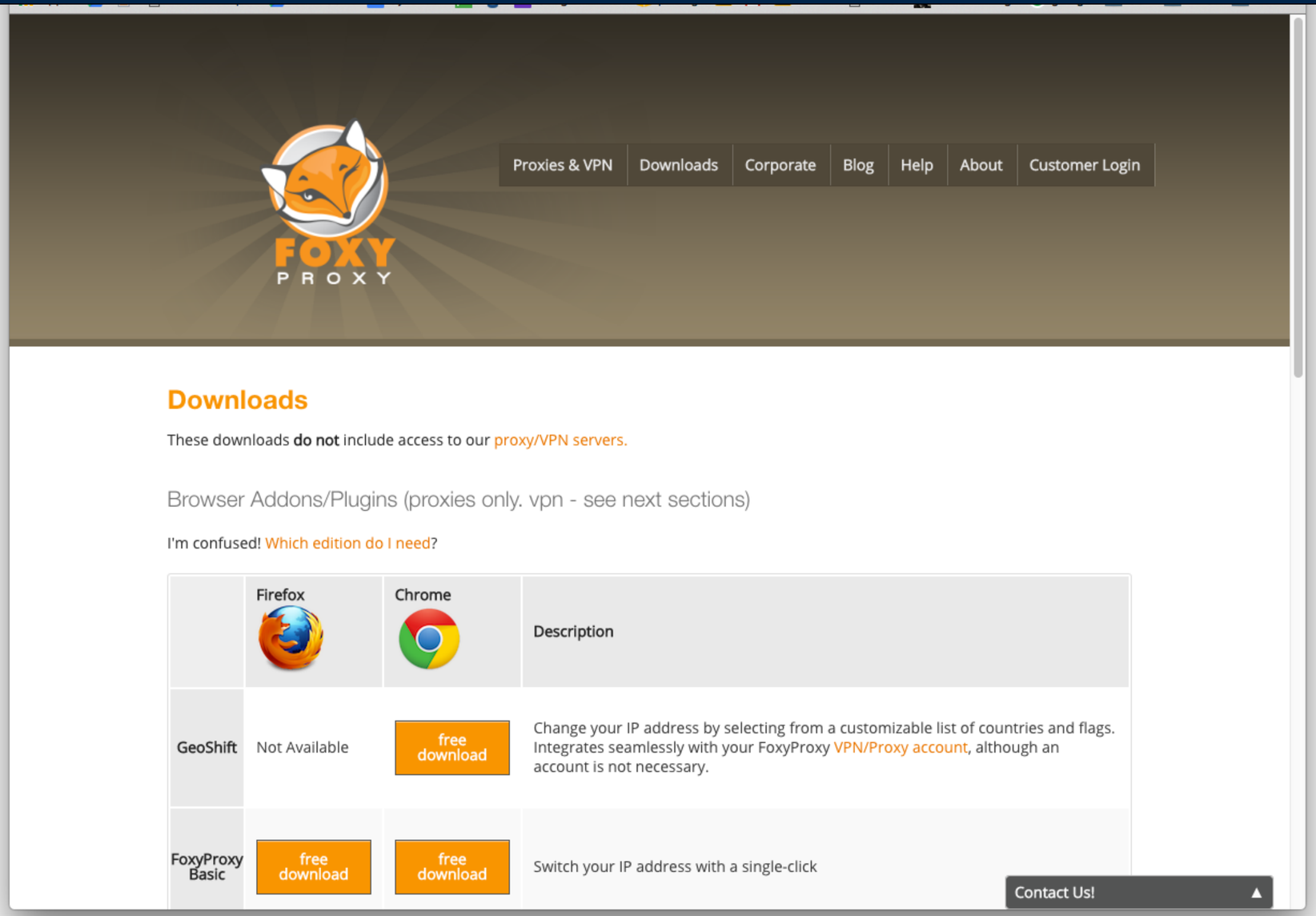

AWS Elastic MapReduce
ANLY502
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LRH4KU2B3C6G
Enable Web Connection
Setup Web Connection
Hadoop, Ganglia, and other applications publish user interfaces as web sites hosted on the master node. For security reasons, these web sites are only available on the master node's local web server.
To reach the web interfaces, you must establish an SSH tunnel with the master node using either dynamic or local port forwarding. If you establish an SSH tunnel using dynamic port forwarding, you must also configure a proxy server to view the web interfaces.
Step 1: Open an SSH Tunnel to the Amazon EMR Master Node - Learn more
Windows
Mac / Linux
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish an SSH tunnel with the master node using dynamic port forwarding, type the following command. Replace ~/mucha.pem with the location and filename of the private key file (.pem) used to launch the cluster.
ssh -i ~/mucha.pem -ND 8157 hadoop@ec2-52-87-233-66.compute-1.amazonaws.com
Note: Port 8157 used in the command is a randomly selected, unused local port.
3. Type yes to dismiss the security warning.
8157 is randomly selected
Always use the same port.
Step 2: Configure a proxy management tool - Learn more
Chrome
Firefox
1. Download and install the standard version of FoxyProxy from:
http://foxyproxy.mozdev.org/downloads.html
2. Restart Chrome after installing FoxyProxy.
3. Using a text editor create a file named foxyproxy-settings.xml containing the following:
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
<proxies>
<proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true"
mode="manual" selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5"
proxyDNS="true" noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false"
clearCookiesBeforeUse="false" rejectCookies="false">
<matches>
<match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegEx="false"
Close
https://docs.aws.amazon.com/console/elasticmapreduce/tunnel

Demo:

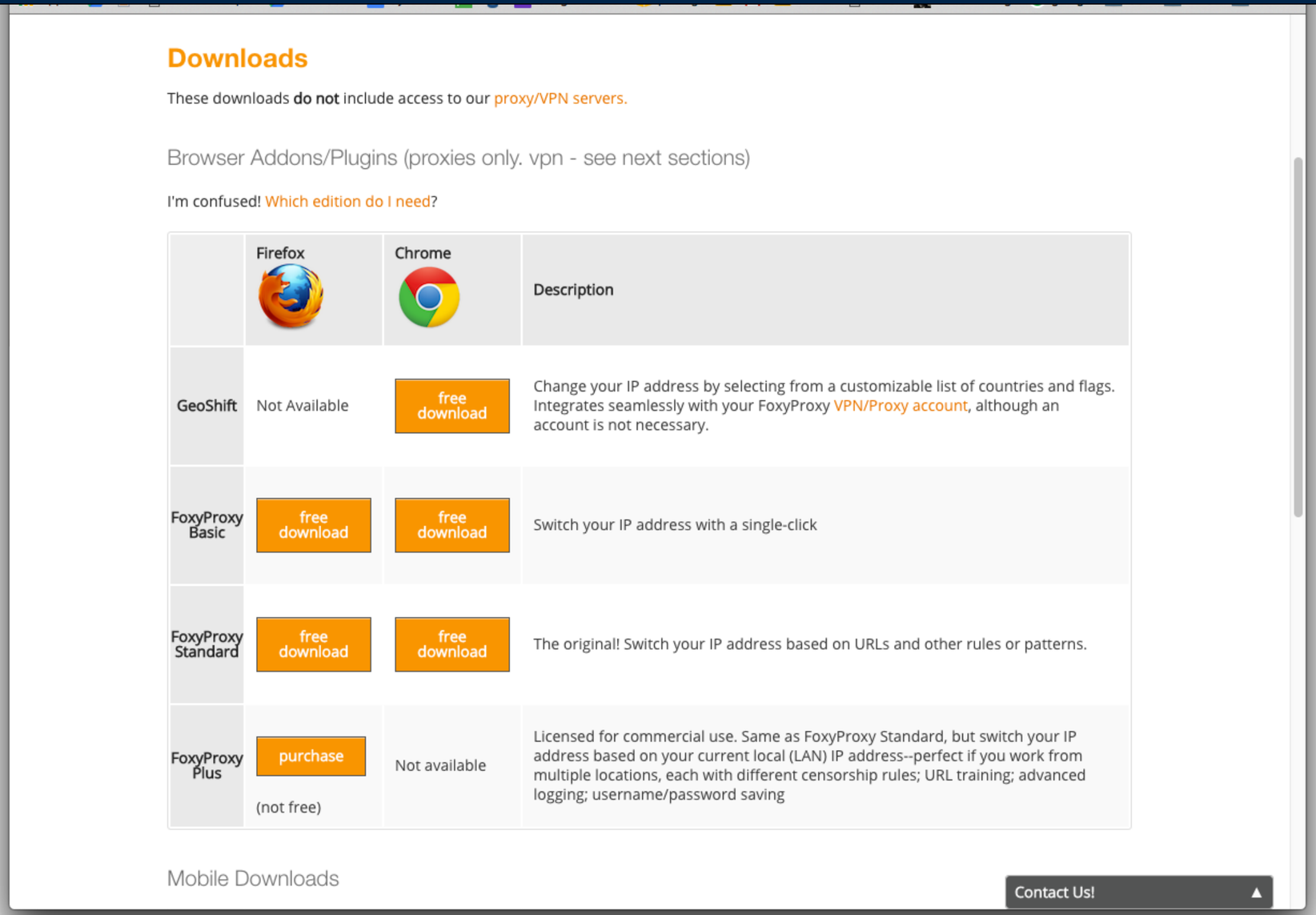

AWS Elastic MapReduce
ANLY502
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LRH4KU2B3C6G
Enable Web Connection
Setup Web Connection
Hadoop, Ganglia, and other applications publish user interfaces as web sites hosted on the master node. For security reasons, these web sites are only available on the master node's local web server.
To reach the web interfaces, you must establish an SSH tunnel with the master node using either dynamic or local port forwarding. If you establish an SSH tunnel using dynamic port forwarding, you must also configure a proxy server to view the web interfaces.
Step 1: Open an SSH Tunnel to the Amazon EMR Master Node - Learn more
Windows
Mac / Linux
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish an SSH tunnel with the master node using dynamic port forwarding, type the following command. Replace ~/mucha.pem with the location and filename of the private key file (.pem) used to launch the cluster.
ssh -i ~/mucha.pem -ND 8157 hadoop@ec2-52-87-233-6 compute-1.amazonaws.com
Note: Port 8157 used in the command is a randomly selected, unused local port.
3. Type yes to dismiss the security warning.
8157 is randomly selected
Always use the same port.
Step 2: Configure a proxy management tool - Learn more
Chrome
Firefox
1. Download and install the standard version of FoxyProxy from:
http://foxyproxy.mozdev.org/downloads.html
2. Restart Chrome after installing FoxyProxy.
3. Using a text editor create a file named foxyproxy-settings.xml containing the following:
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
<proxies>
<proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true" mode="manual" selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5" proxyDNS="true" noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false" clearCookiesBeforeUse="false" rejectCookies="false">
<matches>
<match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegEx="false"
Close
https://docs.aws.amazon.com/console/elasticmapreduce/tunnel

Install foxyproxy

Install foxyproxy



Install foxyproxy



Downloads

These downloads **do not** include access to our proxy/VPN servers.

Browser Addons/Plugins (proxies only. vpn - see next sections)

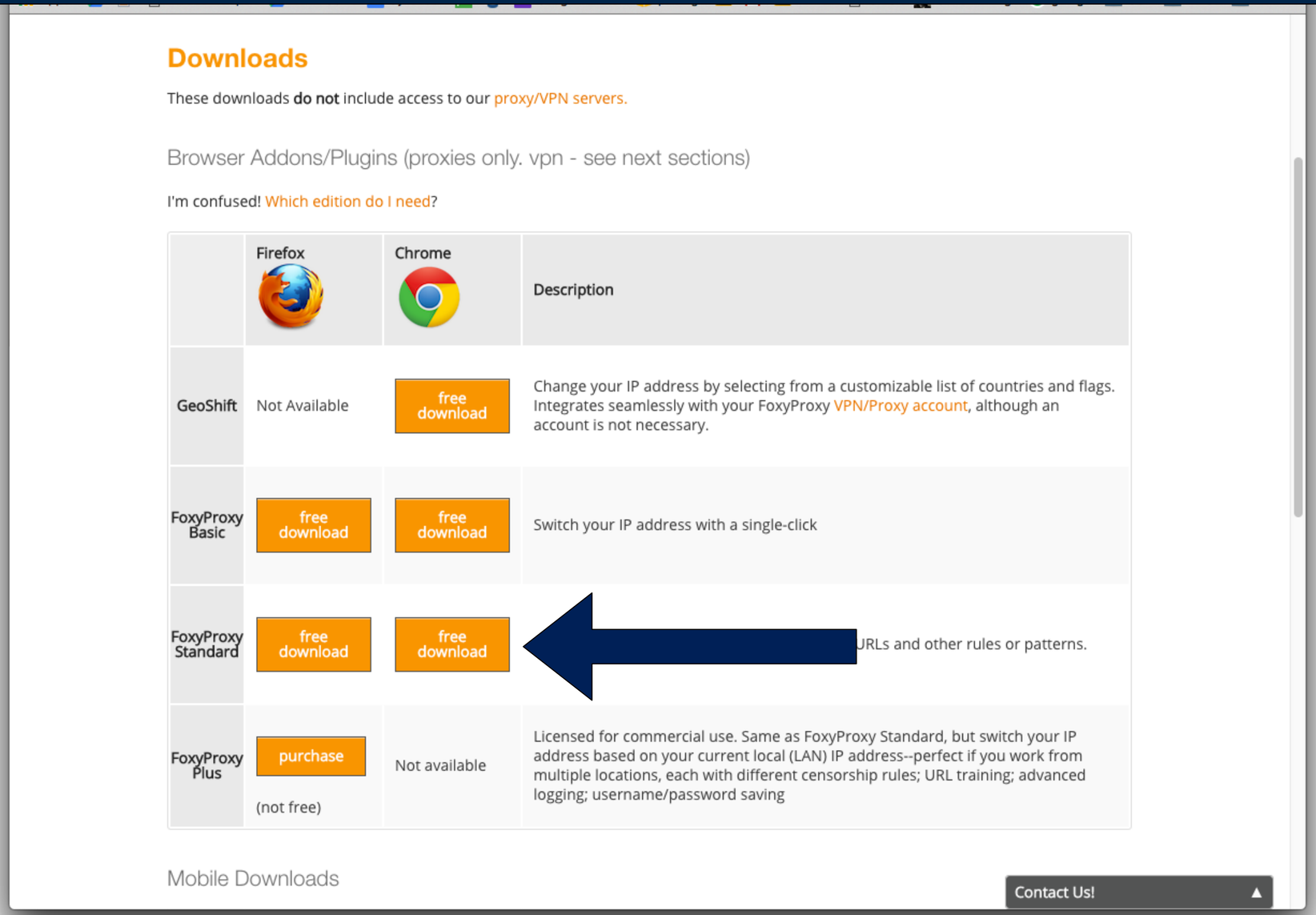
I'm confused! Which edition do I need?

	Firefox	Chrome	Description
GeoShift	Not Available	free download	Change your IP address by selecting from a customizable list of countries and flags. Integrates seamlessly with your FoxyProxy VPN/Proxy account, although an account is not necessary.
FoxyProxy Basic	free download	free download	Switch your IP address with a single-click
FoxyProxy Standard	free download	free download	The original! Switch your IP address based on URLs and other rules or patterns.
FoxyProxy Plus	purchase (not free)	Not available	Licensed for commercial use. Same as FoxyProxy Standard, but switch your IP address based on your current local (LAN) IP address--perfect if you work from multiple locations, each with different censorship rules; URL training; advanced logging; username/password saving

Mobile Downloads

Contact Us! ▲

Install foxyproxy



Downloads

These downloads **do not** include access to our proxy/VPN servers.

Browser Addons/Plugins (proxies only. vpn - see next sections)

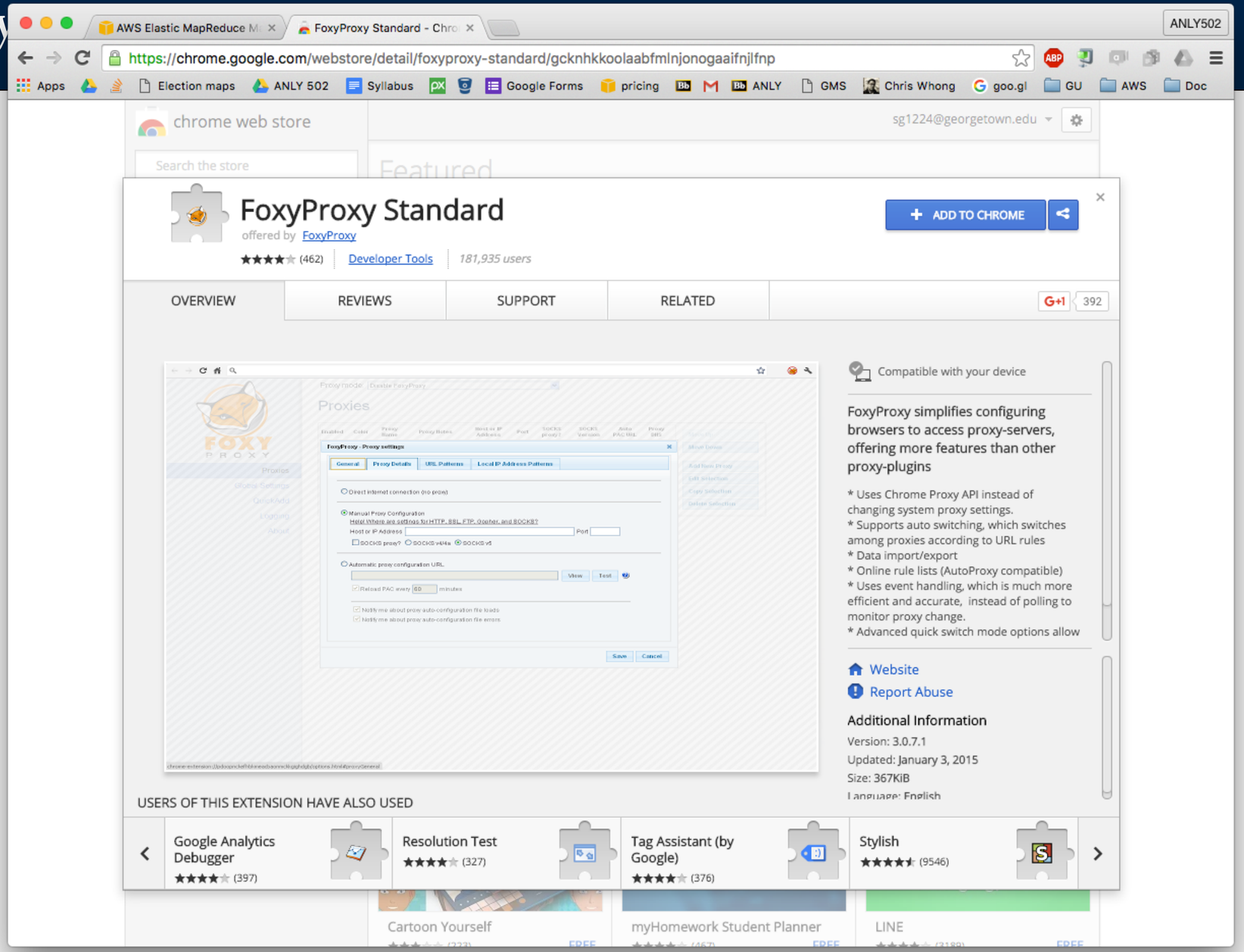
I'm confused! Which edition do I need?

	Firefox	Chrome	Description
GeoShift	Not Available	free download	Change your IP address by selecting from a customizable list of countries and flags. Integrates seamlessly with your FoxyProxy VPN/Proxy account, although an account is not necessary.
FoxyProxy Basic	free download	free download	Switch your IP address with a single-click
FoxyProxy Standard	free download	free download	URLs and other rules or patterns.
FoxyProxy Plus	purchase (not free)	Not available	Licensed for commercial use. Same as FoxyProxy Standard, but switch your IP address based on your current local (LAN) IP address--perfect if you work from multiple locations, each with different censorship rules; URL training; advanced logging; username/password saving

