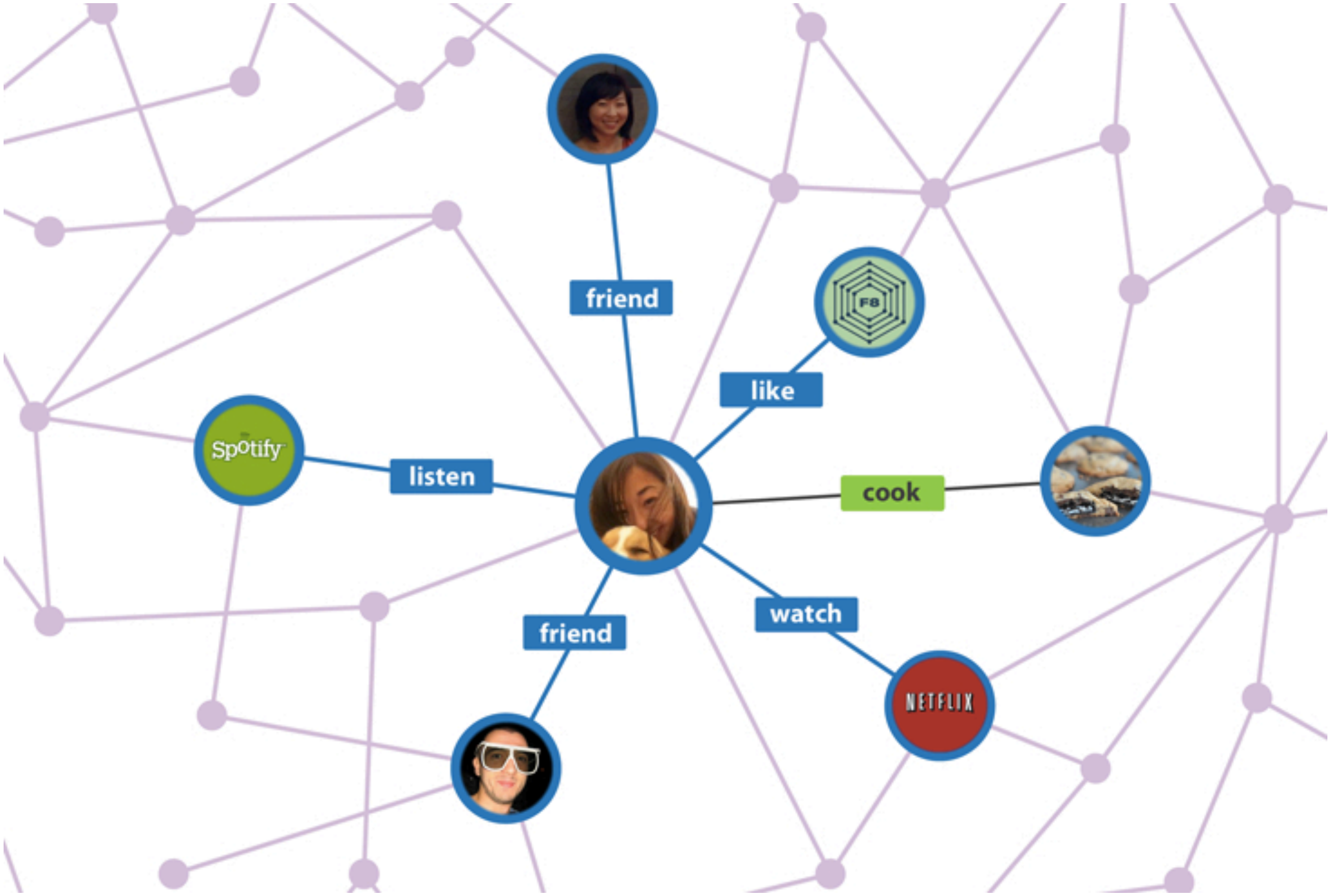
CSE 344 Introduction to Data Management

Section 9: AWS, Hadoop, Pig Latin

Srini (sviyer@cs)