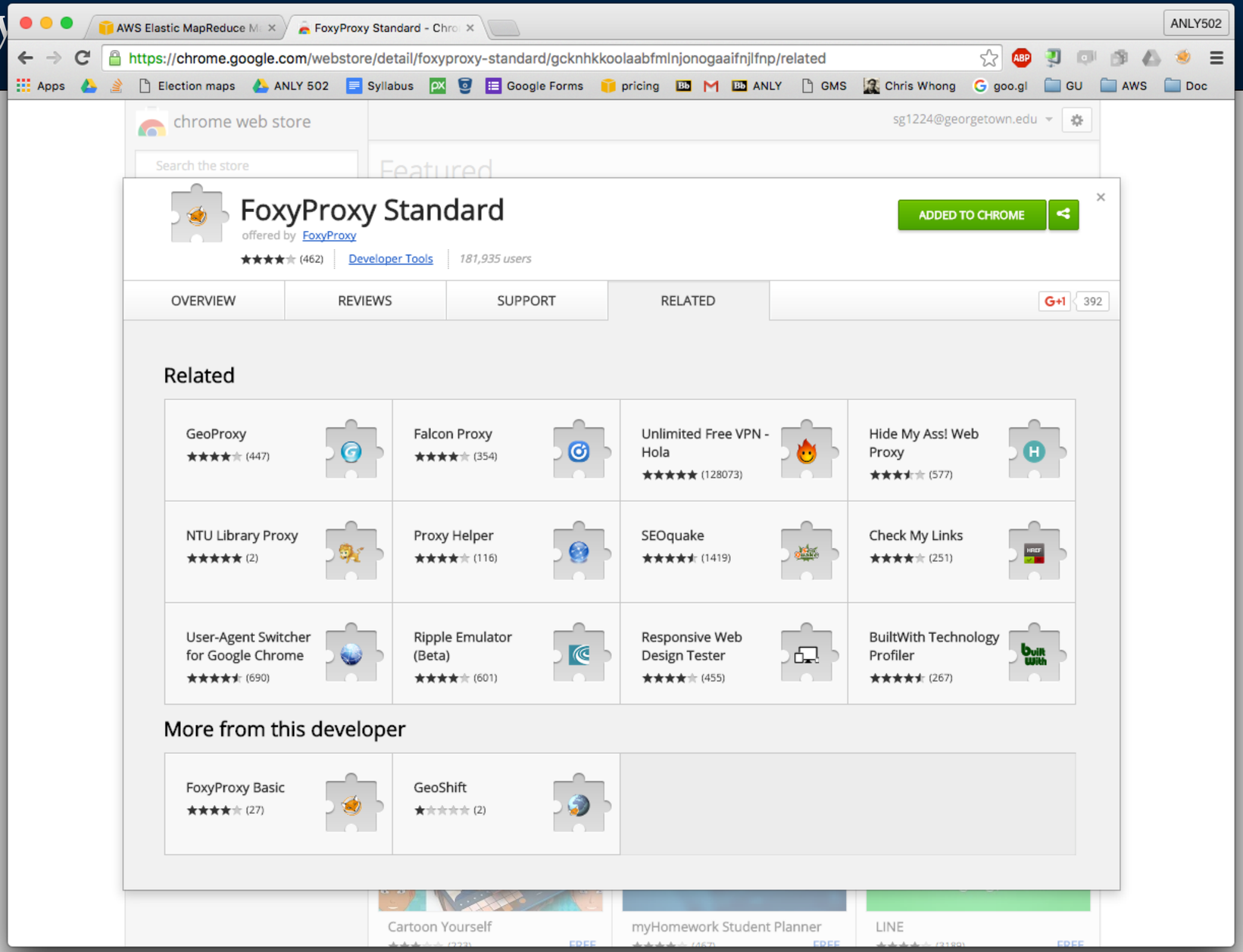
Mobile Downloads

Contact Us! ▲

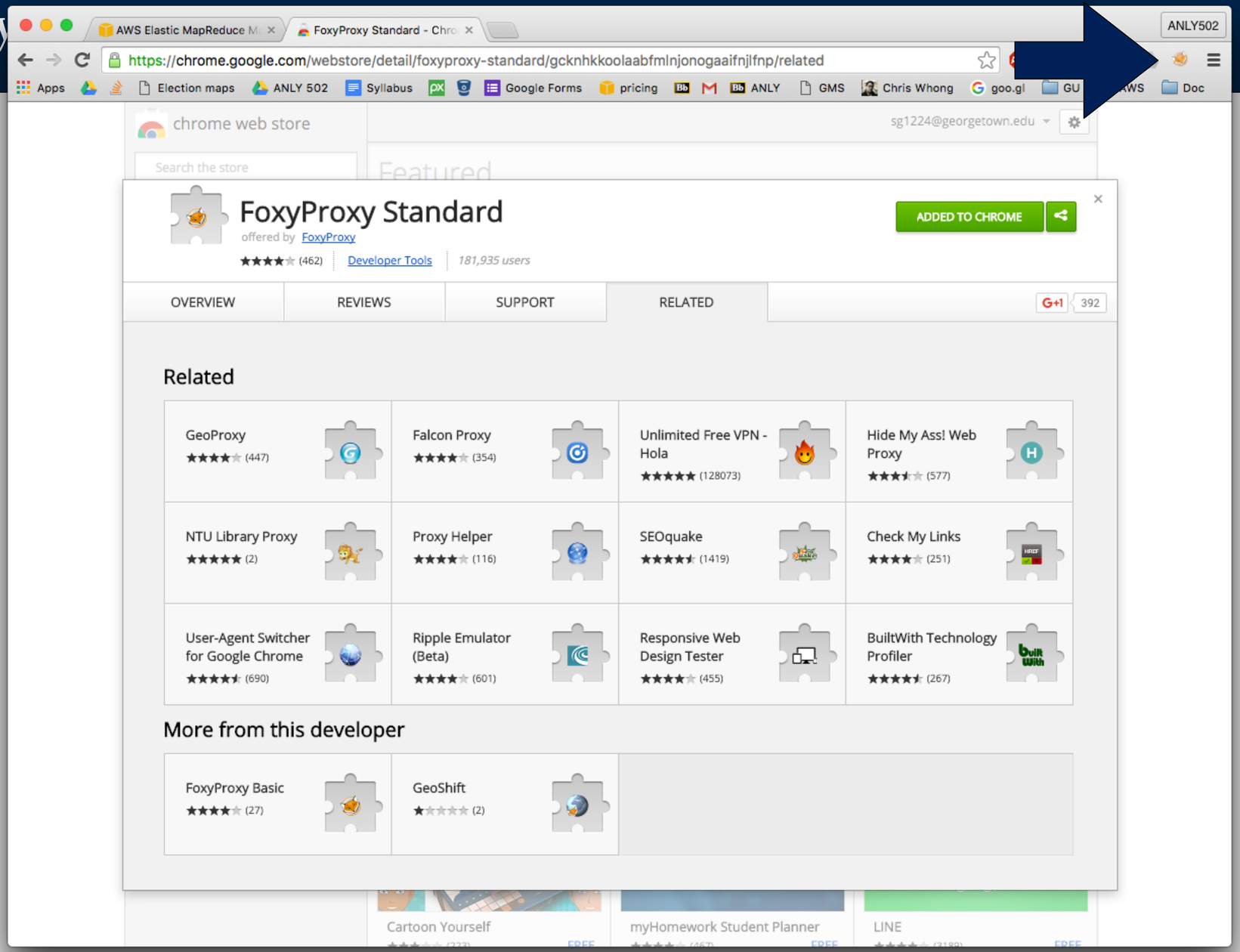
Install foxyproxy



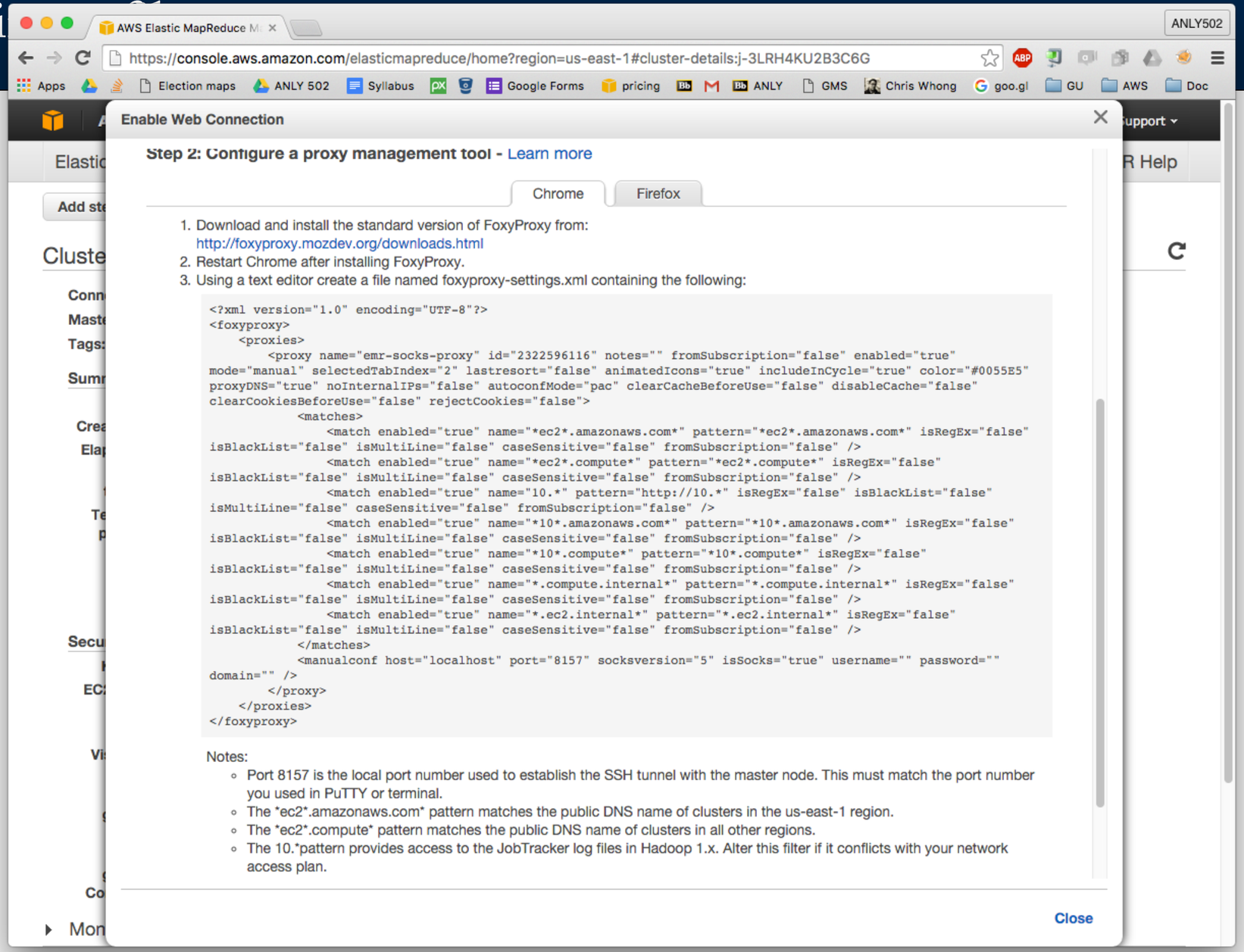
Install foxyproxy

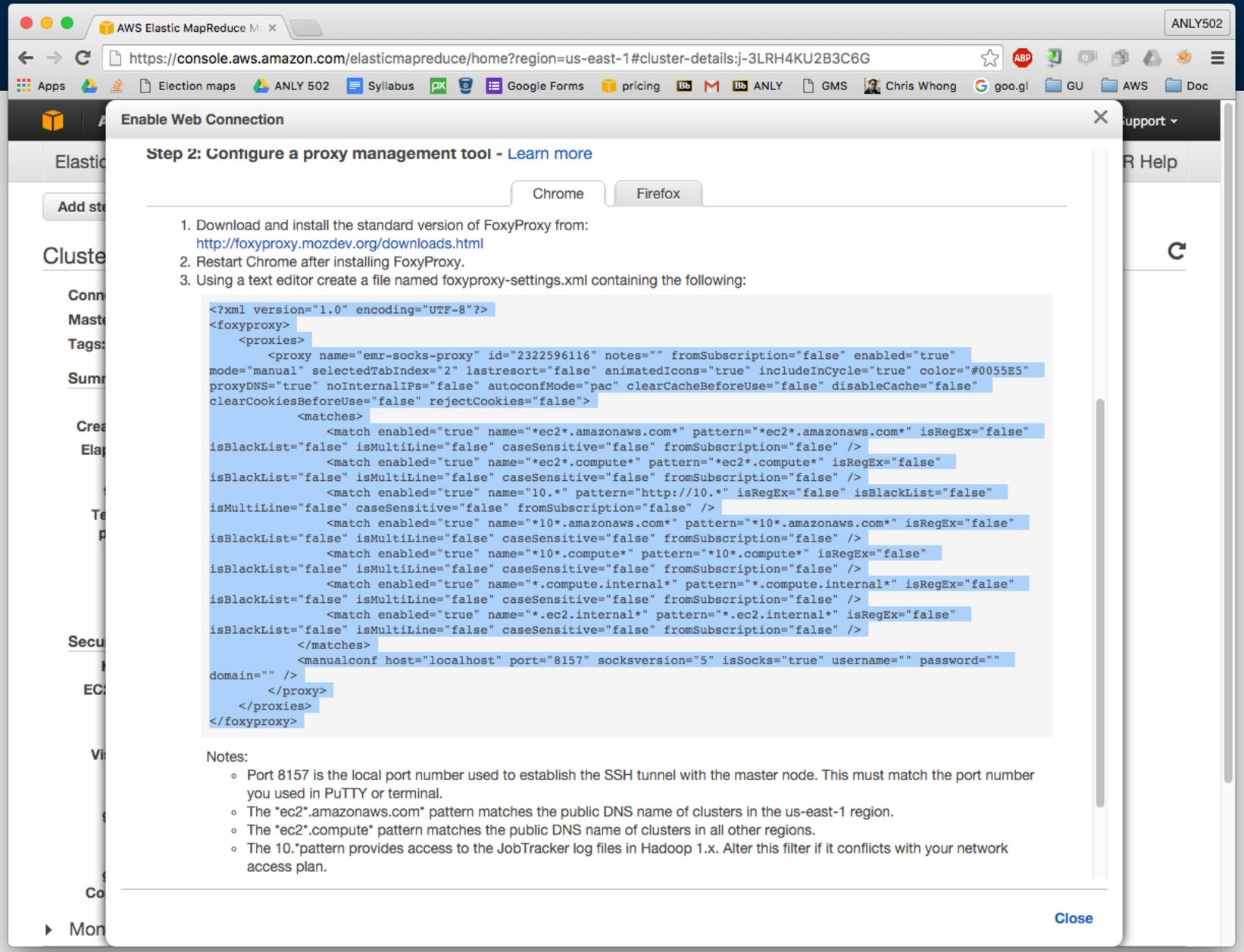


Install foxyproxy



Save the XML i

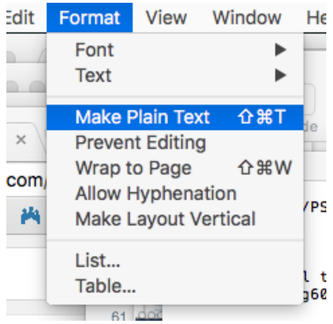


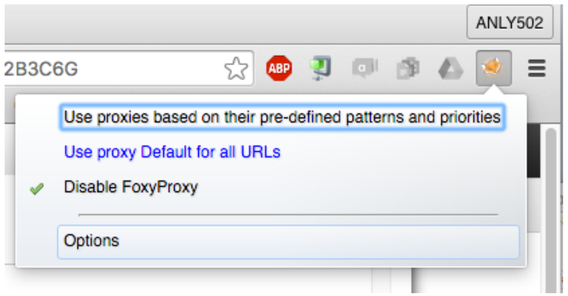

AWS Elastic MapReduce
ANLY502
https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-3LRH4KU2B3C6G
Enable Web Connection
Step 2: Configure a proxy management tool - Learn more
Chrome
Firefox
1. Download and install the standard version of FoxyProxy from:
http://foxyproxy.mozdev.org/downloads.html
2. Restart Chrome after installing FoxyProxy.
3. Using a text editor create a file named foxyproxy-settings.xml containing the following:
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
<proxies>
<proxy name="emr-socks-proxy" id="2322596116" notes="" fromSubscription="false" enabled="true"
mode="manual" selectedTabIndex="2" lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5"
proxyDNS="true" noInternalIPs="false" autoconfMode="pac" clearCacheBeforeUse="false" disableCache="false"
clearCookiesBeforeUse="false" rejectCookies="false">
<matches>
<match enabled="true" name="*ec2*.amazonaws.com*" pattern="*ec2*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="*ec2*.compute*" pattern="*ec2*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="10.*" pattern="http://10.*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="*10*.amazonaws.com*" pattern="*10*.amazonaws.com*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="*10*.compute*" pattern="*10*.compute*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="*.compute.internal*" pattern="*.compute.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
<match enabled="true" name="*.ec2.internal*" pattern="*.ec2.internal*" isRegEx="false"
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscription="false" />
</matches>
<manualconf host="localhost" port="8157" socksversion="5" isSocks="true" username="" password=""
domain="" />
</proxy>
</proxies>
</foxyproxy>
Notes:
Port 8157 is the local port number used to establish the SSH tunnel with the master node. This must match the port number you used in PuTTY or terminal.
The *ec2*.amazonaws.com* pattern matches the public DNS name of clusters in the us-east-1 region.
The *ec2*.compute* pattern matches the public DNS name of clusters in all other regions.
The 10.*pattern provides access to the JobTracker log files in Hadoop 1.x. Alter this filter if it conflicts with your network access plan.
Close

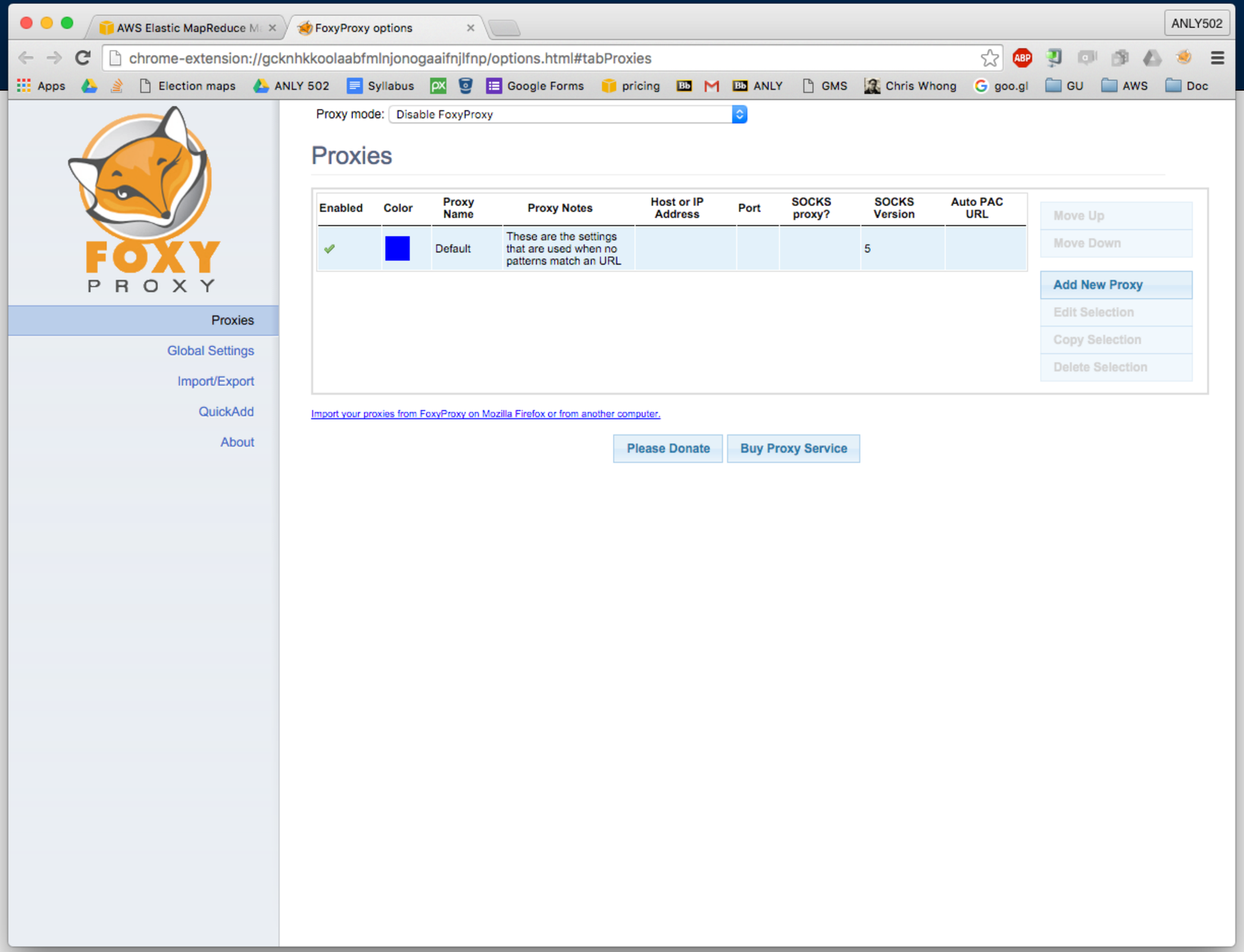
(be sure you have a TEXT file;

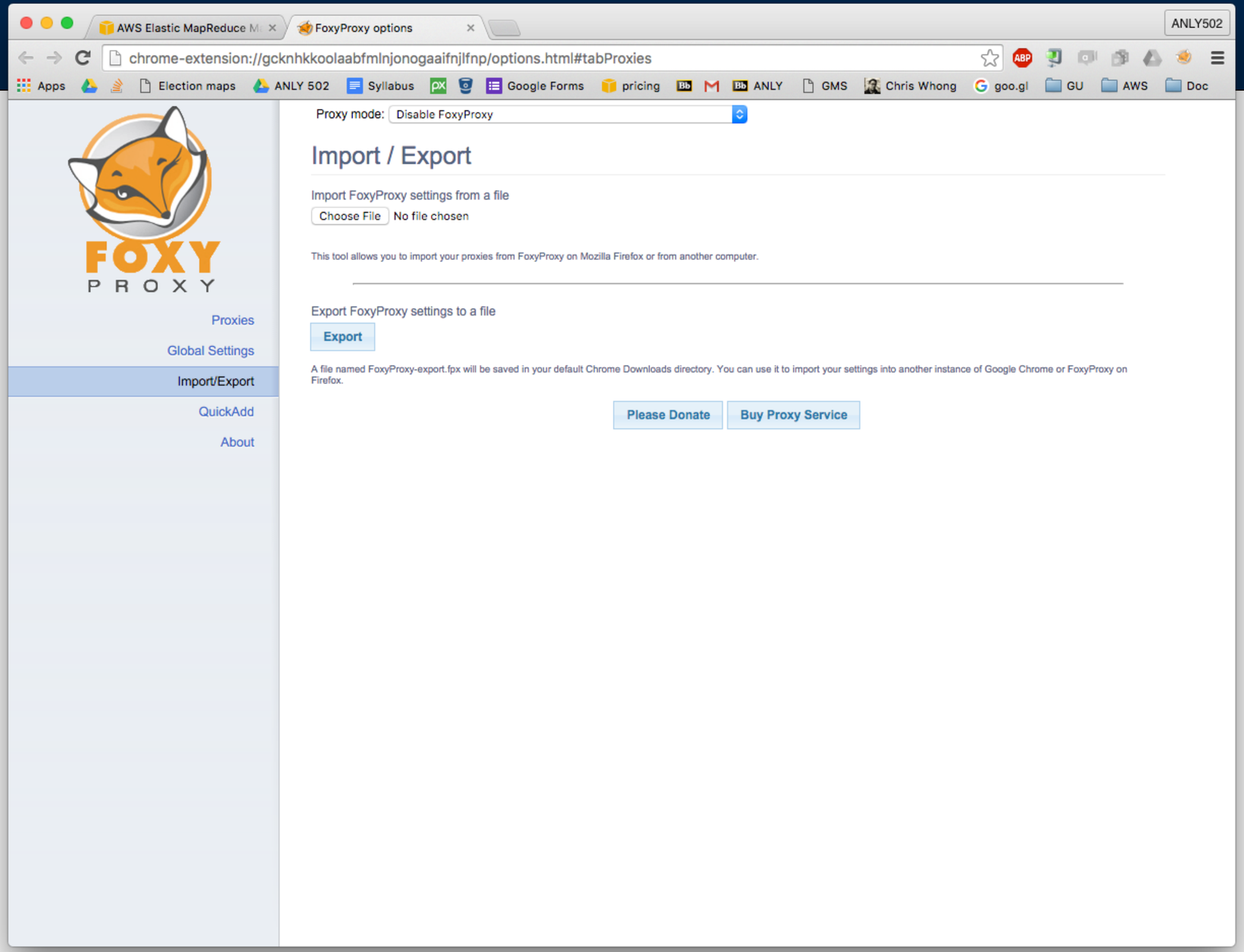
foxyproxy-settings.xml.txt

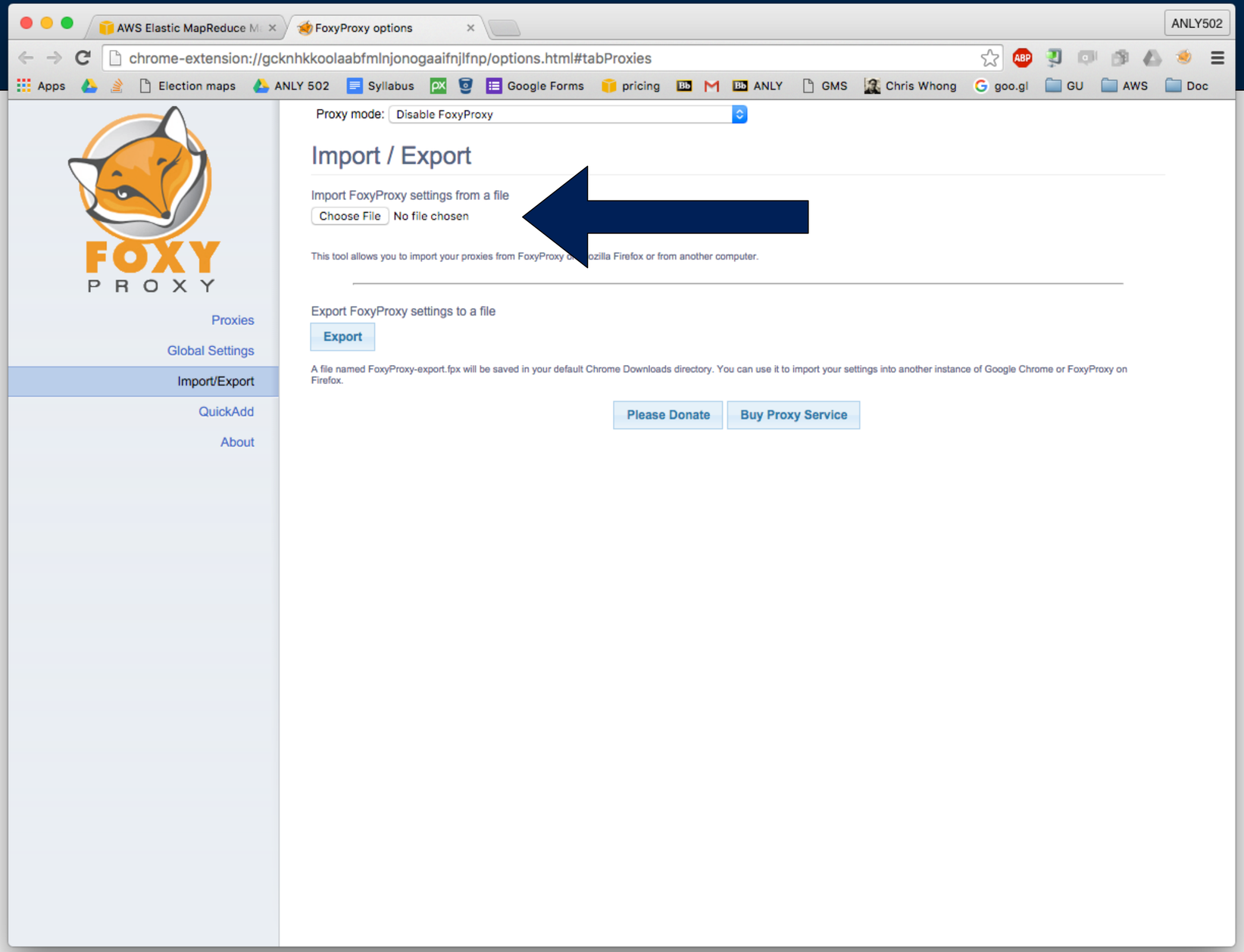
```
<?xml version="1.0" encoding="UTF-8"?>
<foxyproxy>
    <proxies>
        <proxy name="emr-socks-proxy" id="2322596116" notes=""
fromSubscription="false" enabled="true" mode="manual" selectedTabIndex="2"
lastresort="false" animatedIcons="true" includeInCycle="true" color="#0055E5"
proxyDNS="true" noInternalIPs="false" autoconfMode="pac"
clearCacheBeforeUse="false" disableCache="false"
clearCookiesBeforeUse="false" rejectCookies="false">
            <matches>
                <match enabled="true" name="*ec2*.amazonaws.com*"
pattern="*ec2*.amazonaws.com*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*ec2*.compute*"
pattern="*ec2*.compute*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="10.*" pattern="http://10.*"
isRegEx="false" isBlackList="false" isMultiLine="false" caseSensitive="false"
fromSubscription="false" />
                <match enabled="true" name="*10*.amazonaws.com*"
pattern="*10*.amazonaws.com*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*10*.compute*"
pattern="*10*.compute*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.compute.internal*"
pattern="*.compute.internal*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
                <match enabled="true" name="*.ec2.internal*"
pattern="*.ec2.internal*" isRegEx="false" isBlackList="false"
isMultiLine="false" caseSensitive="false" fromSubscription="false" />
            </matches>
            <manualconf host="localhost" port="8157" socksversion="5"
isSocks="true" username="" password="" domain="" />
        </proxy>
    </proxies>
</foxyproxy>
```

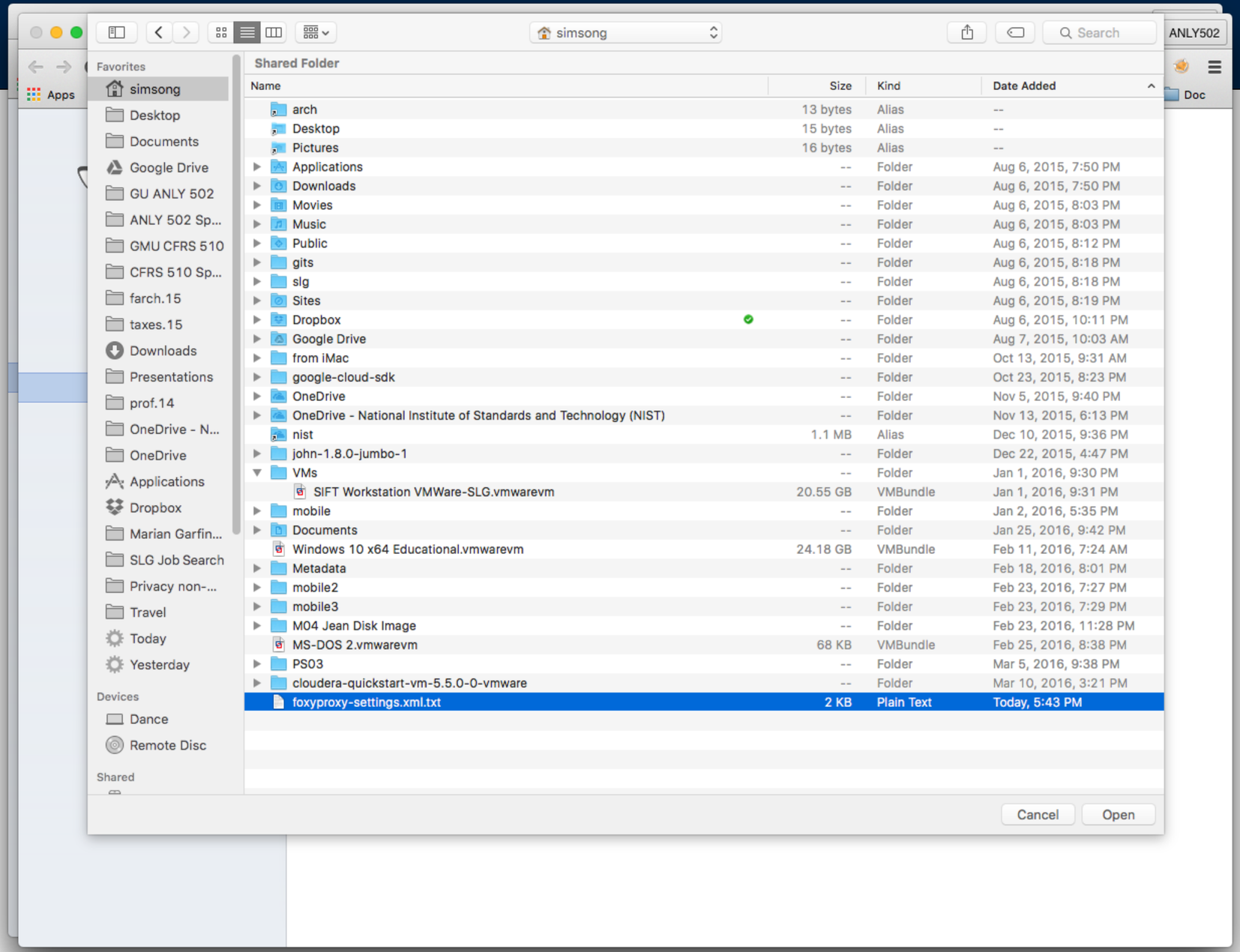





AWS Elastic MapReduce M
FoxyProxy options
ANLY502
chrome-extension://gcknhkkoolaabfmlnjonogaaifnjlfnp/options.html#tabProxies
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
FOXY PROXY
Proxies
Global Settings
Import/Export
QuickAdd
About
Proxy mode: Disable FoxyProxy
Proxies
Enabled
Color
Proxy Name
Proxy Notes
Host or IP Address
Port
SOCKS proxy?
SOCKS Version
Auto PAC URL
Default
These are the settings that are used when no patterns match an URL
5
Move Up
Move Down
Add New Proxy
Edit Selection
Copy Selection
Delete Selection
Import your proxies from FoxyProxy on Mozilla Firefox or from another computer.
Please Donate
Buy Proxy Service

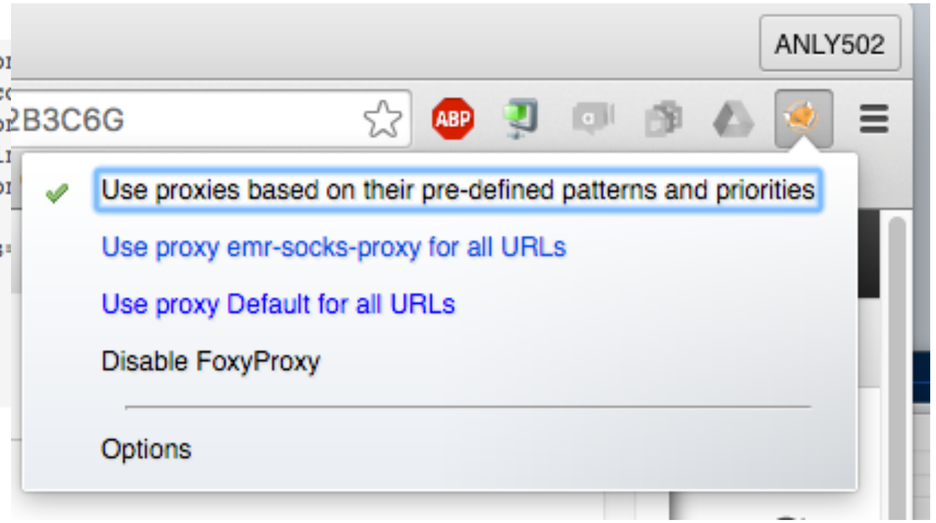

AWS Elastic MapReduce M
FoxyProxy options
ANLY502
chrome-extension://gcknhkkoolaabfmlnjonogaaifnjlfnp/options.html#tabProxies
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
FOXY PROXY
Proxies
Global Settings
Import/Export
QuickAdd
About
Proxy mode: Disable FoxyProxy
Import / Export
Import FoxyProxy settings from a file
Choose File
No file chosen
This tool allows you to import your proxies from FoxyProxy on Mozilla Firefox or from another computer.
Export FoxyProxy settings to a file
Export
A file named FoxyProxy-export.fpx will be saved in your default Chrome Downloads directory. You can use it to import your settings into another instance of Google Chrome or FoxyProxy on Firefox.
Please Donate
Buy Proxy Service


AWS Elastic MapReduce M
FoxyProxy options
ANLY502
chrome-extension://gcknhkkoolaabfmlnjonogaaifnjlfnp/options.html#tabProxies
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
FOXY PROXY
Proxies
Global Settings
Import/Export
QuickAdd
About
Proxy mode: Disable FoxyProxy
Import / Export
Import FoxyProxy settings from a file
Choose File
No file chosen
This tool allows you to import your proxies from FoxyProxy
ozilla Firefox or from another computer.
Export FoxyProxy settings to a file
Export
A file named FoxyProxy-export.fpx will be saved in your default Chrome Downloads directory. You can use it to import your settings into another instance of Google Chrome or FoxyProxy on Firefox.
Please Donate
Buy Proxy Service


simsong
Search
ANLY502
Favorites
simsong
Desktop
Documents
Google Drive
GU ANLY 502
ANLY 502 Sp...
GMU CFRS 510
CFRS 510 Sp...
farch.15
taxes.15
Downloads
Presentations
prof.14
OneDrive - N...
OneDrive
Applications
Dropbox
Marian Garfin...
SLG Job Search
Privacy non-...
Travel
Today
Yesterday
Devices
Dance
Remote Disc
Shared
Shared Folder
Name
Size
Kind
Date Added
arch 13 bytes Alias --
Desktop 15 bytes Alias --
Pictures 16 bytes Alias --
Applications -- Folder Aug 6, 2015, 7:50 PM
Downloads -- Folder Aug 6, 2015, 7:50 PM
Movies -- Folder Aug 6, 2015, 8:03 PM
Music -- Folder Aug 6, 2015, 8:03 PM
Public -- Folder Aug 6, 2015, 8:12 PM
gits -- Folder Aug 6, 2015, 8:18 PM
slg -- Folder Aug 6, 2015, 8:18 PM
Sites -- Folder Aug 6, 2015, 8:19 PM
Dropbox -- Folder Aug 6, 2015, 10:11 PM
Google Drive -- Folder Aug 7, 2015, 10:03 AM
from iMac -- Folder Oct 13, 2015, 9:31 AM
google-cloud-sdk -- Folder Oct 23, 2015, 8:23 PM
OneDrive -- Folder Nov 5, 2015, 9:40 PM
OneDrive - National Institute of Standards and Technology (NIST) -- Folder Nov 13, 2015, 6:13 PM
nist 1.1 MB Alias Dec 10, 2015, 9:36 PM
john-1.8.0-jumbo-1 -- Folder Dec 22, 2015, 4:47 PM
VMs -- Folder Jan 1, 2016, 9:30 PM
SIFT Workstation VMWare-SLG.vmwarevm 20.55 GB VMBundle Jan 1, 2016, 9:31 PM
mobile -- Folder Jan 2, 2016, 5:35 PM
Documents -- Folder Jan 25, 2016, 9:42 PM
Windows 10 x64 Educational.vmwarevm 24.18 GB VMBundle Feb 11, 2016, 7:24 AM
Metadata -- Folder Feb 18, 2016, 8:01 PM
mobile2 -- Folder Feb 23, 2016, 7:27 PM
mobile3 -- Folder Feb 23, 2016, 7:29 PM
M04 Jean Disk Image -- Folder Feb 23, 2016, 11:28 PM
MS-DOS 2.vmwarevm 68 KB VMBundle Feb 25, 2016, 8:38 PM
PS03 -- Folder Mar 5, 2016, 9:38 PM
cloudera-quickstart-vm-5.5.0-0-vmware -- Folder Mar 10, 2016, 3:21 PM
foxyproxy-settings.xml.txt 2 KB Plain Text Today, 5:43 PM
Cancel
Open
Apps
Doc

Don't forget to set FoxyProxy to "Use proxies based on their pre-defined patterns and priorities."

```
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
                <match enabled="true" name="*.compute.internal*" pattern="*.c
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
                <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.i
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
            </matches>
            <manualconf host="localhost" port="8157" socksversion="5" isSocks
domain="" />
        </proxy>
    </proxies>
</foxyproxy>
```



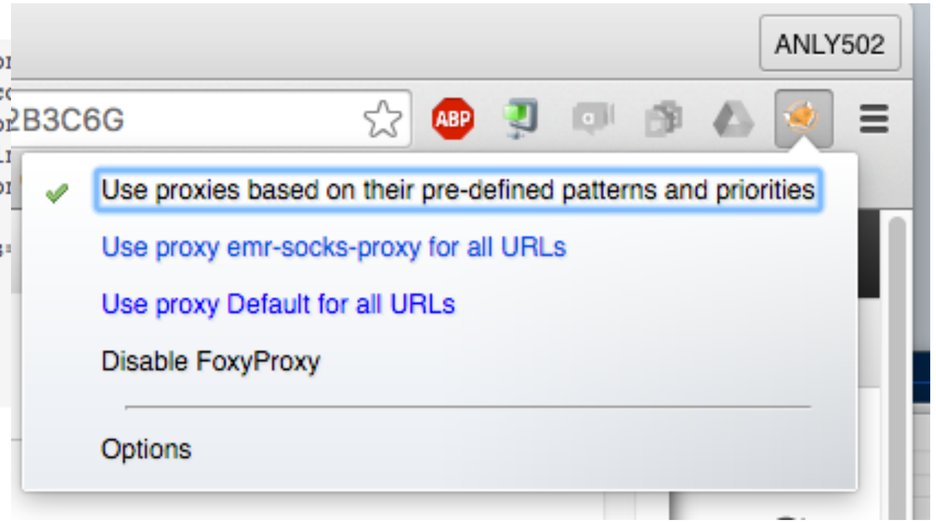
Finally:

```
[Dance ~ 17:47:46]$ ssh -ND 8157 hadoop@ec2-52-87-233-66.compute-1.amazonaws.com
```

There won't be any prompt; run this in a window by itself, or with "&"

Don't forget to set FoxyProxy to "Use proxies based on their pre-defined patterns and priorities."

```
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
                <match enabled="true" name="*.compute.internal*" pattern="*.c
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
                <match enabled="true" name="*.ec2.internal*" pattern="*.ec2.i
isBlackList="false" isMultiLine="false" caseSensitive="false" fromSubscriptio
            </matches>
            <manualconf host="localhost" port="8157" socksversion="5" isSocks
domain="" />
        </proxy>
    </proxies>
</foxyproxy>
```

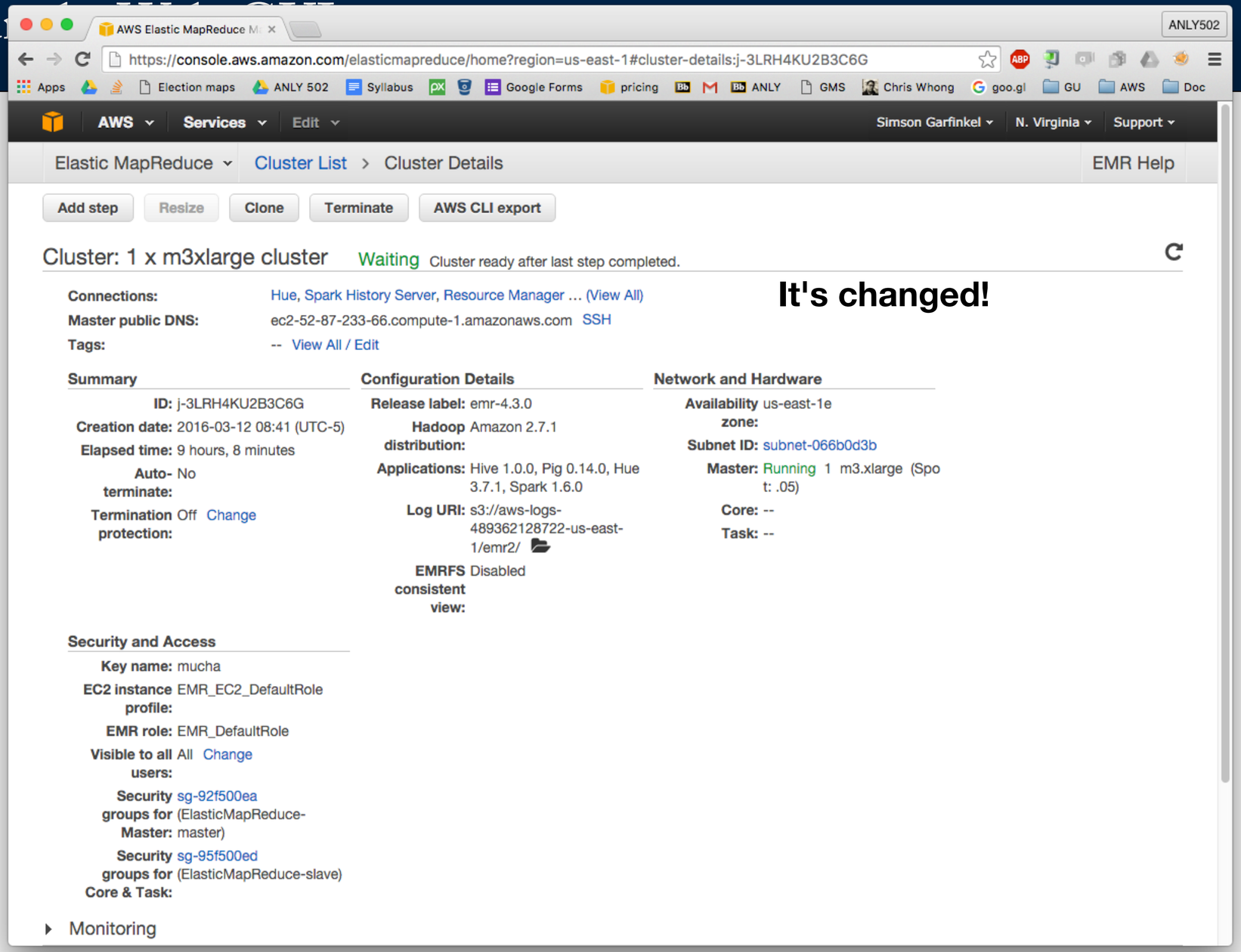


Finally:

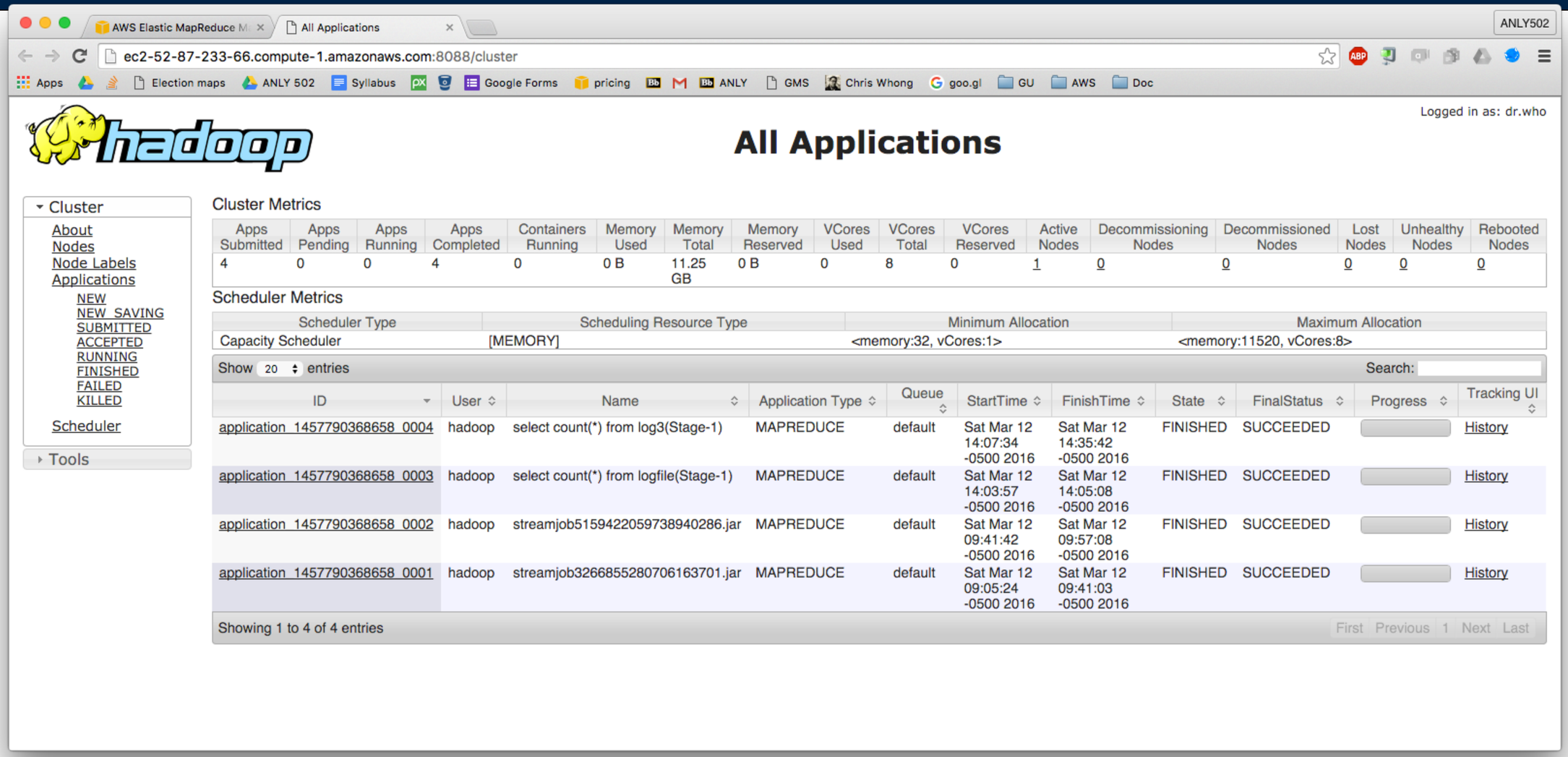
```
[Dance ~ 17:47:46]$ ssh -ND 8157 hadoop@ec2-52-87-233-66.compute-1.amazonaws.com
```

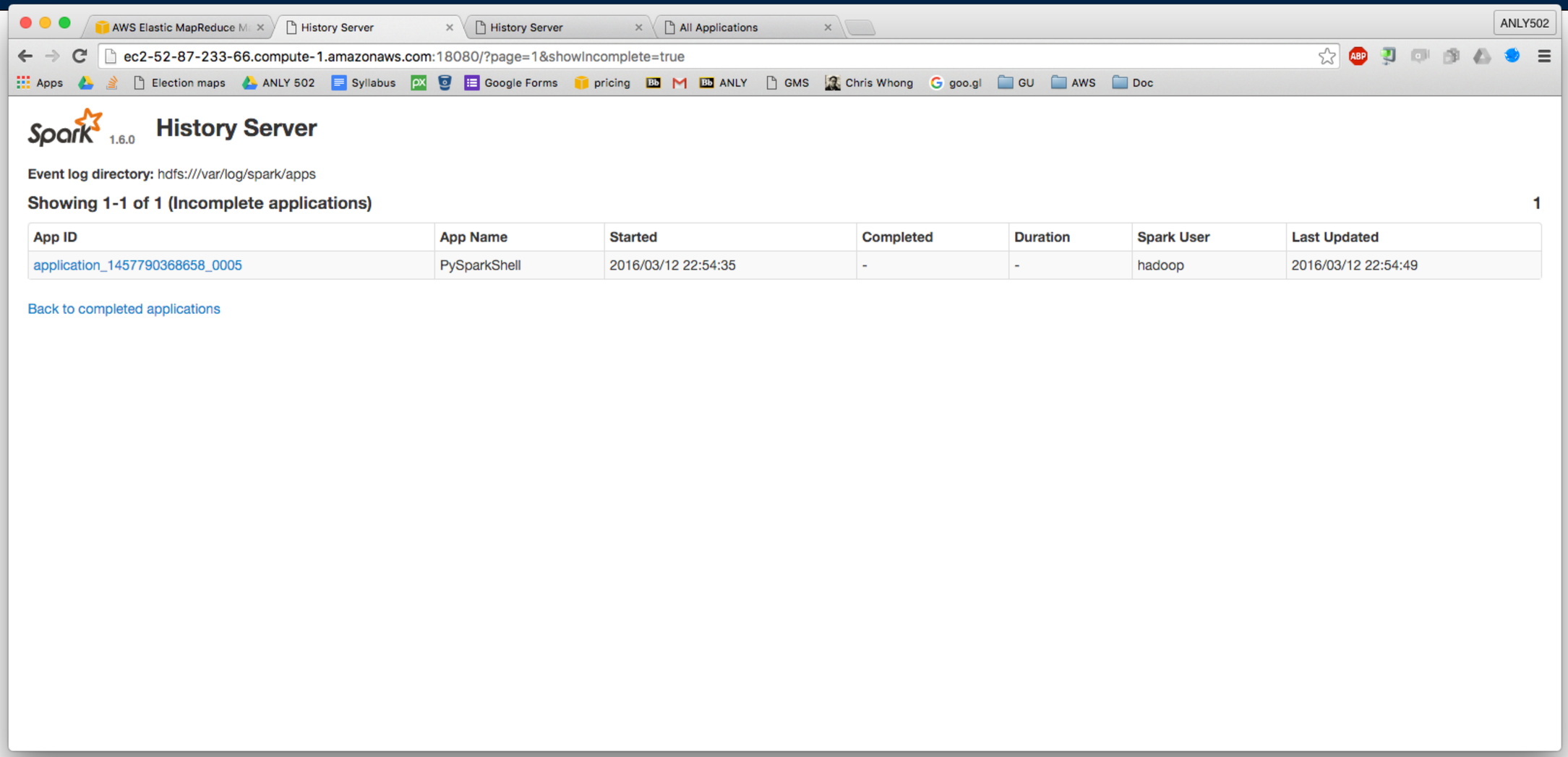
There won't be any prompt; run this in a window by itself, or with "&"

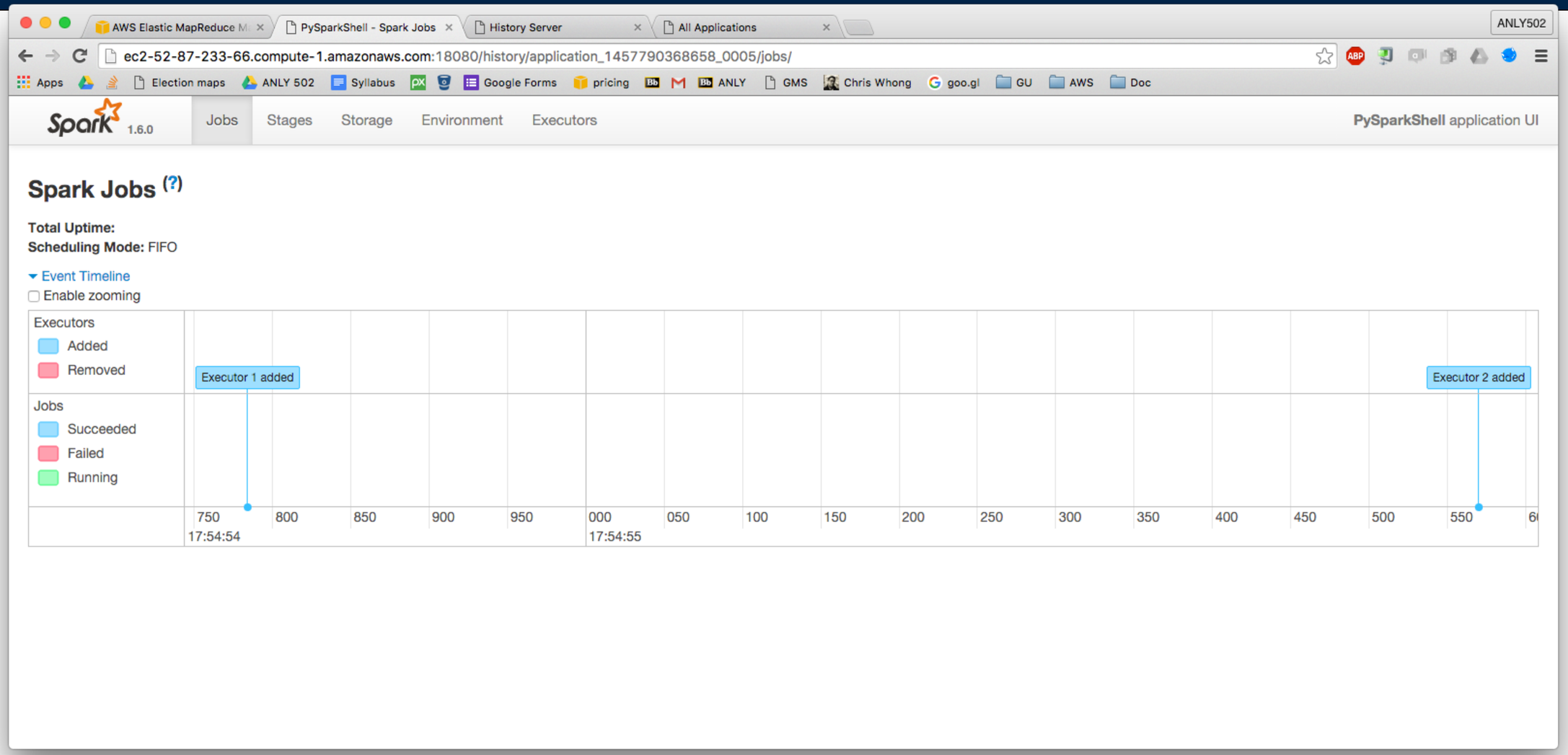
Now you can ru



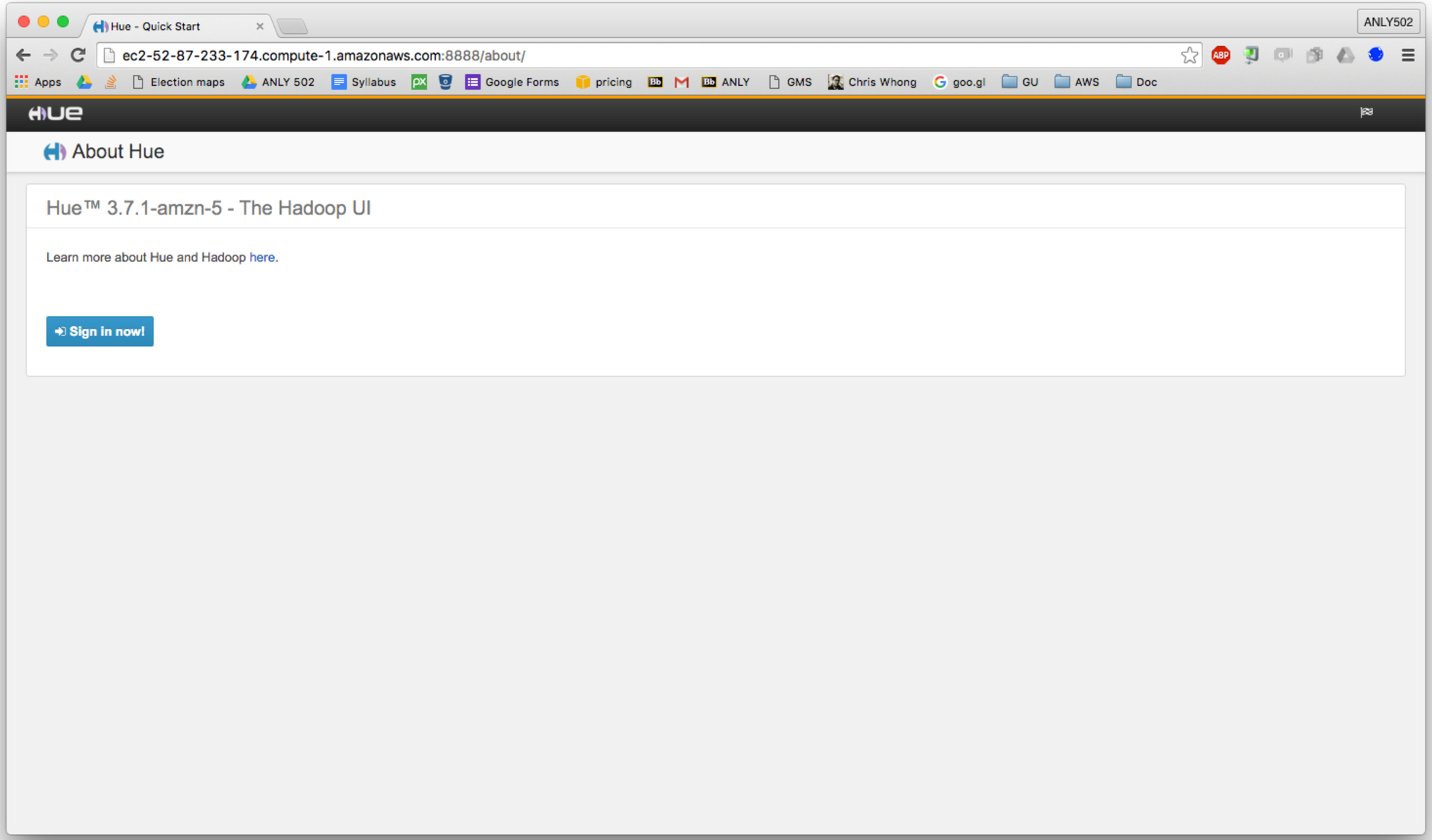
"Resource manager" is Yarn — it shows the running MapReduce jobs