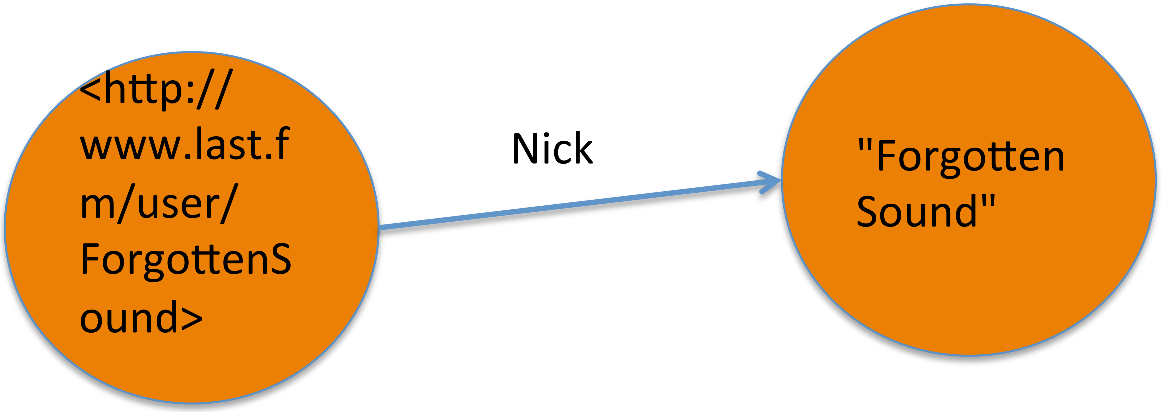
Homework 8 (Last hw ☺)

Huge Graphs out there!



Billion Triple Set:

contains web information, obtained by a crawler



subject predicate object [context]

<http://www.last.fm/user/ForgottenSound>

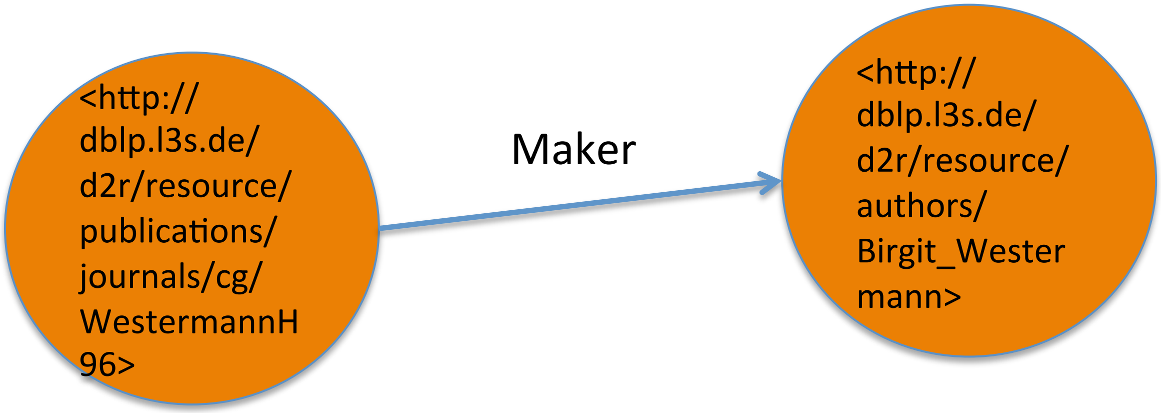
<http://xmlns.com/foaf/0.1/nick>

"ForgottenSound"

<http://rdf.opiumfield.com/lastfm/friends/life-exe> .

Billion Triple Set:

contains web information, obtained by a crawler



<http://dblp.l3s.de/d2r/resource/publications/journals/cg/WestermannH96>

<http://xmlns.com/foaf/0.1/maker>

<http://dblp.l3s.de/d2r/resource/authors/Birgit_Westermann>

<http://dblp.l3s.de/d2r/data/publications/journals/cg/WestermannH96> .

Homework 8 (Last hw ☺)

- 0.5 TB (yes, TeraBytes!) of data
- 251 files of ~ 2GB each

  btc-2010-chunk-000 to btc-2010-chunk-317
- You will write pig queries for each task and use MapReduce to perform data analysis.
- Due ~ 2 weeks from now

- Problem 1:

  select object, count(object) as cnt group by obj order by cnt desc;

- Problem 2 (on 2GB):
  - 1) subject, count(subject) as cnt group by subject

    ```
    spotify.com     50
    last.fm          50
    ```

  - 2) cnt, count(cnt) as cnt1 group by cnt1;

    ```
    50        2
    ```

  - 3) Plot using excel/gnuplot

- Problem 3:

  all (subject, predicate, object, subject2, predicate2, object2)
  where subject contains “rdfabout.com” / others…

- Problem 4 (on 0.5 TB):

  Run Problem 2 on all of the data (use upto 19 machines. Takes ~4 hours)

Amazon web services (AWS)

EC2 (Elastic Computing Cluster): virtual servers in the cloud

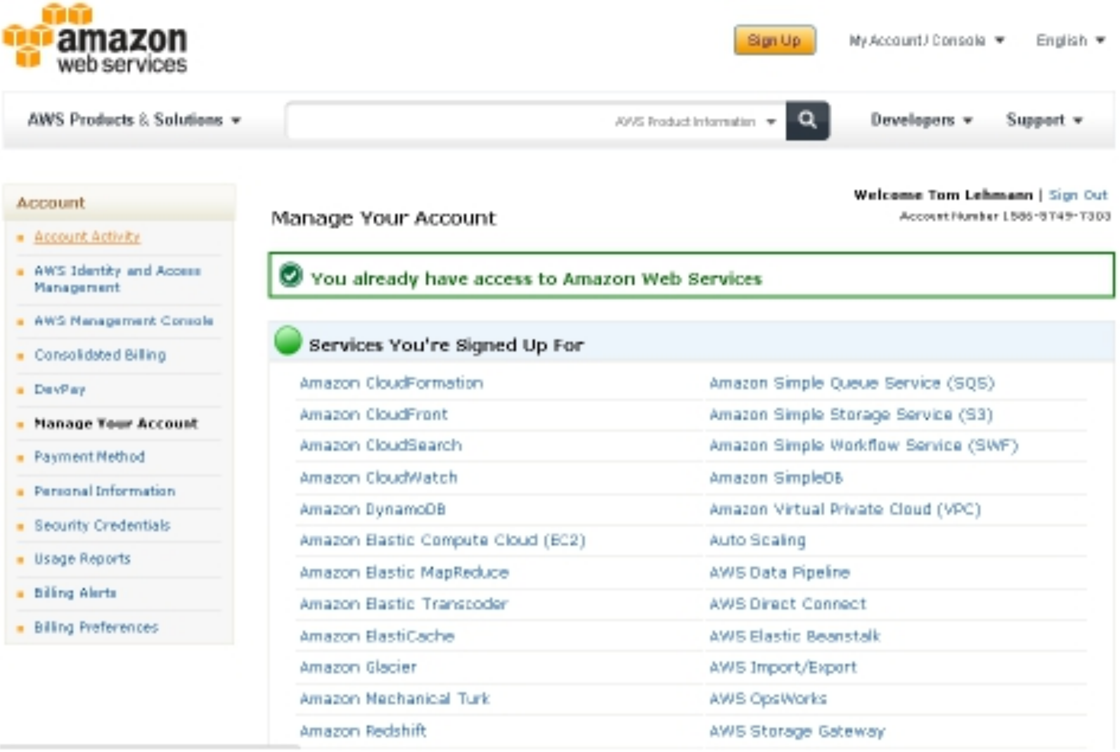
amazon

S3 (Simple Storage Service): scalable storage in the cloud

Elastic MapReduce: Managed Hadoop Framework

1. Setting up AWS account

- Sign up/in: https://aws.amazon.com/
- Make sure you are signed up for (1) Elastic MapReduce (2) EC2 (3) S3