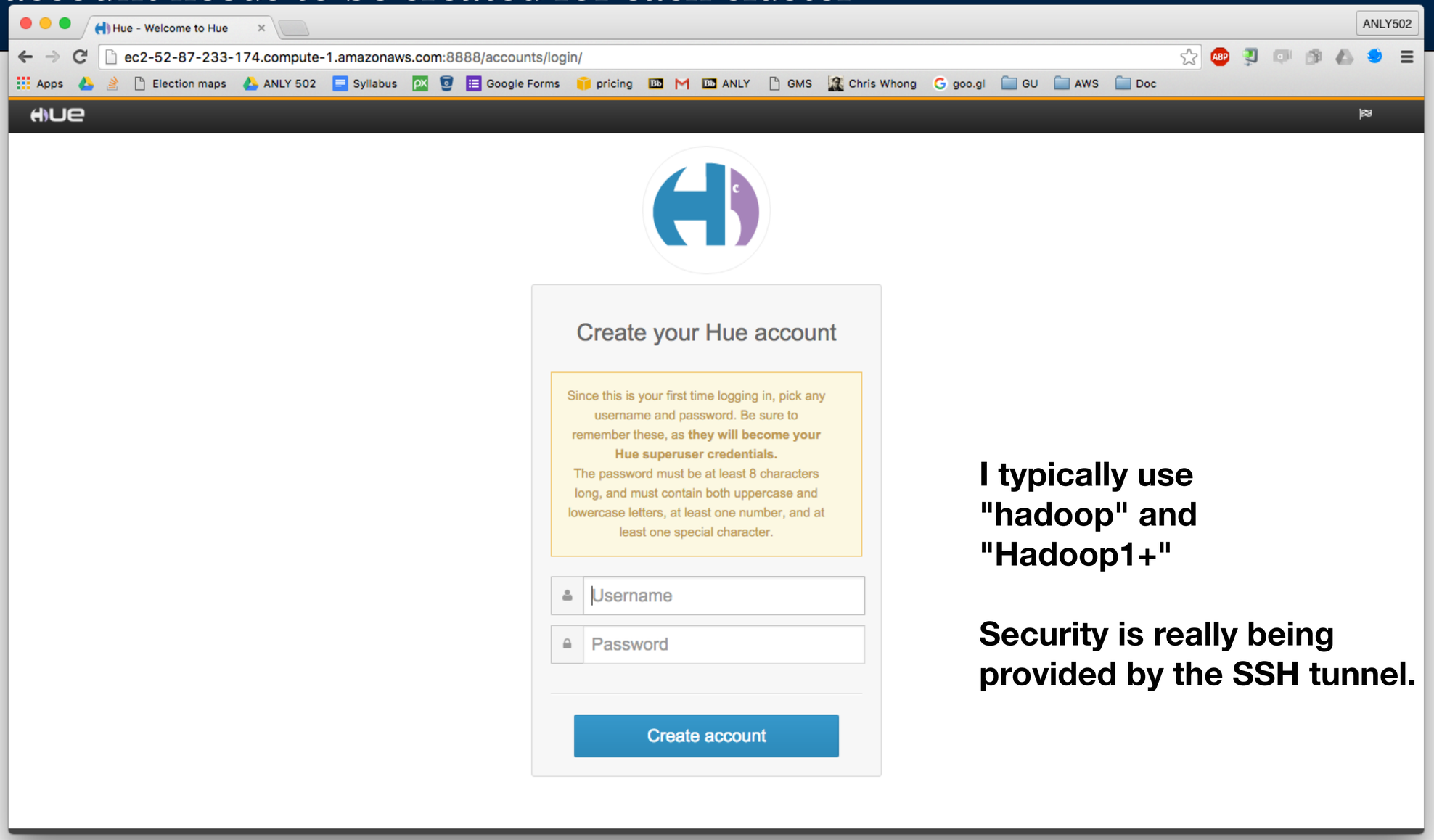

AWS Elastic MapReduce M
History Server
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:18080/?page=1&showIncomplete=true
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
Spark 1.6.0
History Server
Event log directory: hdfs:///var/log/spark/apps
Showing 1-1 of 1 (Incomplete applications)
1
App ID
App Name
Started
Completed
Duration
Spark User
Last Updated
application_1457790368658_0005
PySparkShell
2016/03/12 22:54:35
-
-
hadoop
2016/03/12 22:54:49
Back to completed applications


AWS Elastic MapReduce M
PySparkShell - Spark Jobs
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:18080/history/application_1457790368658_0005/jobs/
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
Spark 1.6.0
Jobs
Stages
Storage
Environment
Executors
PySparkShell application UI
Spark Jobs (?)
Total Uptime:
Scheduling Mode: FIFO
Event Timeline
Enable zooming
Executors
Added
Removed
Executor 1 added
Executor 2 added
Jobs
Succeeded
Failed
Running
750
17:54:54
800
850
900
950
000
17:54:55
050
100
150
200
250
300
350
400
450
500
550

Hue is the Hadoop User Environment

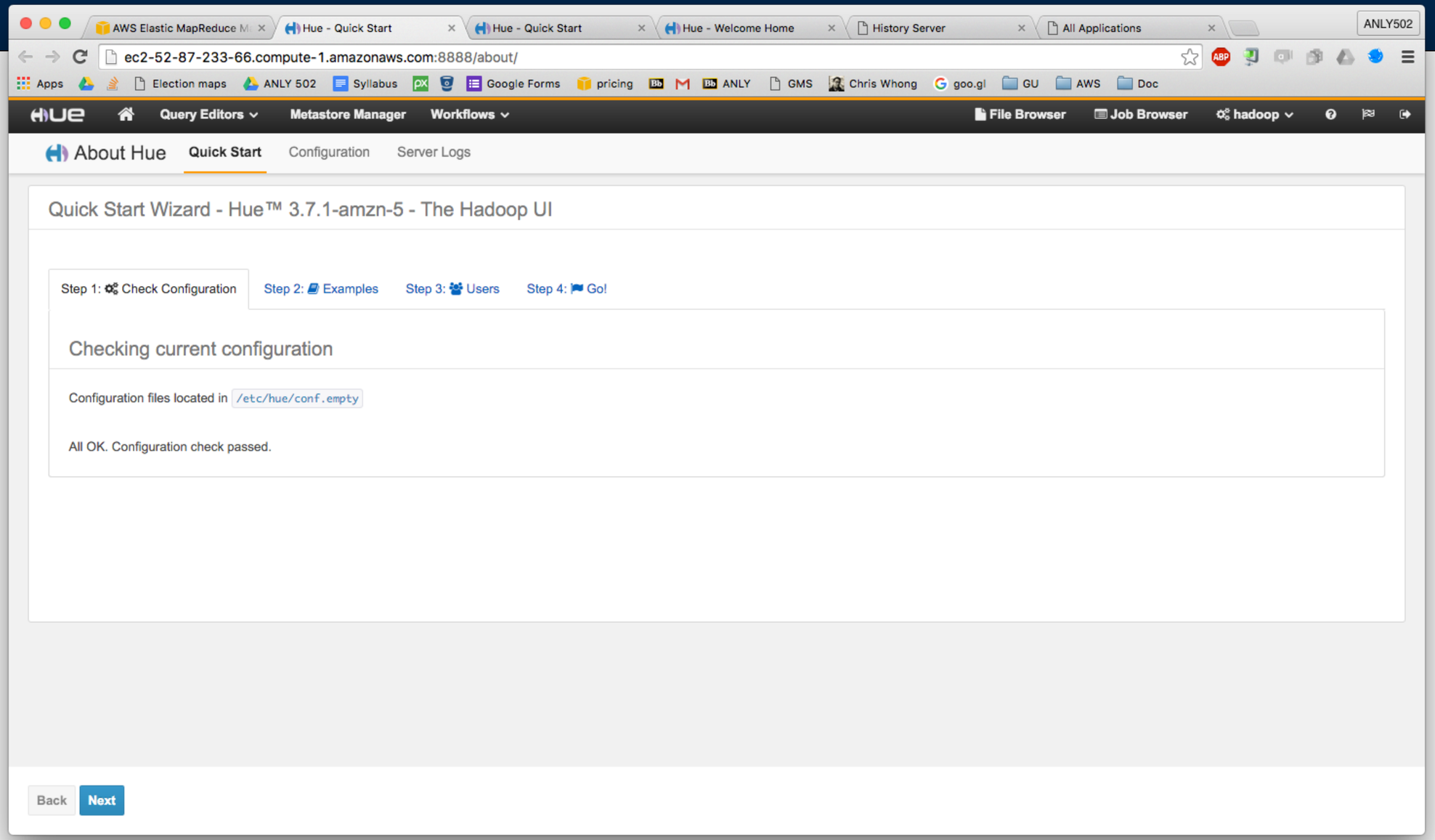


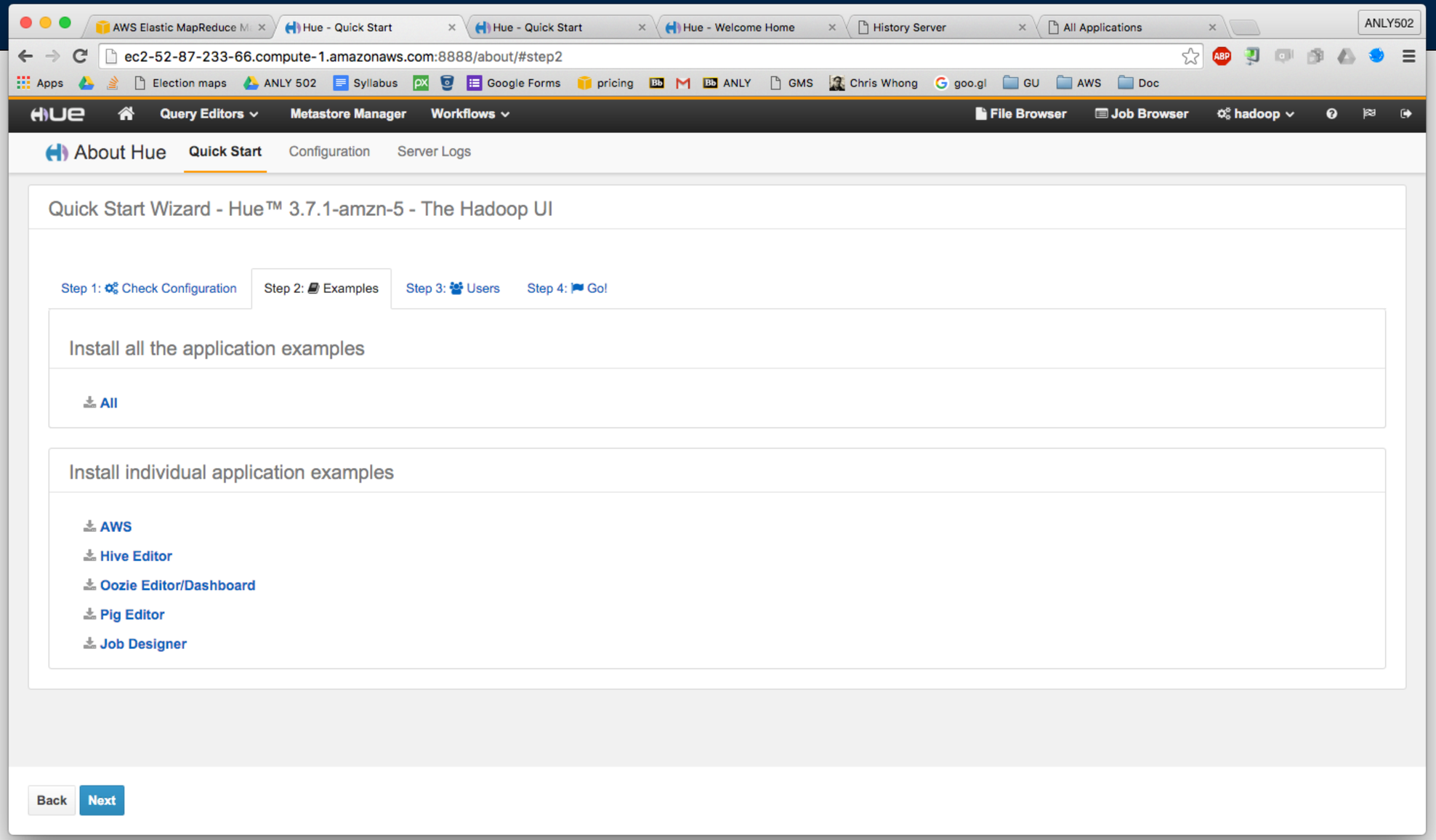
The Hue account needs to be created for each cluster

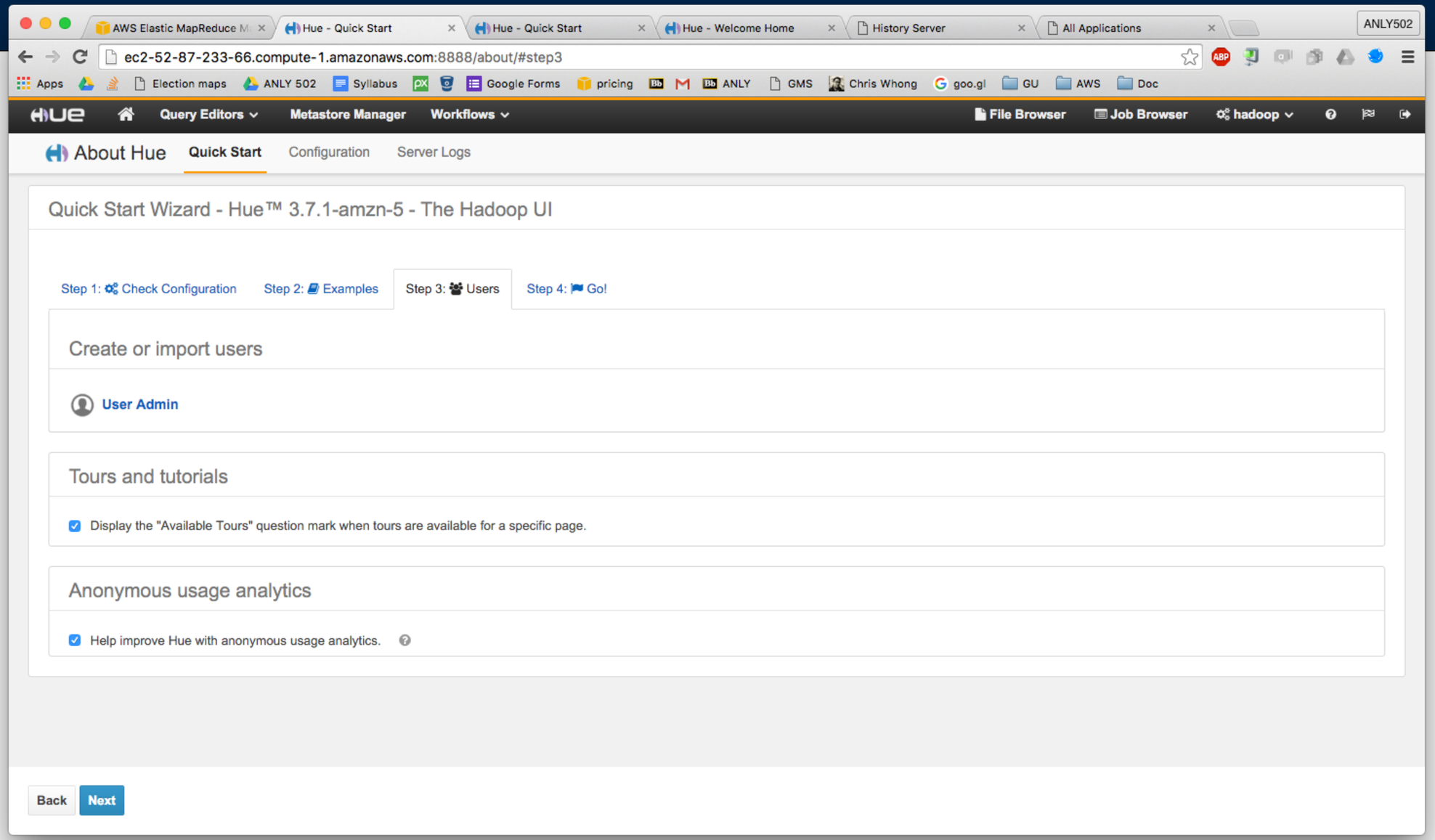


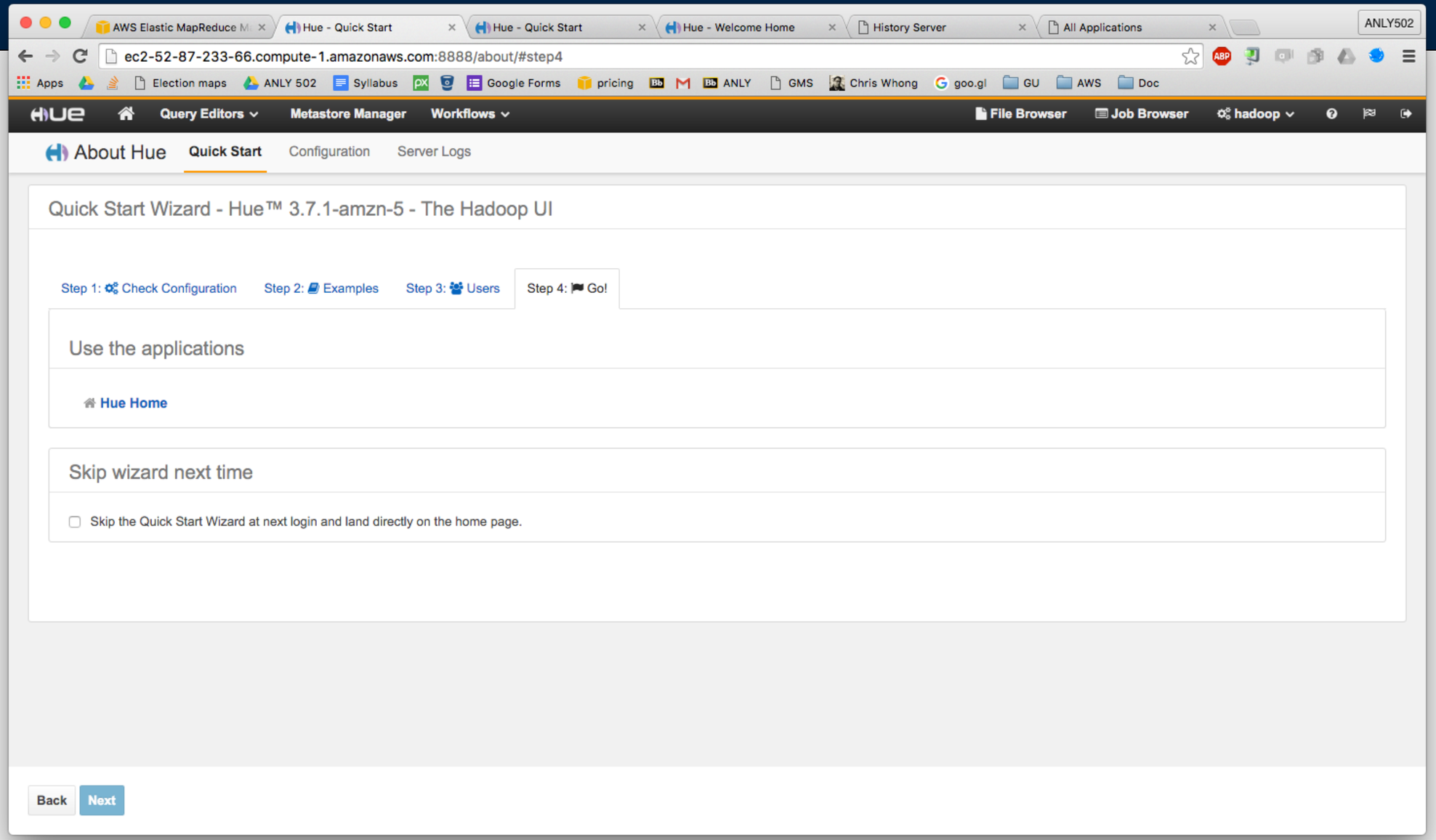
I typically use "hadoop" and "Hadoop1+"

Security is really being provided by the SSH tunnel.

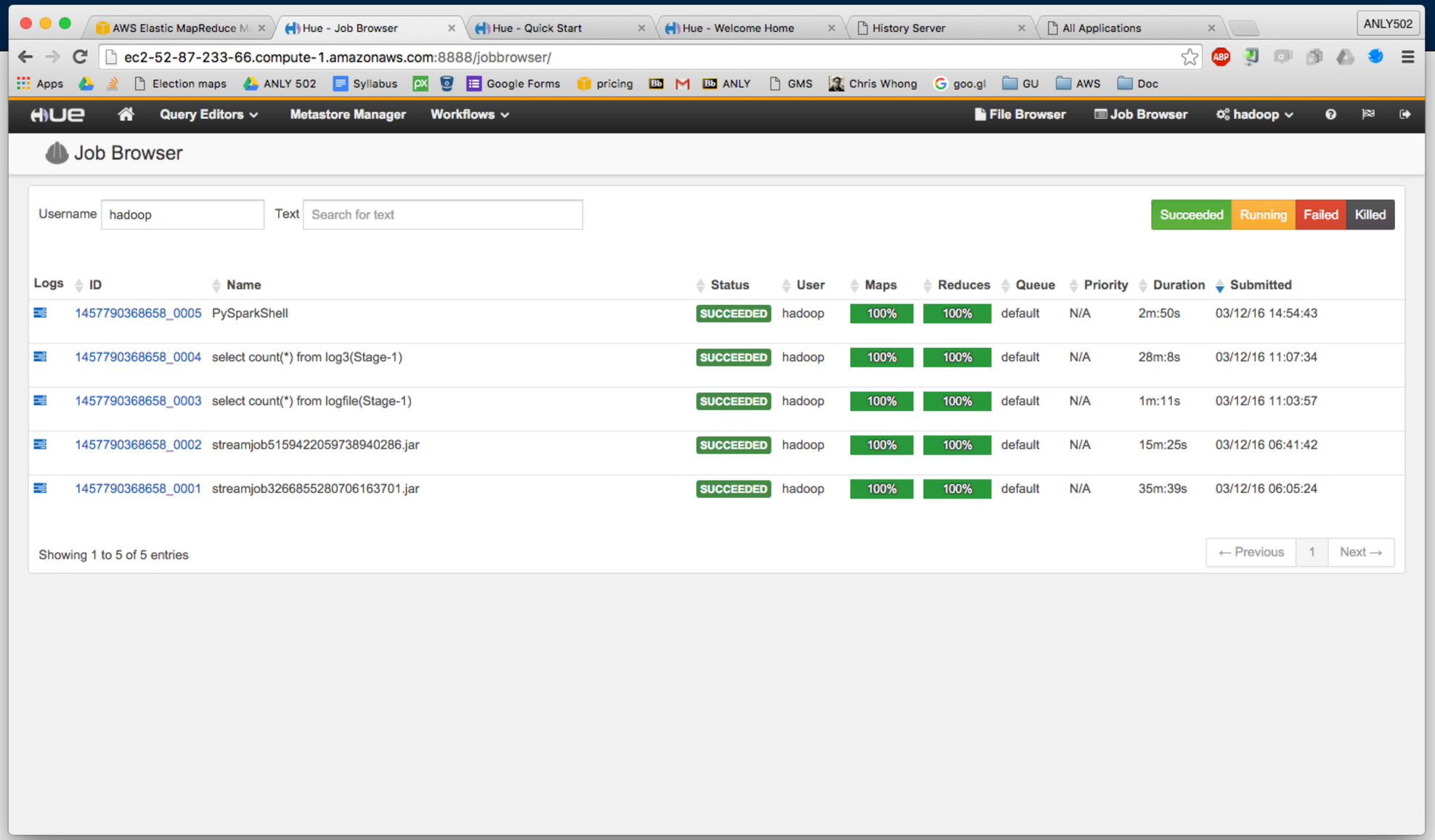

AWS Elastic MapReduce
Hue - Quick Start
Hue - Quick Start
Hue - Welcome Home
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/about/
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
About Hue
Quick Start
Configuration
Server Logs
Quick Start Wizard - Hue™ 3.7.1-amzn-5 - The Hadoop UI
Step 1: Check Configuration
Step 2: Examples
Step 3: Users
Step 4: Go!
Checking current configuration
Configuration files located in /etc/hue/conf.empty
All OK. Configuration check passed.
Back
Next


AWS Elastic MapReduce
Hue - Quick Start
Hue - Quick Start
Hue - Welcome Home
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/about/#step2
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
About Hue
Quick Start
Configuration
Server Logs
Quick Start Wizard - Hue™ 3.7.1-amzn-5 - The Hadoop UI
Step 1: Check Configuration
Step 2: Examples
Step 3: Users
Step 4: Go!
Install all the application examples
All
Install individual application examples
AWS
Hive Editor
Oozie Editor/Dashboard
Pig Editor
Job Designer
Back
Next

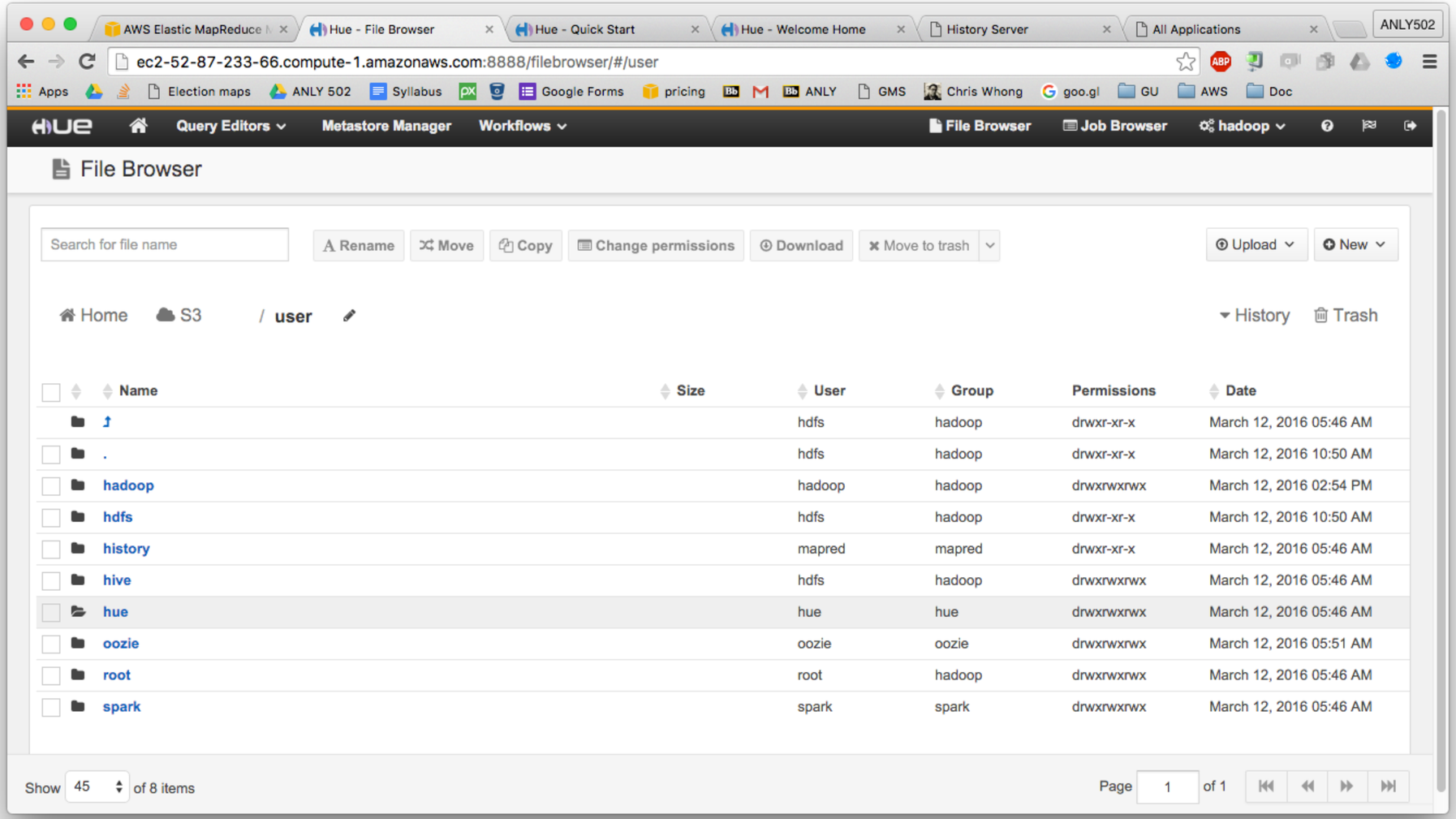

AWS Elastic MapReduce
Hue - Quick Start
Hue - Quick Start
Hue - Welcome Home
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/about/#step3
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
About Hue
Quick Start
Configuration
Server Logs
Quick Start Wizard - Hue™ 3.7.1-amzn-5 - The Hadoop UI
Step 1: Check Configuration
Step 2: Examples
Step 3: Users
Step 4: Go!
Create or import users
User Admin
Tours and tutorials
Display the "Available Tours" question mark when tours are available for a specific page.
Anonymous usage analytics
Help improve Hue with anonymous usage analytics.
Back
Next


AWS Elastic MapReduce
Hue - Quick Start
Hue - Quick Start
Hue - Welcome Home
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/about/#step4
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
About Hue
Quick Start
Configuration
Server Logs
Quick Start Wizard - Hue™ 3.7.1-amzn-5 - The Hadoop UI
Step 1: Check Configuration
Step 2: Examples
Step 3: Users
Step 4: Go!
Use the applications
Hue Home
Skip wizard next time
Skip the Quick Start Wizard at next login and land directly on the home page.
Back
Next

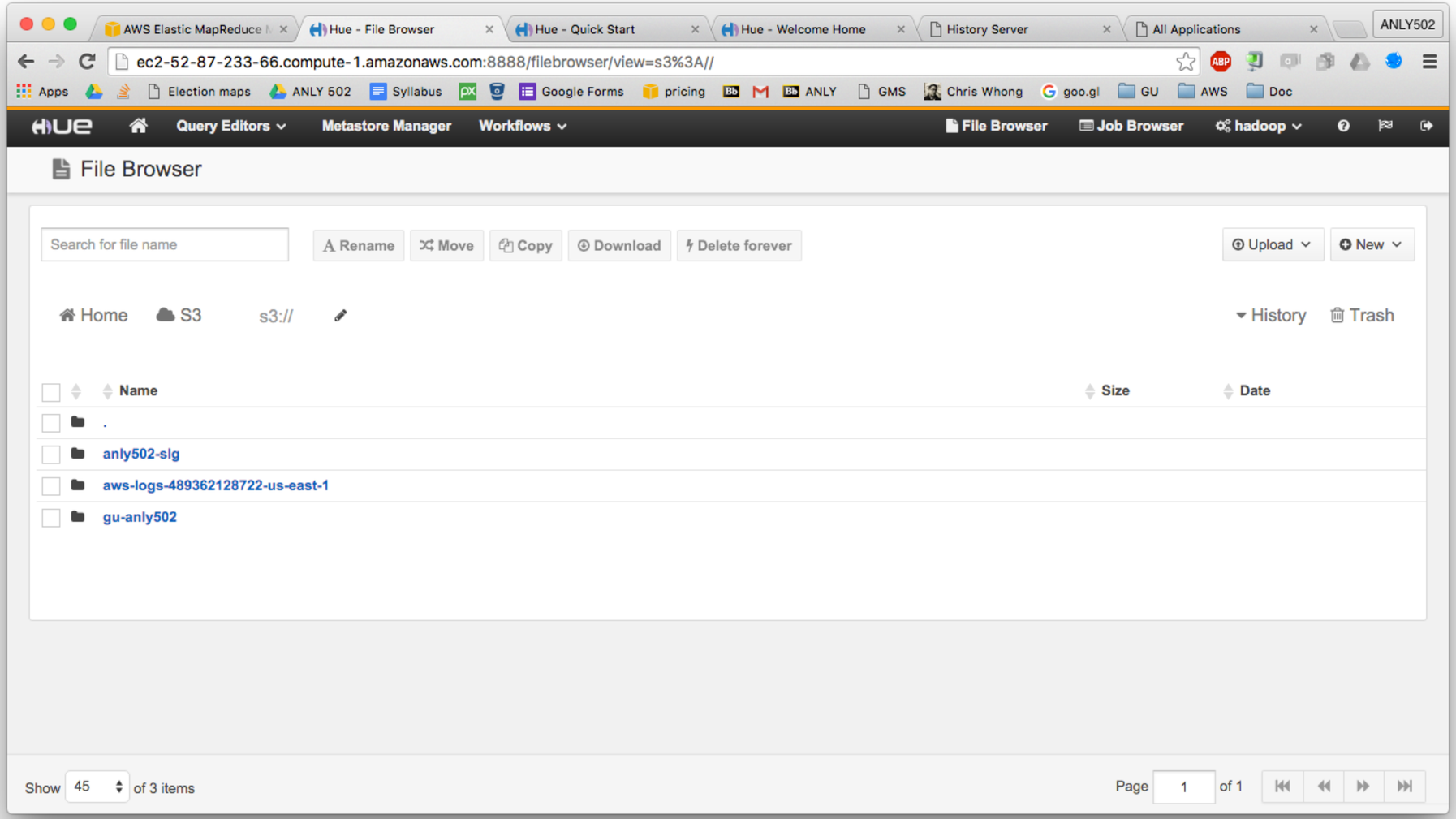
With Hue, you can easily monitor the progress of your MapReduce jobs.

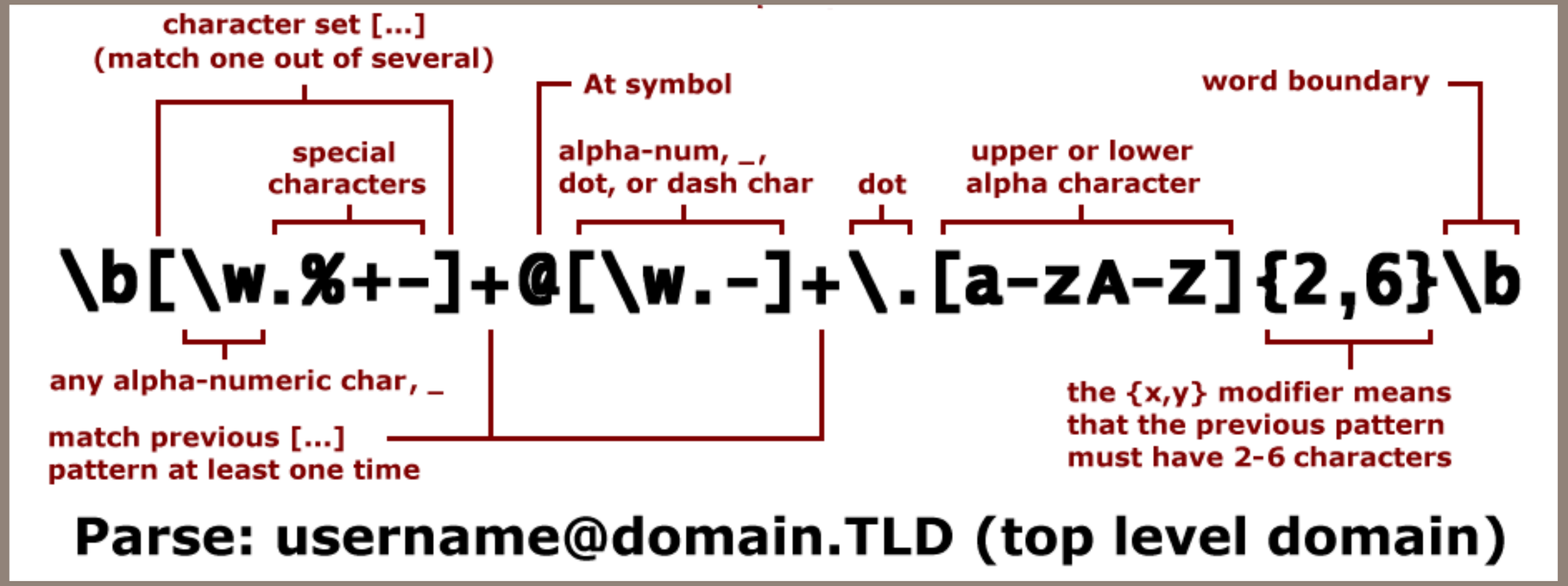


Hue has a HDFS browser



The Hue browser will also browse S3





Regular Expressions

Regular expressions are text strings that describe search patterns

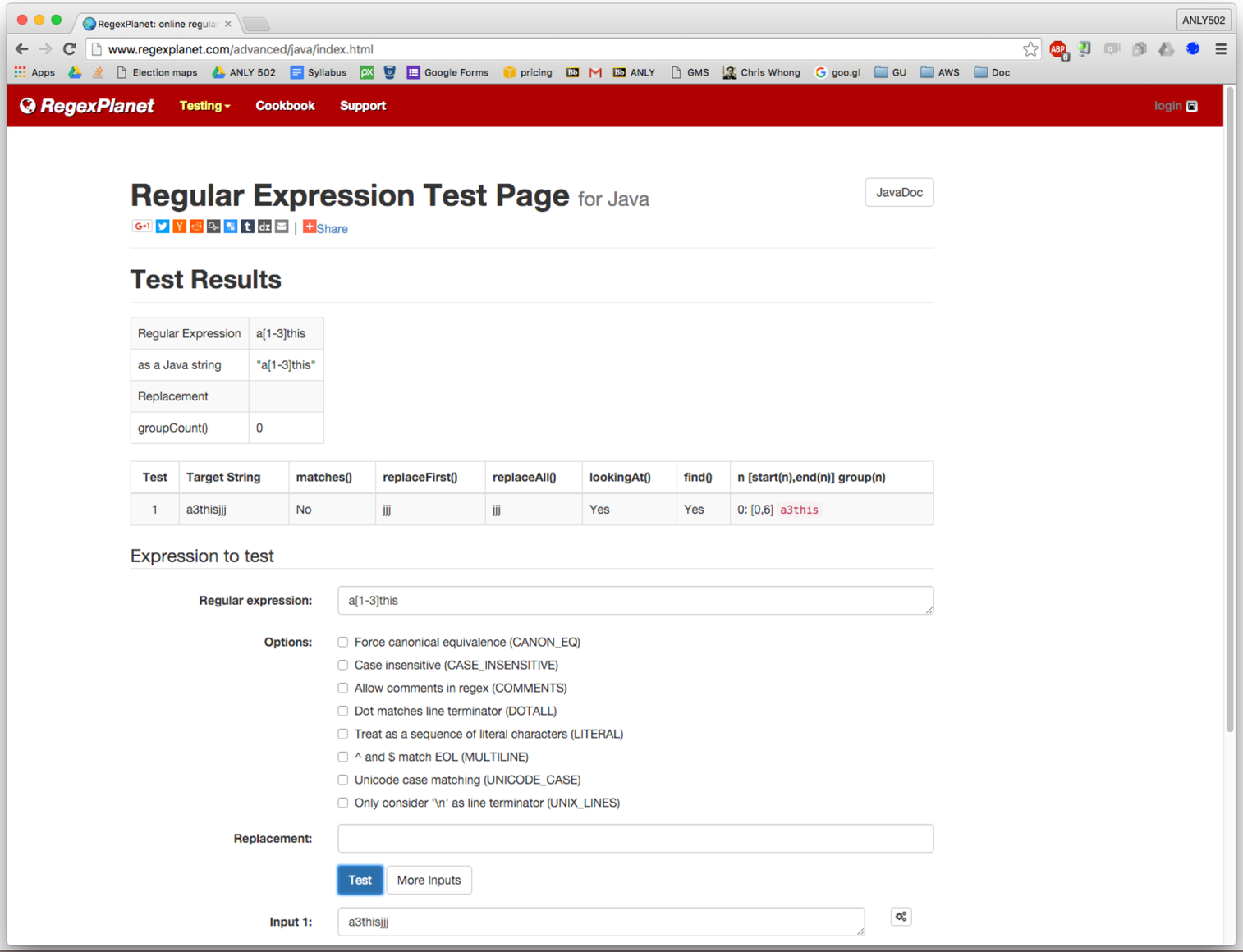
Python and Java both use regular expressions. They are *mostly* the same languages.

Common to both Java and Python:

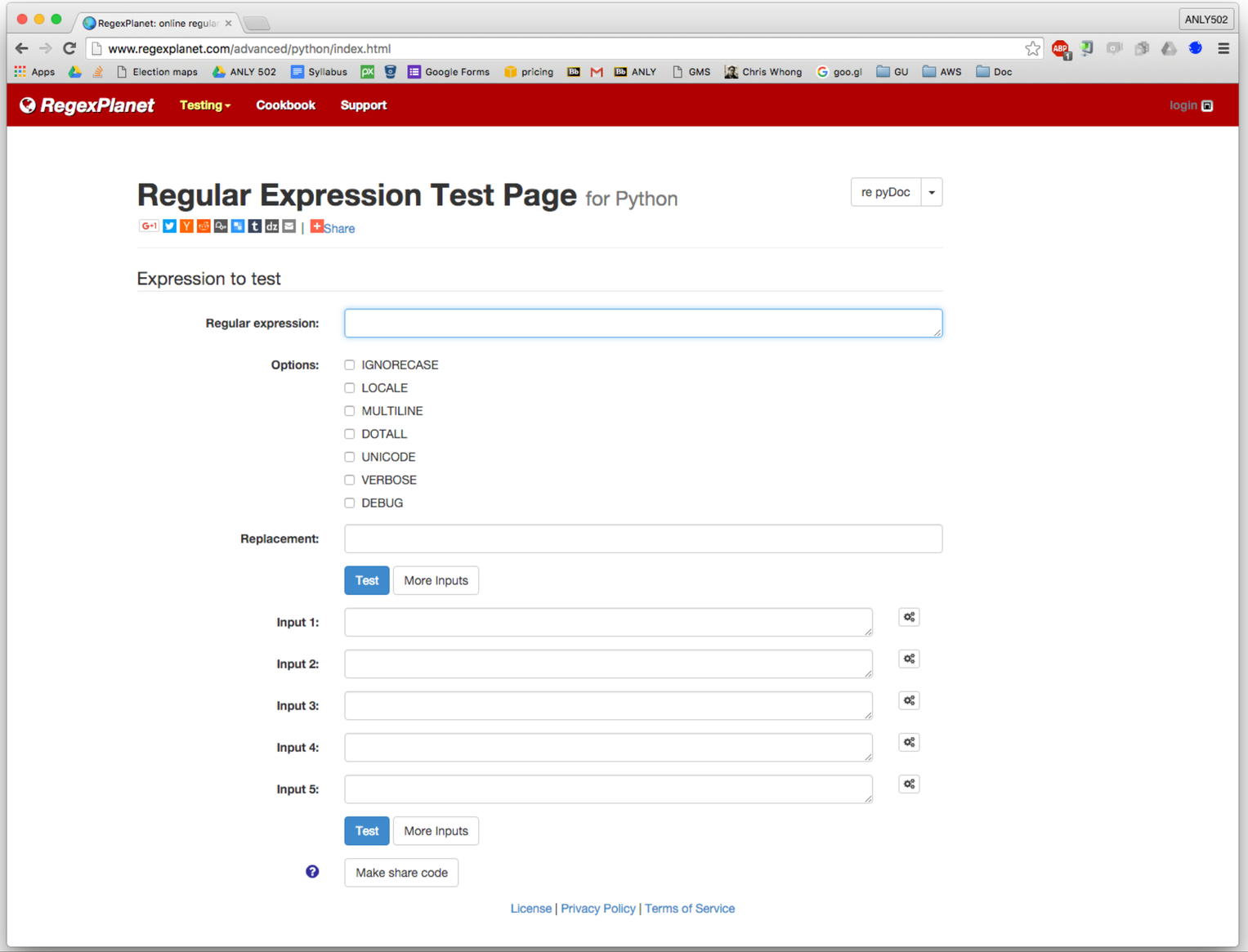
String	Matches
x	*Match character* x *(unless x is special)*
.	*any character*
[abc]	a, b *or* c
[a-z]	*characters* a-z
x?	*nothing* or x
x*	*nothing or any number of* x
x+	*one or more* x
xIy	x *or* y
^x	x *at the beginning of a line*
x$	x *at the end of a line*

RegexPlanet has an online regular expression tester

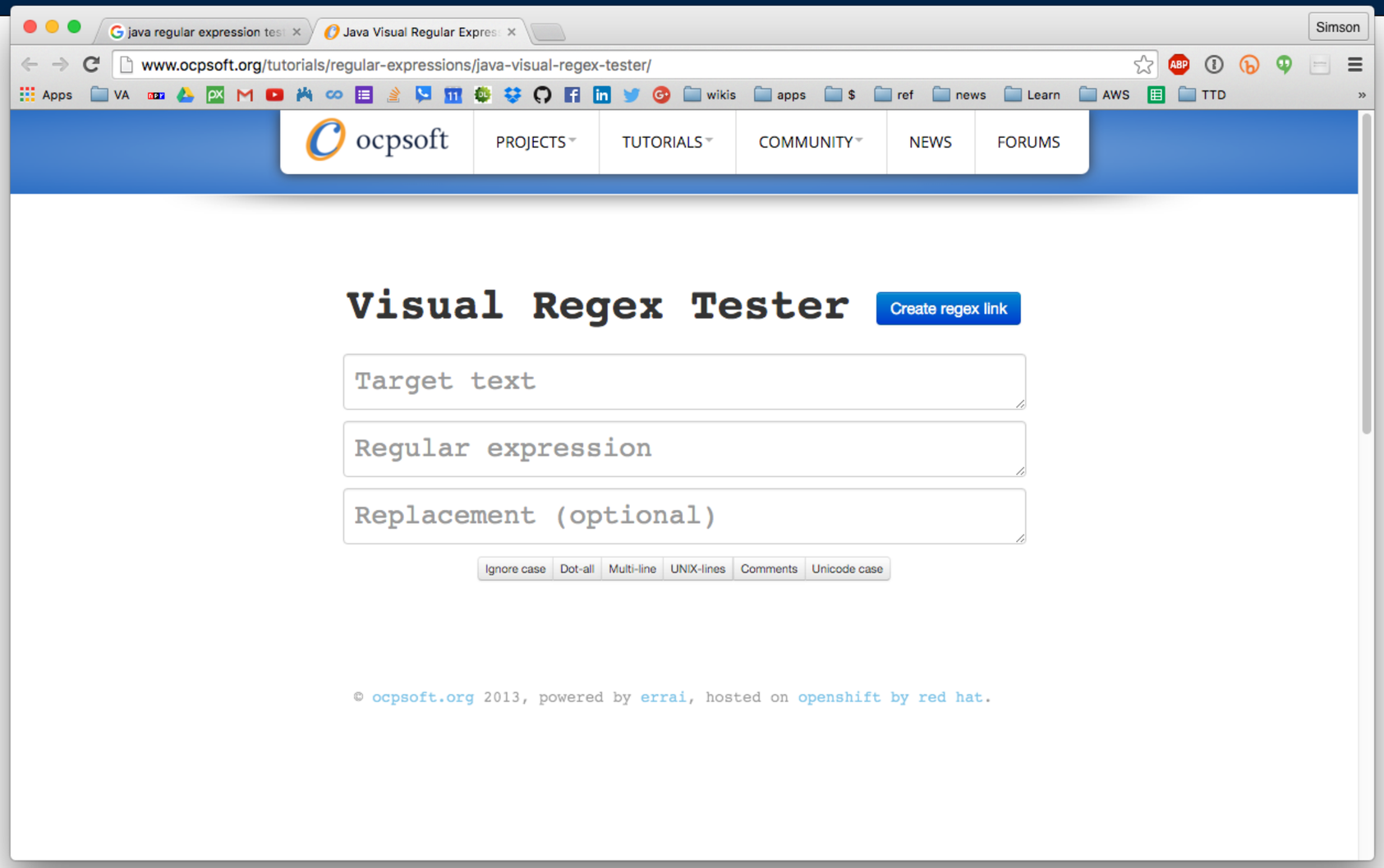
Java:

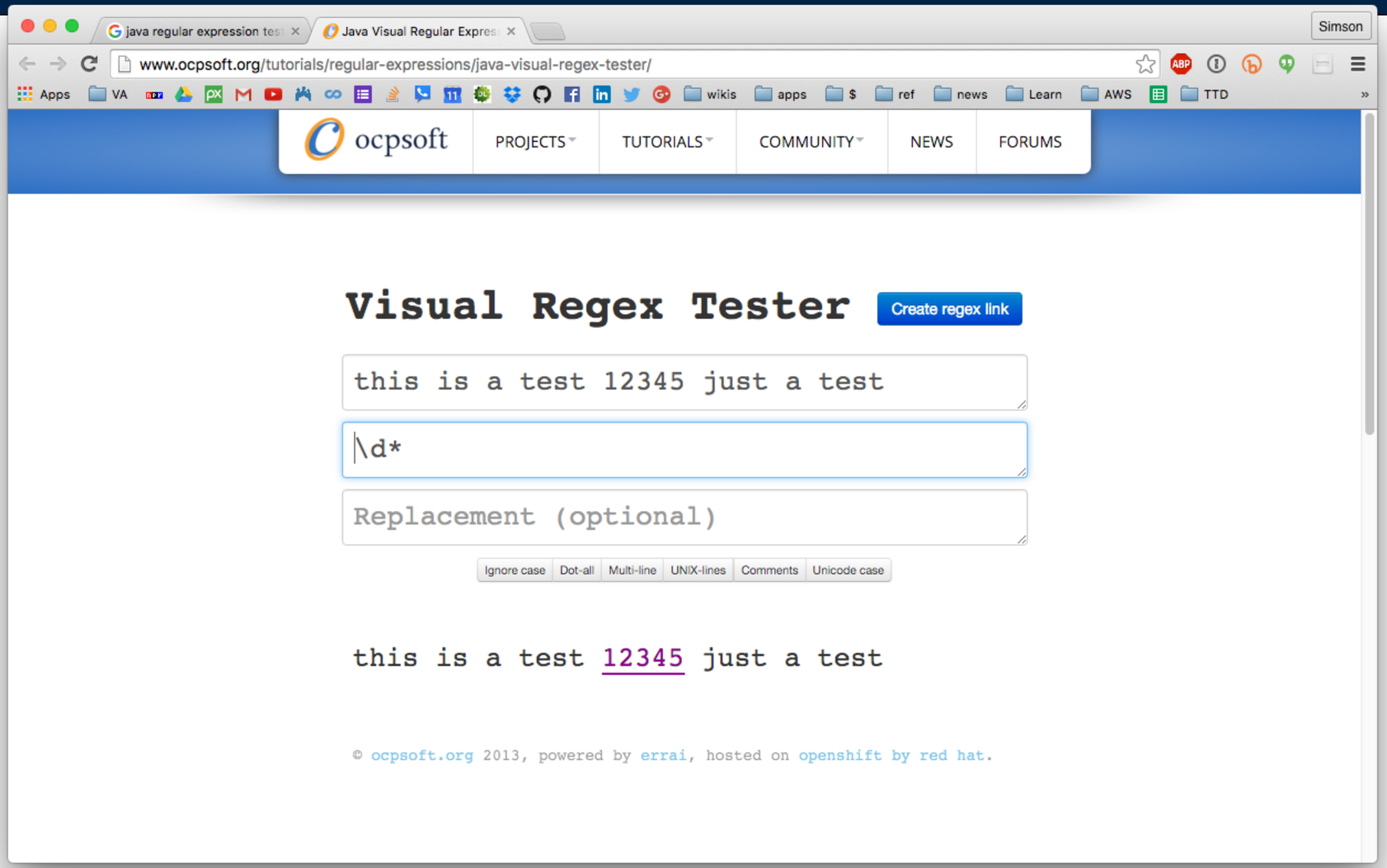


Python


RegexPlanet: online regular
ANLY502
www.regexplanet.com/advanced/python/index.html
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
RegexPlanet
Testing
Cookbook
Support
login
Regular Expression Test Page for Python
re pyDoc
Share
Expression to test
Regular expression:
Options:
IGNORECASE
LOCALE
MULTILINE
DOTALL
UNICODE
VERBOSE
DEBUG
Replacement:
Test
More Inputs
Input 1:
Input 2:
Input 3:
Input 4:
Input 5:
Test
More Inputs
Make share code
License | Privacy Policy | Terms of Service

http://www.ocpsoft.org/tutorials/regular-expressions/java-visual-regex-tester/




java regular expression tes
Java Visual Regular Expres
Simson
www.ocpsoft.org/tutorials/regular-expressions/java-visual-regex-tester/
Apps
VA
wikis
apps
$
ref
news
Learn
AWS
TTD
ocpsoft
PROJECTS
TUTORIALS
COMMUNITY
NEWS
FORUMS
Visual Regex Tester
Create regex link
this is a test 12345 just a test
\d*
Replacement (optional)
Ignore case
Dot-all
Multi-line
UNIX-lines
Comments
Unicode case
this is a test 12345 just a test
© ocpsoft.org 2013, powered by errai, hosted on openshift by red hat.

Introducing Hive

Apache Hive

A standard interface for data analysis in Hadoop

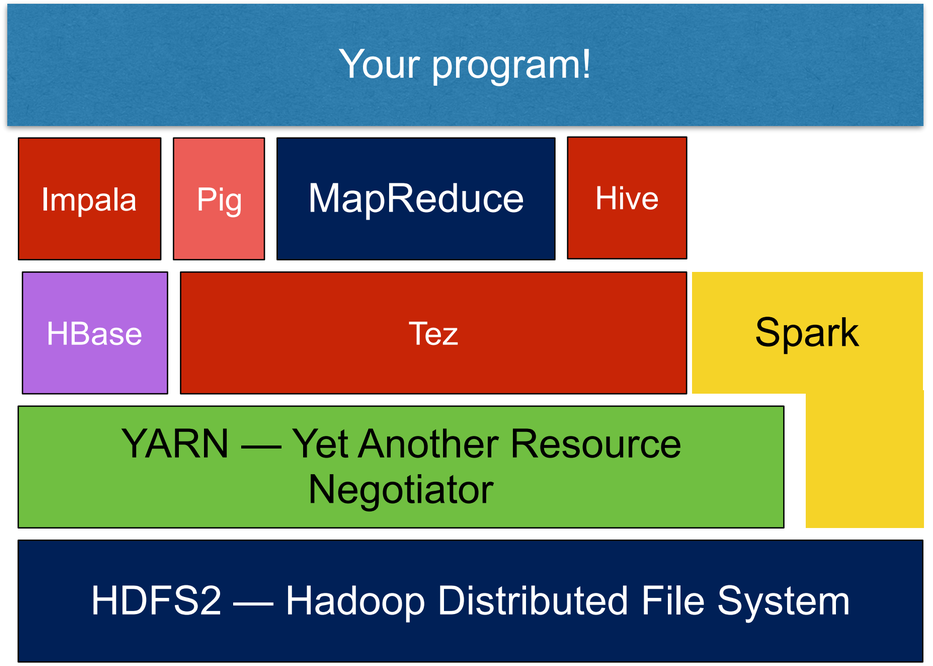
SQL-like language called HiveQL for querying data.

Scalable — Turns HQL queries into MapReduce Jobs

Extensible

- Plug-ins for new data sources
- User Defined Functions
- Also works with HBase, Spark, etc.

Hive and Impala both provide SQL interfaces.
Hive may scale better.
See:
http://hortonworks.com/blog/impala-vs-hive-performance-benchmark/



Hive is Hadoop's "data warehouse."

Hive gives an SQL-like interface to "big data"

Hadoop provides:

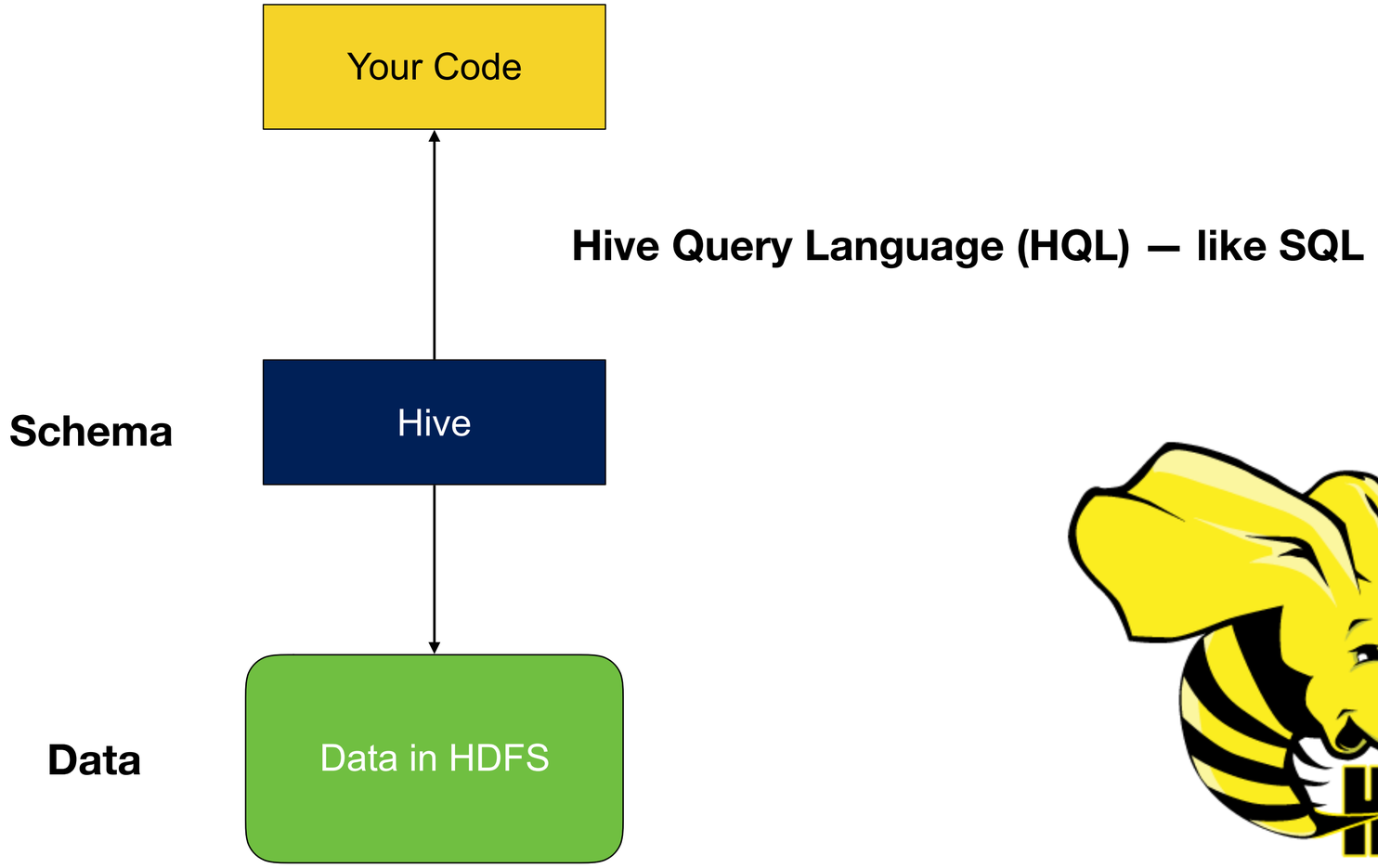
- Massive scale
- Fault tolerance

Hive provides:

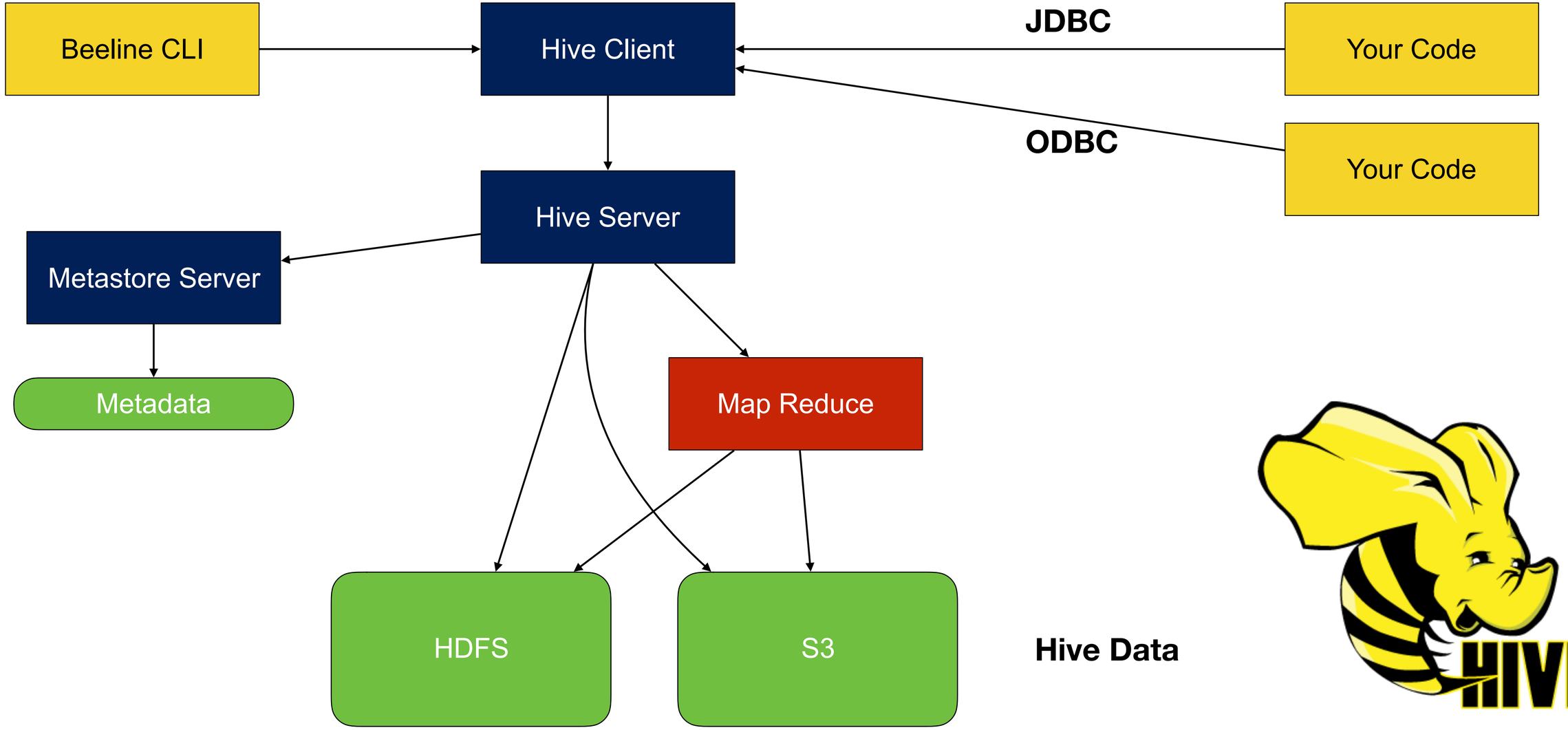
- Data summarization
- ad-hoc querying
- Simple query interface

But it's still map reduce:

- Batch processing



Hive is Hadoop's "data warehouse."



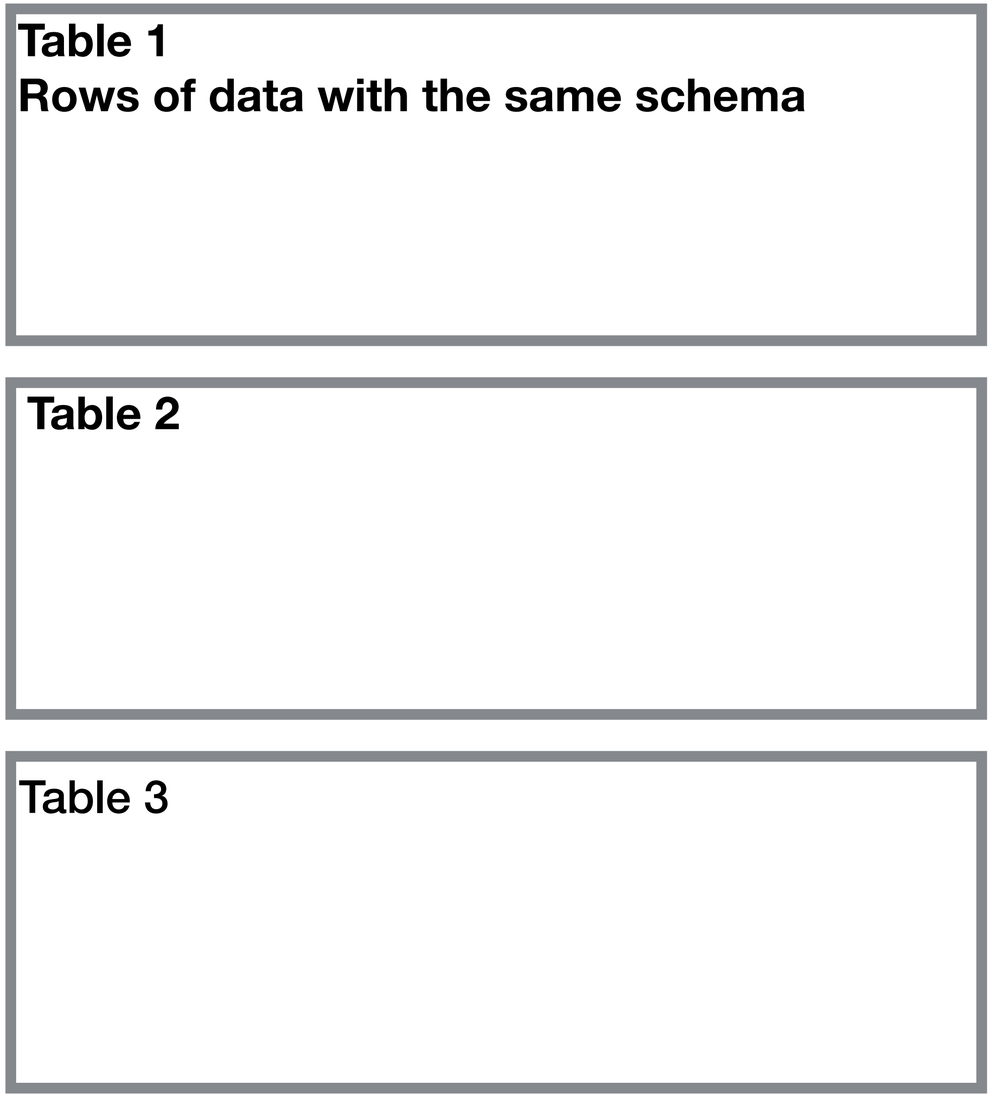
Hive does not store data. Hive points to data.

Hive does not store data. Hive points to data.

Database: Multiple Tables

Hive does not store data. Hive points to data.

Database: Multiple Tables



Hive does not store data. Hive points to data.

Database: Multiple Tables

Table 1
Rows of data with the same schema

Table 2

Table 3

Partitions.
Refers to a subset of data in the table

Hive data types — similar to Pig

"Types are associated with the columns in the tables. The following Primitive types are supported:"

- Integers
  - *TINYINT - 1 byte integer*
  - *SMALLINT - 2 byte integer*
  - *INT - 4 byte integer*
  - *BIGINT - 8 byte integer*
- Boolean type
  - *BOOLEAN - TRUE/FALSE*
- Floating point numbers
  - *FLOAT - single precision*
  - *DOUBLE - Double precision*
- String type
  - *STRING - sequence of characters in a specified character set*

Command line interrface

Hive 1 — "hive" command

- hive command can run as either client or server.
- Supports interactive & batch modes (like pig)

```
$ hive
```

Works on Cloudera VM
Does not work on EMR

Hive 2 — "Hive" server; many clients.

- Clients communicate with server via JODBC or Thrift

```
$ beeline
beeline> !connect jdbc:hive2://localhost:10000
scan complete in 6ms
Connecting to jdbc:hive2://localhost:10000
Enter username for jdbc:hive2://localhost:10000: hadoop
Enter password for jdbc:hive2://localhost:10000: ******
Connected to: Apache Hive (version 1.0.0-amzn-2)
Driver: Hive JDBC (version 1.0.0-amzn-2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000>
```

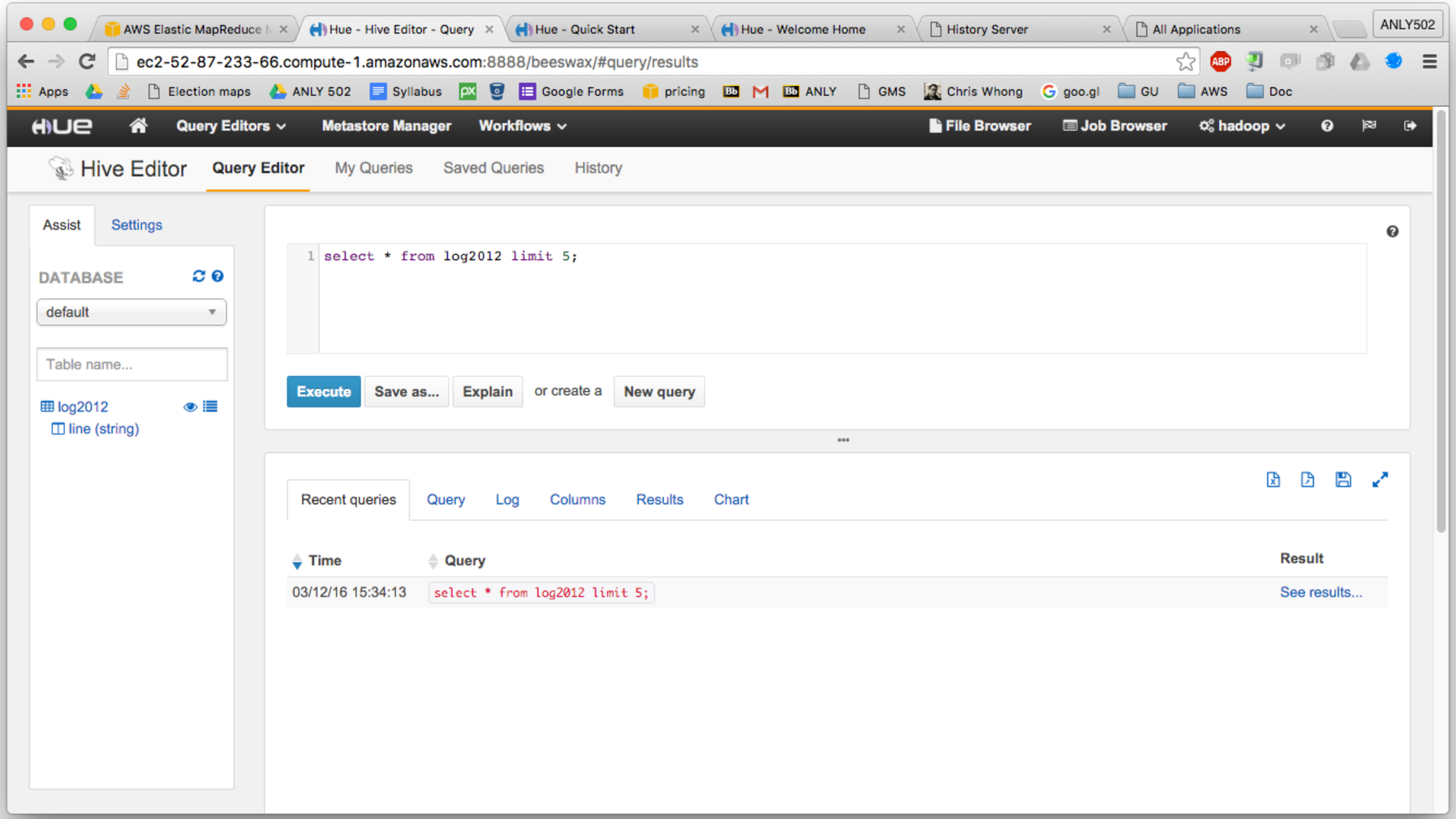
EMR: hadoop/hadoop
Cloudera VM: cloudera/cloudera

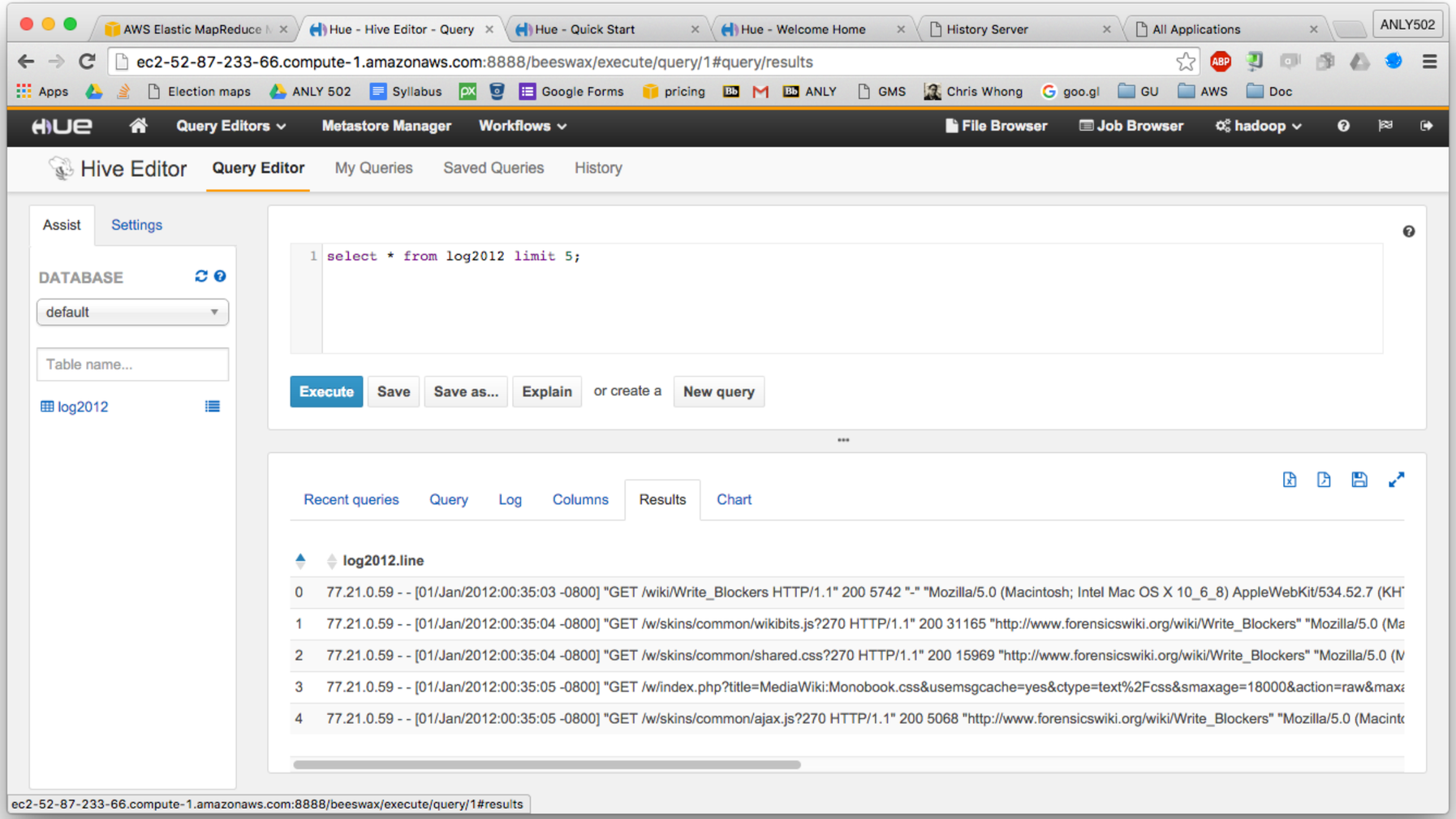
You can have multiple SQL connections at a time.

Create a table from the forensicswiki logfile

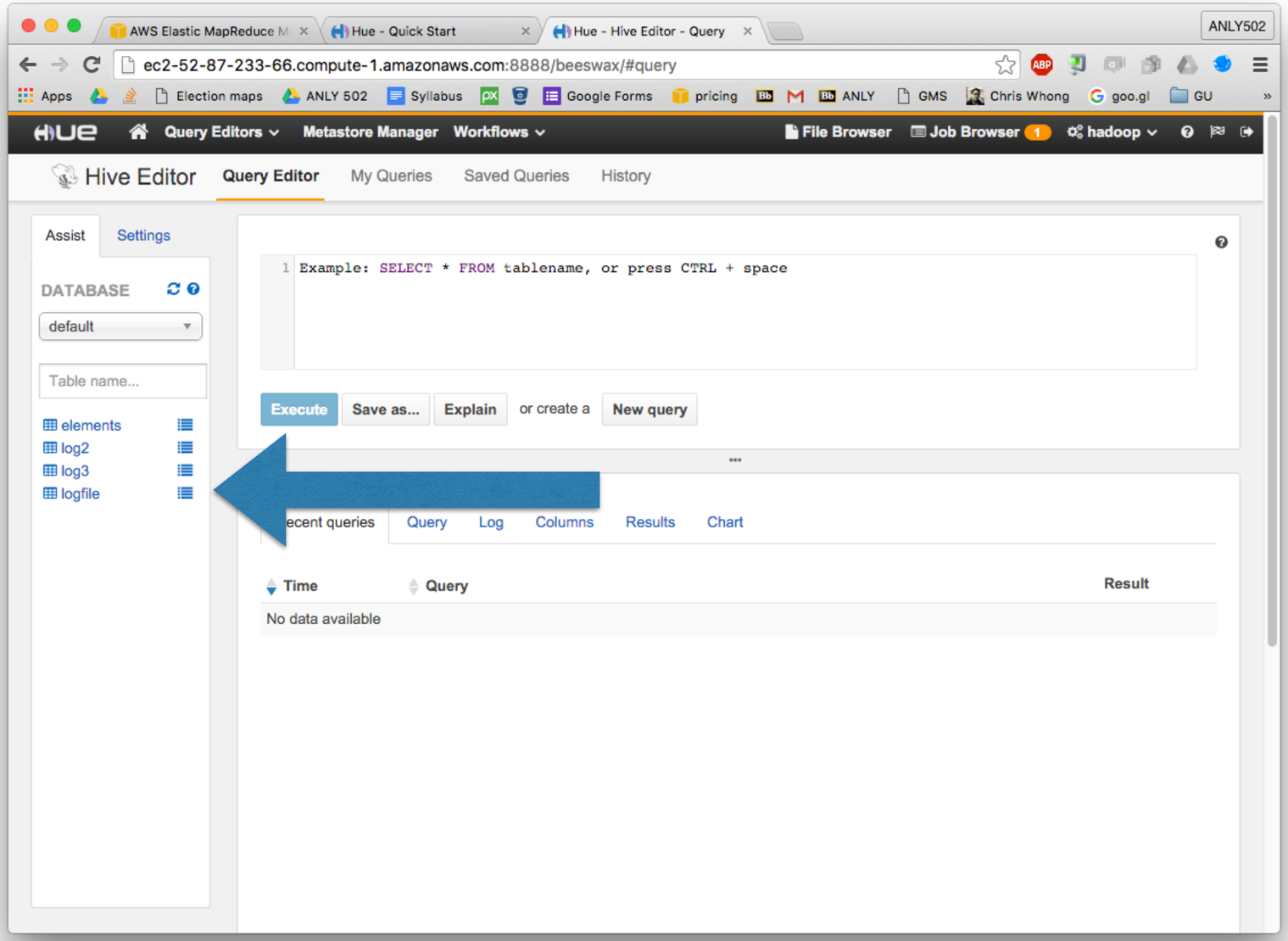
```
$ beeline
beeline> !connect jdbc:hive2://localhost:10000
...
0: jdbc:hive2://localhost:10000> create external table log2012 (line string) location 's3://gu-anly502/ps03/forensicswiki/';
No rows affected (0.721 seconds)
0: jdbc:hive2://localhost:10000> select * from log2012 limit 3;
+-----------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
--------------+--+
|
log2012.line
|
+-----------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
--------------+--+
| 77.21.0.59 - - [01/Jan/2012:00:35:03 -0800] "GET /wiki/Write_Blockers HTTP/1.1" 200 5742 "-" "Mozilla/5.0 (Macintosh; Intel Mac
OS X 10_6_8) AppleWebKit/534.52.7 (KHTML, like Gecko) Version/5.1.2 Safari/534.52.7"
|
| 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/wikibits.js?270 HTTP/1.1" 200 31165 "http://
www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/534.52.7 (KHTML, like Gecko)
Version/5.1.2 Safari/534.52.7"   |
| 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/shared.css?270 HTTP/1.1" 200 15969 "http://
www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/534.52.7 (KHTML, like Gecko)
Version/5.1.2 Safari/534.52.7"    |
+-----------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
--------------+--+
3 rows selected (6.632 seconds)
0: jdbc:hive2://localhost:10000>
```

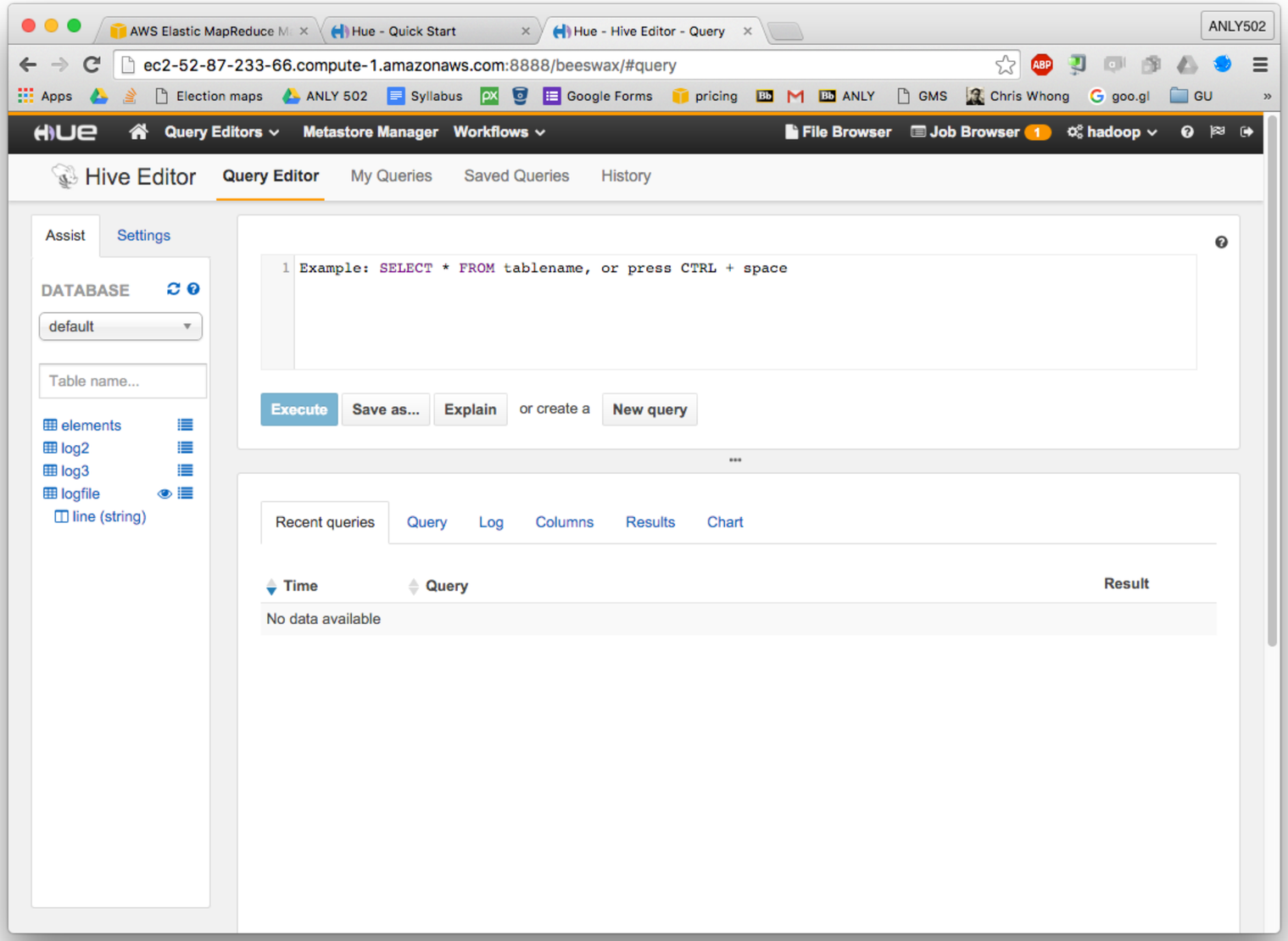
You can run this query from Hue

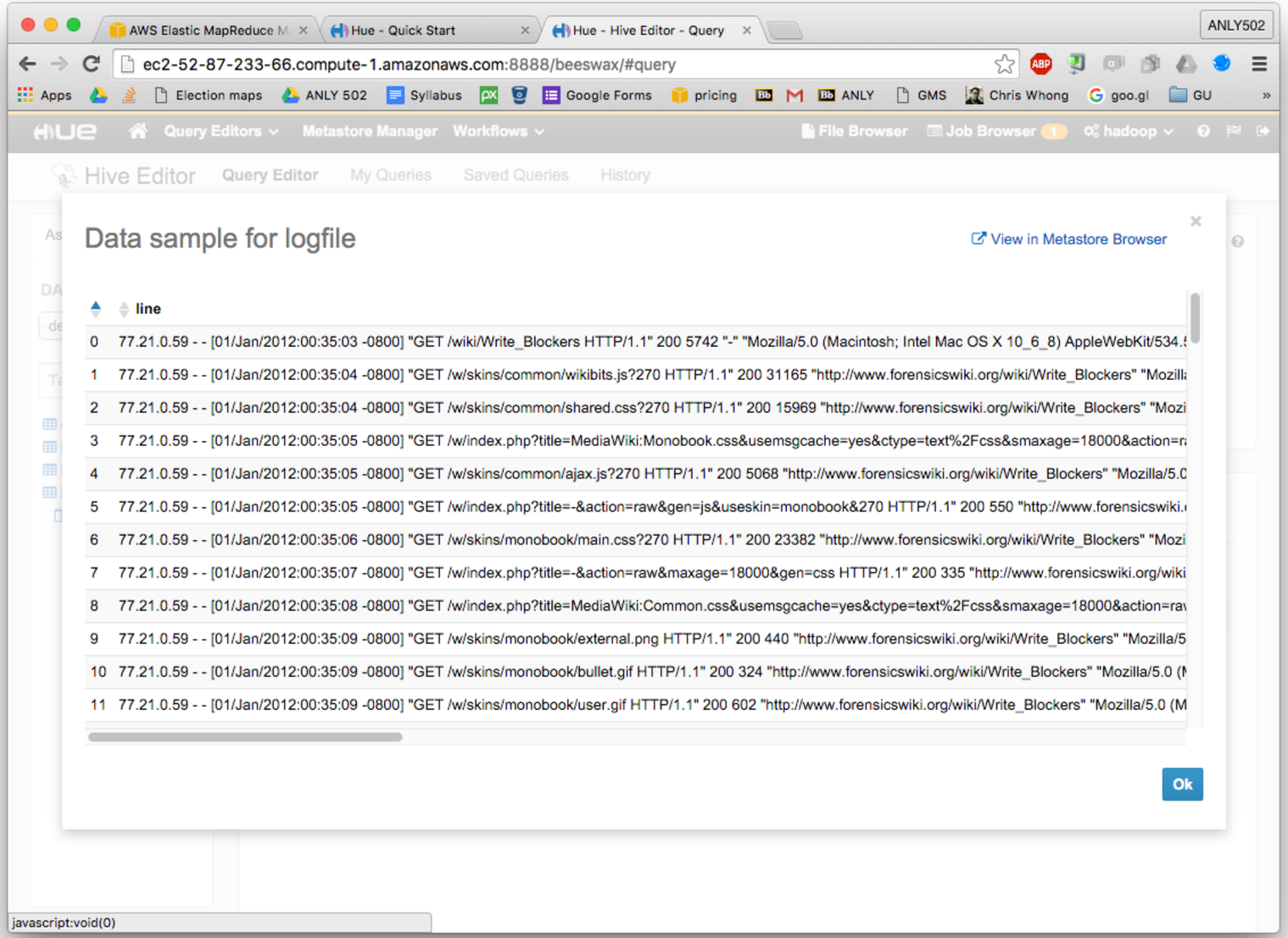



AWS Elastic MapReduce
Hue - Hive Editor - Query
Hue - Quick Start
Hue - Welcome Home
History Server
All Applications
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/beeswax/execute/query/1#query/results
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
AWS
Doc
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
Hive Editor
Query Editor
My Queries
Saved Queries
History
Assist
Settings
DATABASE
default
Table name...
log2012
1 select * from log2012 limit 5;
Execute
Save
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
log2012.line
0 77.21.0.59 - - [01/Jan/2012:00:35:03 -0800] "GET /wiki/Write_Blockers HTTP/1.1" 200 5742 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/534.52.7 (KH
1 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/wikibits.js?270 HTTP/1.1" 200 31165 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (Ma
2 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/shared.css?270 HTTP/1.1" 200 15969 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (M
3 77.21.0.59 - - [01/Jan/2012:00:35:05 -0800] "GET /w/index.php?title=MediaWiki:Monobook.css&usemsgcache=yes&ctype=text%2Fcss&smaxage=18000&action=raw&maxa
4 77.21.0.59 - - [01/Jan/2012:00:35:05 -0800] "GET /w/skins/common/ajax.js?270 HTTP/1.1" 200 5068 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (Macinto
ec2-52-87-233-66.compute-1.amazonaws.com:8888/beeswax/execute/query/1#results

You can also browse the tables directly




AWS Elastic MapReduce M
Hue - Quick Start
Hue - Hive Editor - Query
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/beeswax/#query
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
HUE
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
Hive Editor
Query Editor
My Queries
Saved Queries
History
Assist
Settings
DATABASE
default
Table name...
elements
log2
log3
logfile
line (string)
1 Example: SELECT * FROM tablename, or press CTRL + space
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
Time
Query
Result
No data available


AWS Elastic MapReduce
Hue - Quick Start
Hue - Hive Editor - Query
ANLY502
ec2-52-87-233-66.compute-1.amazonaws.com:8888/beeswax/#query
Apps
Election maps
ANLY 502
Syllabus
Google Forms
pricing
ANLY
GMS
Chris Whong
goo.gl
GU
Query Editors
Metastore Manager
Workflows
File Browser
Job Browser
hadoop
Hive Editor
Query Editor
My Queries
Saved Queries
History
Data sample for logfile
View in Metastore Browser
line
0 77.21.0.59 - - [01/Jan/2012:00:35:03 -0800] "GET /wiki/Write_Blockers HTTP/1.1" 200 5742 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/534.
1 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/wikibits.js?270 HTTP/1.1" 200 31165 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozill
2 77.21.0.59 - - [01/Jan/2012:00:35:04 -0800] "GET /w/skins/common/shared.css?270 HTTP/1.1" 200 15969 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozi
3 77.21.0.59 - - [01/Jan/2012:00:35:05 -0800] "GET /w/index.php?title=MediaWiki:Monobook.css&usemsgcache=yes&ctype=text%2Fcss&smaxage=18000&action=r
4 77.21.0.59 - - [01/Jan/2012:00:35:05 -0800] "GET /w/skins/common/ajax.js?270 HTTP/1.1" 200 5068 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0
5 77.21.0.59 - - [01/Jan/2012:00:35:05 -0800] "GET /w/index.php?title=-&action=raw&gen=js&useskin=monobook&270 HTTP/1.1" 200 550 "http://www.forensicswiki.
6 77.21.0.59 - - [01/Jan/2012:00:35:06 -0800] "GET /w/skins/monobook/main.css?270 HTTP/1.1" 200 23382 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozi
7 77.21.0.59 - - [01/Jan/2012:00:35:07 -0800] "GET /w/index.php?title=-&action=raw&maxage=18000&gen=css HTTP/1.1" 200 335 "http://www.forensicswiki.org/wiki
8 77.21.0.59 - - [01/Jan/2012:00:35:08 -0800] "GET /w/index.php?title=MediaWiki:Common.css&usemsgcache=yes&ctype=text%2Fcss&smaxage=18000&action=ra
9 77.21.0.59 - - [01/Jan/2012:00:35:09 -0800] "GET /w/skins/monobook/external.png HTTP/1.1" 200 440 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5
10 77.21.0.59 - - [01/Jan/2012:00:35:09 -0800] "GET /w/skins/monobook/bullet.gif HTTP/1.1" 200 324 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (
11 77.21.0.59 - - [01/Jan/2012:00:35:09 -0800] "GET /w/skins/monobook/user.gif HTTP/1.1" 200 602 "http://www.forensicswiki.org/wiki/Write_Blockers" "Mozilla/5.0 (M
Ok
javascript:void(0)

Things to note:

Hive will load:

- HDFS "directories" e.g. hdfs:///user/hadoop/forensicswiki/"
- HDFS files e.g. hdfs:///user/hadoop/forensicswiki.2012.txt
- S3 "directories" e.g. s3://gu-anly502/ps03/forensicswiki/

But Hive will *not* load a single file for S3. (Appears to be a bug)

Hive will load directories *recursively* if you set two variables:

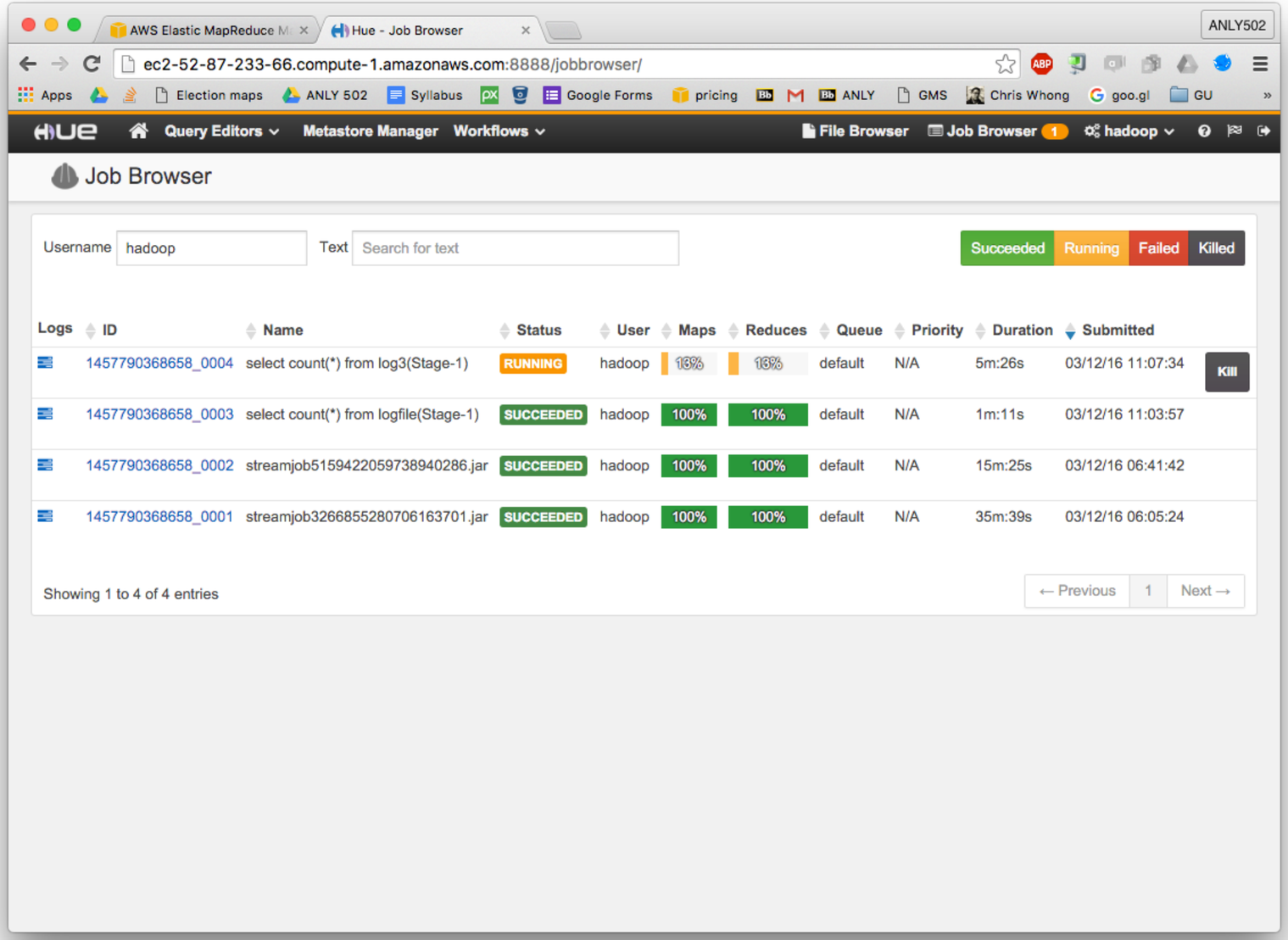
```
0: jdbc:hive2://localhost:10000> SET mapred.input.dir.recursive=true;
0: jdbc:hive2://localhost:10000> SET hive.mapred.supports.subdirectories=true
```

Let's use Hive to count the number of lines in s3://gu-anly502/ps03/forensicswiki/ (the 365-file version)

```
$ beeline
beeline> !connect jdbc:hive2://localhost:10000
...
0: jdbc:hive2://localhost:10000> create external table 2012 (line string) location 's3://gu-anly502/ps03/forensicswiki/';
No rows affected (0.721 seconds)
0: jdbc:hive2://localhost:10000> select count(*) from log2012;
...
```

```
0: jdbc:hive2://localhost:10000> select count(*) from log3;
INFO  : Number of reduce tasks determined at compile time: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:365
INFO  : Submitting tokens for job: job_1457790368658_0004
INFO  : The url to track the job: http://ip-172-31-44-166.ec2.internal:20888/proxy/application_1457790368658_0004/
INFO  : Starting Job = job_1457790368658_0004, Tracking URL = http://ip-172-31-44-166.ec2.internal:20888/proxy/application_1457790368658_0004/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1457790368658_0004
INFO  : Hadoop job information for Stage-1: number of mappers: 365; number of reducers: 1
INFO  : 2016-03-12 19:07:40,540 Stage-1 map = 0%,  reduce = 0%
INFO  : 2016-03-12 19:08:05,631 Stage-1 map = 1%,  reduce = 0%, Cumulative CPU 13.79 sec
INFO  : 2016-03-12 19:08:12,479 Stage-1 map = 2%,  reduce = 0%, Cumulative CPU 41.36 sec
INFO  : 2016-03-12 19:08:36,114 Stage-1 map = 3%,  reduce = 0%, Cumulative CPU 76.79 sec
INFO  : 2016-03-12 19:08:52,007 Stage-1 map = 4%,  reduce = 0%, Cumulative CPU 90.08 sec
INFO  : 2016-03-12 19:09:02,345 Stage-1 map = 5%,  reduce = 0%, Cumulative CPU 118.71 sec
INFO  : 2016-03-12 19:09:31,063 Stage-1 map = 6%,  reduce = 0%, Cumulative CPU 154.39 sec
INFO  : 2016-03-12 19:09:34,425 Stage-1 map = 7%,  reduce = 0%, Cumulative CPU 169.24 sec
INFO  : 2016-03-12 19:09:59,468 Stage-1 map = 8%,  reduce = 0%, Cumulative CPU 197.72 sec
INFO  : 2016-03-12 19:10:21,173 Stage-1 map = 9%,  reduce = 0%, Cumulative CPU 225.28 sec
INFO  : 2016-03-12 19:10:28,107 Stage-1 map = 10%,  reduce = 0%, Cumulative CPU 246.17 sec
INFO  : 2016-03-12 19:10:50,493 Stage-1 map = 11%,  reduce = 0%, Cumulative CPU 274.83 sec
INFO  : 2016-03-12 19:10:57,153 Stage-1 map = 12%,  reduce = 0%, Cumulative CPU 297.25 sec
INFO  : 2016-03-12 19:11:18,439 Stage-1 map = 13%,  reduce = 0%, Cumulative CPU 326.12 sec
INFO  : 2016-03-12 19:11:43,724 Stage-1 map = 14%,  reduce = 0%, Cumulative CPU 354.71 sec
INFO  : 2016-03-12 19:11:49,329 Stage-1 map = 15%,  reduce = 0%, Cumulative CPU 377.45 sec
INFO  : 2016-03-12 19:12:14,303 Stage-1 map = 16%,  reduce = 0%, Cumulative CPU 407.04 sec
INFO  : 2016-03-12 19:12:28,107 Stage-1 map = 17%,  reduce = 0%, Cumulative CPU 439.22 sec
INFO  : 2016-03-12 19:12:48,660 Stage-1 map = 18%,  reduce = 0%, Cumulative CPU 473.09 sec
INFO  : 2016-03-12 19:13:12,521 Stage-1 map = 19%,  reduce = 0%, Cumulative CPU 502.13 sec
```

View running jobs…



Why is this taking so long?

```
INFO  : number of splits:365
```

365 splits?

```
INFO  : 2016-03-12 19:19:44,879 Stage-1 map = 43%,  reduce = 14%, Cumulative CPU 1168.69 sec
INFO  : 2016-03-12 19:20:03,511 Stage-1 map = 44%,  reduce = 14%, Cumulative CPU 1191.84 sec
INFO  : 2016-03-12 19:20:04,577 Stage-1 map = 44%,  reduce = 15%, Cumulative CPU 1200.04 sec
INFO  : 2016-03-12 19:20:21,996 Stage-1 map = 45%,  reduce = 15%, Cumulative CPU 1220.86 sec
INFO  : 2016-03-12 19:20:40,499 Stage-1 map = 46%,  reduce = 15%, Cumulative CPU 1254.78 sec
INFO  : 2016-03-12 19:20:49,226 Stage-1 map = 47%,  reduce = 15%, Cumulative CPU 1268.63 sec
INFO  : 2016-03-12 19:20:50,314 Stage-1 map = 47%,  reduce = 16%, Cumulative CPU 1268.67 sec
INFO  : 2016-03-12 19:21:08,951 Stage-1 map = 48%,  reduce = 16%, Cumulative CPU 1300.92 sec
INFO  : 2016-03-12 19:21:26,297 Stage-1 map = 49%,  reduce = 16%, Cumulative CPU 1331.44 sec
INFO  : 2016-03-12 19:21:40,455 Stage-1 map = 50%,  reduce = 16%, Cumulative CPU 1351.64 sec
INFO  : 2016-03-12 19:21:42,669 Stage-1 map = 50%,  reduce = 17%, Cumulative CPU 1351.67 sec
INFO  : 2016-03-12 19:21:57,839 Stage-1 map = 51%,  reduce = 17%, Cumulative CPU 1379.21 sec
INFO  : 2016-03-12 19:22:14,363 Stage-1 map = 52%,  reduce = 17%, Cumulative CPU 1400.65 sec
INFO  : 2016-03-12 19:22:31,819 Stage-1 map = 53%,  reduce = 17%, Cumulative CPU 1430.1 sec
INFO  : 2016-03-12 19:22:33,963 Stage-1 map = 53%,  reduce = 18%, Cumulative CPU 1436.9 sec
INFO  : 2016-03-12 19:22:50,276 Stage-1 map = 54%,  reduce = 18%, Cumulative CPU 1458.65 sec
INFO  : 2016-03-12 19:22:57,997 Stage-1 map = 55%,  reduce = 18%, Cumulative CPU 1478.61 sec
INFO  : 2016-03-12 19:23:15,373 Stage-1 map = 56%,  reduce = 18%, Cumulative CPU 1504.5 sec
```

Lots of mapping and reducing?

Why is this taking so long? "top"

```
top - 19:26:25 up  5:43,  3 users,  load average: 9.84, 9.63, 7.32
Tasks: 172 total,   1 running, 169 sleeping,   2 stopped,   0 zombie
Cpu(s): 88.2%us, 11.4%sy,  0.0%ni,  0.4%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  15407244k total, 14758868k used,   648376k free,    70120k buffers
Swap:        0k total,        0k used,        0k free,  6949444k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
17588 yarn      20   0 1826m 151m  27m S 111.7  1.0   0:04.53 java
17465 yarn      20   0 1849m 236m  27m S 96.1  1.6   0:11.56 java
17423 yarn      20   0 1843m 274m  27m S 75.9  1.8   0:12.75 java
17350 yarn      20   0 1853m 493m  27m S 71.9  3.3   0:15.66 java
 4477 yarn      20   0 3422m 466m  28m S  8.0  3.1   3:02.65 java
 5363 hadoop    20   0 6333m 441m  14m S  6.0  2.9   2:28.49 java
 2315 hadoop    20   0 3496m 171m  14m S  4.3  1.1   1:57.17 java
 4906 yarn      20   0 2918m 256m  28m S  4.3  1.7   5:30.34 java
 8980 hive      20   0 1653m 271m  28m S  4.0  1.8   1:20.78 java
10221 oozie     20   0 3511m 438m  17m S  3.3  2.9   8:33.79 java
12759 hue       20   0 1088m 117m  12m S  1.3  0.8   0:29.26 python2.6
24364 yarn      20   0 3219m 604m  27m S  1.3  4.0   0:47.14 java
 3810 hdfs      20   0 2522m 638m  27m S  0.7  4.2   1:49.90 java
 3938 hdfs      20   0 1507m 275m  27m S  0.7  1.8   1:33.83 java
 5711 yarn      20   0 3054m 242m  27m S  0.7  1.6   0:15.50 java
    3 root      20   0     0    0    0 S  0.3  0.0   0:00.95 ksoftirqd/0
 1580 root      20   0  110m 3084 2872 S  0.3  0.0   0:00.08 dump-instance-s
 3446 kms       20   0 5736m 164m  14m S  0.3  1.1   0:25.42 java
12730 root      20   0  357m  15m 4056 S  0.3  0.1   0:05.88 python2.6
17492 hadoop    20   0 15292 2392 2016 R  0.3  0.0   0:00.03 top
17579 yarn      20   0  110m 2820 2668 S  0.3  0.0   0:00.01 bash
17646 hadoop    20   0 15288 2192 1896 S  0.3  0.0   0:00.01 top
20964 hadoop    20   0 1611m 117m  27m S  0.3  0.8   0:02.83 java
    1 root      20   0 19776 2744 2264 S  0.0  0.0   0:01.44 init
```

Load of 9?

4 cores processing
15 million lines?

We eventually get the result…

```
INFO  : 2016-03-12 19:35:43,436 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2672.63 sec
INFO  : MapReduce Total cumulative CPU time: 44 minutes 32 seconds 630 msec
INFO  : Ended Job = job_1457790368658_0004
+-----------+--+
|    _c0    |
+-----------+--+
| 15949554  |
+-----------+--+
1 row selected (1700.456 seconds)
0: jdbc:hive2://localhost:10000>
```

There are many ways to load data into a table

CREATE EXTERNAL TABLE name LOCATION ;

- *optional: FIELDS TERMINATED BY ','*
- *optional: LINES TERMINATED BY '\n'*

CREATE EXTERNAL TABLE *table...*;
LOAD DATA INPATH '*hdfs_path*' INTO TABLE *table*;

CREATE EXTERNAL TABLE ...;
ALTER TABLE *name* SET LOCATION '*hdfs_or_s3_path_of_directory*'

CREATE [TEMPORARY|EXTERNAL] TABLE *name*
INSERT INTO TABLE *name* SELECT *expression1,expression2,..* FROM *othertable*

CREATE TABLE...
INSERT ... VALUES ...

Create Table DDL and Alter Table DDL

```
CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name    -- (Note: TEMPORARY available in Hive 0.14.0 and later)
  [(col_name data_type [COMMENT col_comment], ...)]
  [COMMENT table_comment]
  [PARTITIONED BY (col_name data_type [COMMENT col_comment], ...)]
  [CLUSTERED BY (col_name, col_name, ...) [SORTED BY (col_name [ASC|DESC], ...)] INTO num_buckets BUCKETS]
  [SKEWED BY (col_name, col_name, ...)                  -- (Note: Available in Hive 0.10.0 and later)]
     ON ((col_value, col_value, ...), (col_value, col_value, ...), ...)
     [STORED AS DIRECTORIES]
  [
   [ROW FORMAT row_format]
   [STORED AS file_format]
     | STORED BY 'storage.handler.class.name' [WITH SERDEPROPERTIES (...)]  -- (Note: Available in Hive 0.6.0 and later)
  ]
  [LOCATION hdfs_path]
  [TBLPROPERTIES (property_name=property_value, ...)]   -- (Note: Available in Hive 0.6.0 and later)
  [AS select_statement];   -- (Note: Available in Hive 0.5.0 and later; not supported for external tables)
```

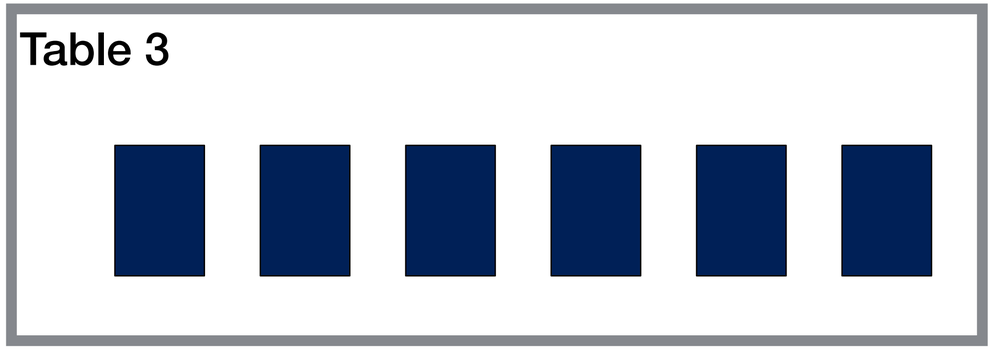
—*https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL*

Partitions split a single table into different sections

For example, we can partition forensicswiki logfiles by YEAR and MONTH.

e.g.:

```
$ aws s3 ls s3://gu-anly502/ps05/
                           PRE forensicswiki/
$ aws s3 ls s3://gu-anly502/ps05/forensicswiki/
                           PRE 2012/
$ aws s3 ls s3://gu-anly502/ps05/forensicswiki/2012/
                           PRE 01/
                           PRE 02/
                           PRE 03/
                           PRE 04/
                           PRE 05/
                           PRE 06/
                           PRE 07/
                           PRE 08/
                           PRE 09/
                           PRE 10/
                           PRE 11/
                           PRE 12/
$ aws s3 ls s3://gu-anly502/ps05/forensicswiki/2012/01/
2016-03-12 14:23:30  402236086 access.log.2012-01
$
```



When a query is issued Hive will only consider the necessary partitions.

Creating partitions:
Create the table; create the partitions; add data to the partitions.

```
0: jdbc:hive2://localhost:10000> create table plogs (line string) partitioned by (year int, month int);
No rows affected (0.032 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs add partition (year=2012,month=1);
No rows affected (0.082 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs partition (year=2012,month=1) set location 's3://gu-anly502/ps05/forensicswiki/
2012/01/';
No rows affected (0.386 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs add partition (year=2012,month=2);
No rows affected (0.052 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs partition (year=2012,month=2) set location 's3://gu-anly502/ps05/forensicswiki/
2012/02/';
No rows affected (0.36 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs add partition (year=2012,month=3);
No rows affected (0.057 seconds)
0: jdbc:hive2://localhost:10000> alter table plogs partition (year=2012,month=3) set location 's3://gu-anly502/ps05/forensicswiki/
2012/03/';
No rows affected (0.356 seconds)
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> select year,month,substr(line,0,30) from plogs limit 3;
+-------+--------+----------------------------------+--+
| year  | month  |               _c2                |
+-------+--------+----------------------------------+--+
| 2012  | 1      | 77.21.0.59 - - [01/Jan/2012:00   |
| 2012  | 1      | 77.21.0.59 - - [01/Jan/2012:00   |
| 2012  | 1      | 77.21.0.59 - - [01/Jan/2012:00   |
+-------+--------+----------------------------------+--+
3 rows selected (0.332 seconds)
0: jdbc:hive2://localhost:10000>
```

Parsing Apache logs in Hive
1. Create a table of extracted fields; 2. Reparse the fields as necessary

```
DROP TABLE IF EXISTS apache_common_log;
CREATE EXTERNAL TABLE apache_common_log (
  host STRING,
  identity STRING,
  user STRING,
  rawdatetime STRING,
  request STRING,
  status STRING,
  size STRING,
  refer STRING,
  agent STRING
  )
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  "input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*) \"([^\"]*)\" \"([^\"]*)\".*",
  "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s"
)
STORED AS TEXTFILE
LOCATION 's3://gu-anly502/ps05/forensicswiki/2012/12/';

create temporary table clean_logs (
  host string,
  datetime string,
  url string
);

insert overwrite table clean_logs
  select host,
         from_unixtime(unix_timestamp(rawdatetime, "[dd/MMM/yyyy:HH:mm:ss Z]")),
         regexp_extract(request,".[^ ]* ([^ ]*)")

  from apache_common_log;


SELECT * FROM clean_logs limit 3;
```

Uses Hive's Regular Expression Serializer/Deserializer

Creates a temporary table clean_logs with the desired fields

reformats the time, extracts the URL from the request.

Running the sample code:

```
$ beeline -u jdbc:hive2://localhost:10000 -n hadoop -p hadoop -f apache-demo.sql
scan complete in 8ms
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 1.0.0-amzn-2)
Driver: Hive JDBC (version 1.0.0-amzn-2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
0: jdbc:hive2://localhost:10000> DROP TABLE IF EXISTS apache_common_log;
No rows affected (0.202 seconds)
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE apache_common_log (
. . . . . .>   host STRING,
. . . . . .>   identity STRING,
. . . . . .>   user STRING,
...
. . . . . .>   from apache_common_log;
INFO  : Number of reduce tasks is set to 0 since there's no reduce operator
INFO  : number of splits:2
INFO  : Submitting tokens for job: job_1457790368658_0014
INFO  : The url to track the job: http://ip-172-31-44-166.ec2.internal:20888/proxy/application_1457790368658_0014/
INFO  : Starting Job = job_1457790368658_0014, Tracking URL = http://ip-172-31-44-166.ec2.internal:20888/proxy/application_1457790368658_0014/
INFO  : Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1457790368658_0014
INFO  : Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 0
INFO  : 2016-03-13 04:12:22,934 Stage-1 map = 0%,  reduce = 0%
...
INFO  : Stage-5 is filtered out by condition resolver.
INFO  : Moving data to: hdfs://ip-172-31-44-166.ec2.internal:8020/tmp/hive/hadoop/5216aa41-8209-48fc-a9ab-f02c2329007c/hive_2016-03-13_04-12-15_999_5818612435067683317-3/-
ext-10000 from hdfs://ip-172-31-44-166.ec2.internal:8020/tmp/hive/hadoop/5216aa41-8209-48fc-a9ab-f02c2329007c/hive_2016-03-13_04-12-15_999_5818612435067683317-3/-ext-10002
INFO  : Loading data to table default.clean_logs from hdfs://ip-172-31-44-166.ec2.internal:8020/tmp/hive/hadoop/5216aa41-8209-48fc-a9ab-f02c2329007c/
hive_2016-03-13_04-12-15_999_5818612435067683317-3/-ext-10000
INFO  : Table default.clean_logs stats: [numFiles=2, numRows=1397115, totalSize=149134460, rawDataSize=147737345]
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> -- SELECT * FROM apache_common_log LIMIT 5;
0: jdbc:hive2://localhost:10000> SELECT * FROM clean_logs limit 3;
+------------------+----------------------+----------------------------------+--+
| clean_logs.host  | clean_logs.datetime  |          clean_logs.url          |
+------------------+----------------------+----------------------------------+--+
| 72.47.85.98      | 2012-12-01 08:19:03  | /                                |
| 69.163.129.236   | 2012-12-01 08:19:04  | /w/extensions/BibTex/bibtex.css  |
| 69.163.129.236   | 2012-12-01 08:19:04  | /w/extensions/BibTex/bibtex.js   |
+------------------+----------------------+----------------------------------+--+
3 rows selected (0.09 seconds)
0: jdbc:hive2://localhost:10000>
```

Other goodies in Hive

Math Functions

Array functions

Map & Reduce functions

Date time functions

XML (xpath) & JSON

String Functions

Statistics, Aggragate

expode() — Takes an array and maps to separate rows

Parsing

MD5, SHA1, CRC, Encryption

—*https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF*

This is easy!

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable
RDBMS	Yes ☹	✗	✔	✔	✔	✔	✗

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable
RDBMS	Yes ☹	✗	✔	✔	✔	✔	✗
MapReduce	No ☺	✗	✗	✗	✗	✗	✔

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable
RDBMS	Yes ☹	✗	✔	✔	✔	✔	✗
MapReduce	No ☺	✗	✗	✗	✗	✗	✔
Pig (procedural)	No ☺	✔	✗	✔	✔	✔	✔

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable
RDBMS	Yes ☹	✗	✔	✔	✔	✔	✗
MapReduce	No ☺	✗	✗	✗	✗	✗	✔
Pig (procedural)	No ☺	✔	✗	✔	✔	✔	✔
Hive (SQL in HDFS)	No ☺	✗	✔	✔	✔	✔	✔

Massive Data meets Analytics

	ETL	procedural	SQL	Interactive	High level	Optimized	Scalable
RDBMS	Yes ☹	✗	✔	✔	✔	✔	✗
MapReduce	No ☺	✗	✗	✗	✗	✗	✔
Pig (procedural)	No ☺	✔	✗	✔	✔	✔	✔
Hive (SQL in HDFS)	No ☺	✗	✔	✔	✔	✔	✔
SparkSQL (procedural & SQL)	No ☺	✔	✔	✔	✔	✔	✔

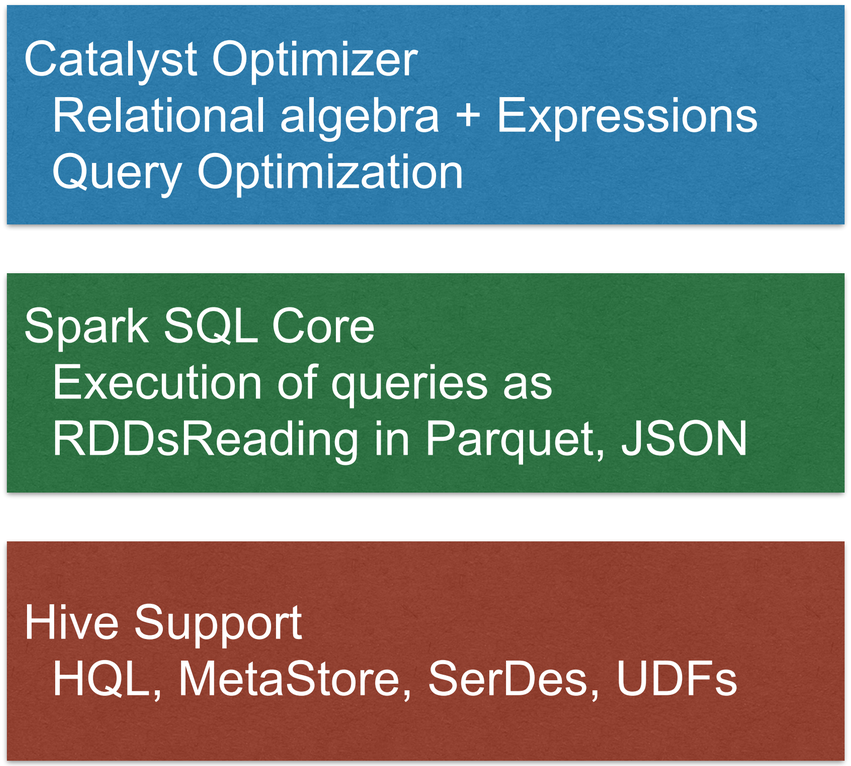
Spark SQL —

Spark Summit 2013 — December 2013

- 1 Developer
- Able to run simple queries on data stored in Hive

Spark Summit 2014

- 44 contributors
- Supports data stored in Hive, Parquet, JSON
- Bindings for Scala, Java & Python



- https://spark-summit.org/2014/wp-content/uploads/2014/07/Performing-Advanced-Analytics-on-Relational-Data-with-Spark-SQL-Michael-Armbrust.pdf

Spark SQL is an SQL-interface to Spark.

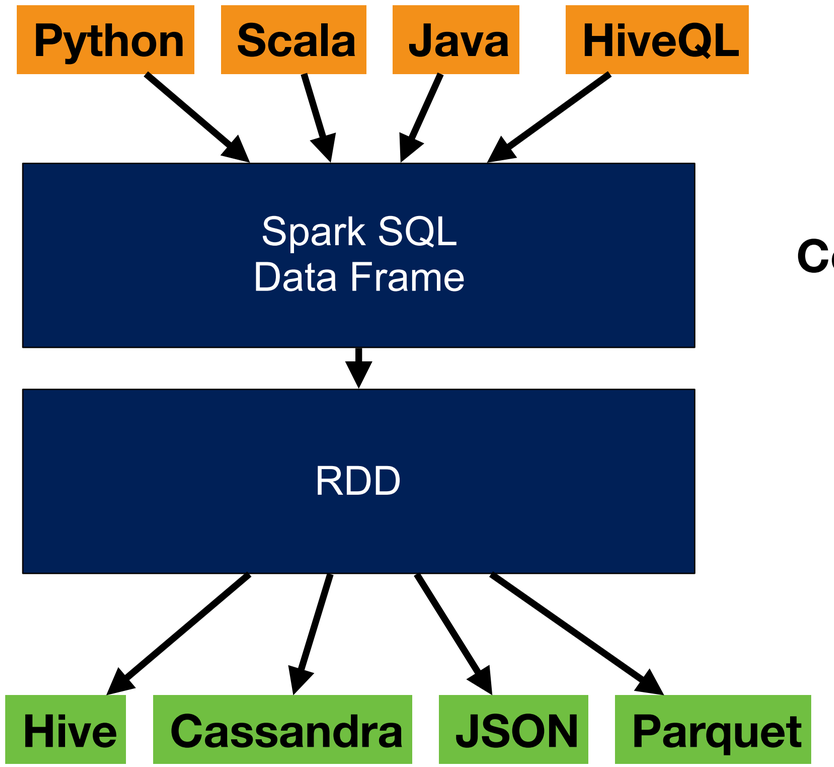
Create a data source (RDD, List of JSON objects, ...)

Bind the data to a Data Frame

- CreateTemporaryTable

Access via SQL

- SELECT * statements return Data Frames



Collection of Row() objects

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.7.10 (default, Dec  8 2015 18:25:23)
SparkContext available as sc, HiveContext available as sqlContext.
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 1.6.0
      /_/

Using Python version 2.7.10 (default, Dec  8 2015 18:25:23)
SparkContext available as sc, HiveContext available as sqlContext.
```

Creating an RDD with a schema from rows of Python dictionaries (e.g. JSON)

presents.csv:

```
num,name,url,begin,end,party,state
1,George Washington,http://en.wikipedia.org/wiki/George_Washington,30/04/1789,4/03/1797,Independent,Virginia
2,John Adams,http://en.wikipedia.org/wiki/John_Adams,4/03/1797,4/03/1801,Federalist,Massachusetts
3,Thomas Jefferson,http://en.wikipedia.org/wiki/Thomas_Jefferson,4/03/1801,4/03/1809,Democratic-Republican,Virginia
4,James Madison,http://en.wikipedia.org/wiki/James_Madison,4/03/1809,4/03/1817,Democratic-Republican,Virginia
...
```

Create an array of President "dicts":

```
import csv
if __name__=="__main__":
    with open("presidents.csv","r") as csvfile:
        reader = csv.DictReader(csvfile)
        rows = [row for row in reader]
```

— *Each row:*

```
{'begin': '30/04/1789','end': '4/03/1797','name': 'George Washington','num': '1','party': 'Independent',
  'state': 'Virginia','url': 'http://en.wikipedia.org/wiki/George_Washington'}
```

— *rows[]* = [row, row, row]

Turn it into an DataFrame / SQL Table:

```
presidentsTable = sqlCtx.createDataFrame(rows)
presidentsTable.registerTempTable("pres")
print(sqlCtx.sql("select name,party from pres").collect())
```

Convert the array of dictionaries into a Data Frame:

sqlCtx is the SparkSQL Context.

— *Use sqlCtx.createDataFrame():*

```
presidentsTable = sqlCtx.createDataFrame(rows)
presidentsTable.registerTempTable("pres")
```

Then issue SQL:

```
rdd = sqlCtx.sql("select name,party from pres")
```

This works for a Data Frame of 45 rows, or 45 million rows.

Convert the array of dictionaries into a Data Frame:

sqlCtx is the SparkSQL Context.

— *Use sqlCtx.createDataFrame():*

```
presidentsTable = sqlCtx.createDataFrame(rows)
presidentsTable.registerTempTable("pres")
```

Registers the table

Then issue SQL:

```
rdd = sqlCtx.sql("select name,party from pres")
```

This works for a Data Frame of 45 rows, or 45 million rows.

Convert the array of dictionaries into a Data Frame:

sqlCtx is the SparkSQL Context.

— *Use sqlCtx.createDataFrame():*

```
presidentsTable = sqlCtx.createDataFrame(rows)
presidentsTable.registerTempTable("pres")
```

Registers the table

Then issue SQL:

Uses registered table

```
rdd = sqlCtx.sql("select name,party from pres")
```

This works for a Data Frame of 45 rows, or 45 million rows.

Convert the array of dictionaries into a Data Frame:

sqlCtx is the SparkSQL Context.

— *Use sqlCtx.createDataFrame():*

```
presidentsTable = sqlCtx.createDataFrame(rows)
presidentsTable.registerTempTable("pres")
```

Registers the table

Then issue SQL:

Uses registered table

```
rdd = sqlCtx.sql("select name,party from pres")
```

This works for a Data Frame of 45 rows, or 45 million rows.

Hive tables are pre-registered

You can do a lot with a data frame!

Arbitrary SQL:

```
rdd = sqlCtx.sql("select name,party from pres where name like 'George%' ")
```

User defined functions! (in any language)

```
>>> from pyspark.sql.types import IntegerType
>>> sqlContext.udf.register("stringLengthInt", lambda x: len(x), IntegerType())
>>> sqlContext.sql("SELECT stringLengthInt('test')").collect()
```

Great functions:

```
df.show()                 – Show the contents of the Data Frame
df['name']                – Refers to the column 'name'
df.filter()               – Filter an DataFrame to product another one
```

Works Great!

JDBC:

- JDBC client to connect to other databases
- JDBC server for clients.

Read the docs!

- http://spark.apache.org/docs/latest/api/python/pyspark.sql.html

https://www.pexels.com/photo/people-apple-iphone-writing-154/

Homework and L08 Preview

Reminder — Upcoming

March 18 — PS04 due

- Validation script works! (I hope!)

March 21 — Midterm

March 22 — Final Project Proposals (2)

- Post to forum

April 1 — Final Project Group Proposal

- Each group member must submit the same proposal on Blackboard!
- Blackboard groups will be created.

Reading!

Read this Apache Hive Documentation:

- https://cwiki.apache.org/confluence/display/Hive/Tutorial

Skim the API

- https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL
- https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients

The Hive documentation is on a wiki.

Recommended blog posts:

- https://blog.cloudera.com/blog/2014/02/migrating-from-hive-cli-to-beeline-a-primer/
- https://www.brentozar.com/archive/2013/03/introduction-to-hive-partitioning/
- http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/
- https://www.qubole.com/blog/big-data/5-tips-for-efficient-hive-queries/