Welcome Tom Lehmann | Sign Out
Account Number 1586-8749-7303

Manage Your Account

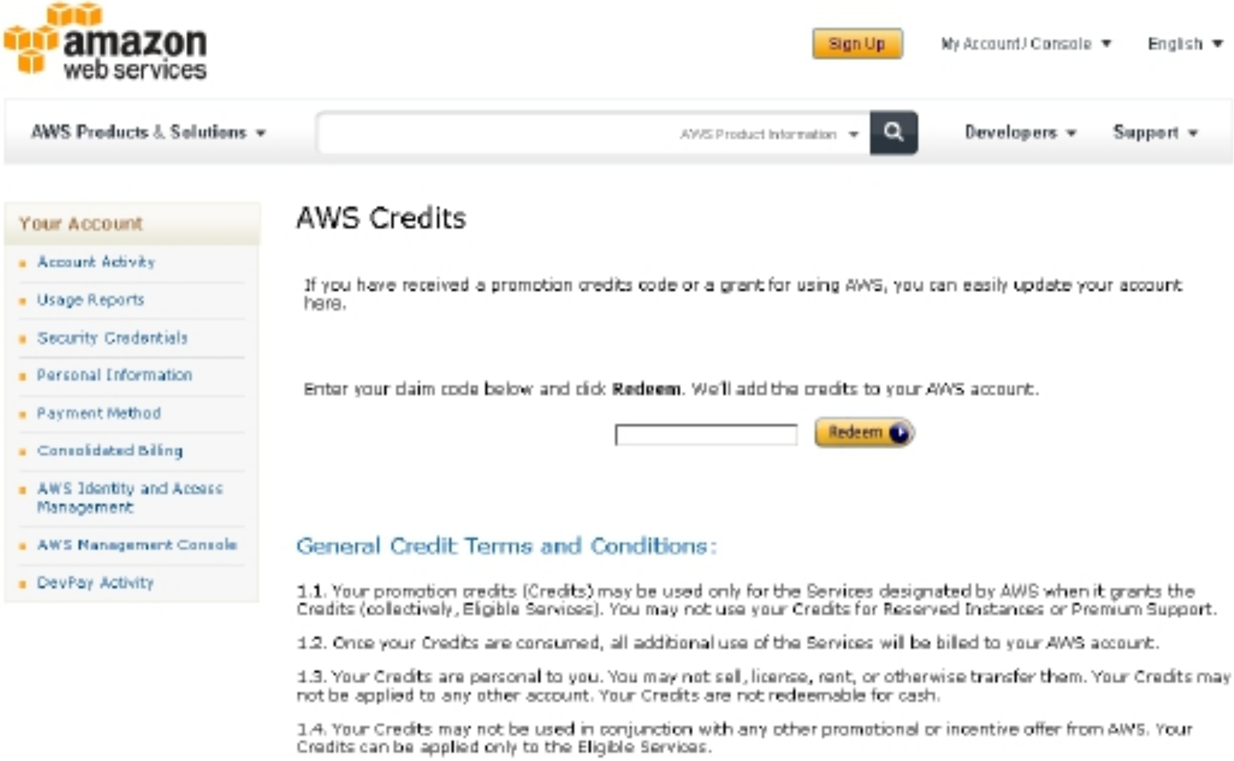
You already have access to Amazon Web Services

Services You're Signed Up For

- Amazon CloudFormation
- Amazon CloudFront
- Amazon CloudSearch
- Amazon CloudWatch
- Amazon DynamoDB
- Amazon Elastic Compute Cloud (EC2)
- Amazon Elastic MapReduce
- Amazon Elastic Transcoder
- Amazon ElastiCache
- Amazon Glacier
- Amazon Mechanical Turk
- Amazon Redshift
- Amazon Simple Queue Service (SQS)
- Amazon Simple Storage Service (S3)
- Amazon Simple Workflow Service (SWF)
- Amazon SimpleDB
- Amazon Virtual Private Cloud (VPC)
- Auto Scaling
- AWS Data Pipeline
- AWS Direct Connect
- AWS Elastic Beanstalk
- AWS Import/Export
- AWS OpsWorks
- AWS Storage Gateway

1. Setting up AWS account

- Free Credit: https://aws.amazon.com/awscredits/
  - Should have received your AWS credit code by email
  - $100 worth of credits should be enough
- Don't forget to terminate your clusters to avoid extra charges!



AWS Credits

If you have received a promotion credits code or a grant for using AWS, you can easily update your account here.

Enter your claim code below and click **Redeem**. We'll add the credits to your AWS account.

General Credit Terms and Conditions:

1.1. Your promotion credits (Credits) may be used only for the Services designated by AWS when it grants the Credits (collectively, Eligible Services). You may not use your Credits for Reserved Instances or Premium Support.

1.2. Once your Credits are consumed, all additional use of the Services will be billed to your AWS account.

1.3. Your Credits are personal to you. You may not sell, license, rent, or otherwise transfer them. Your Credits may not be applied to any other account. Your Credits are not redeemable for cash.

1.4. Your Credits may not be used in conjunction with any other promotional or incentive offer from AWS. Your Credits can be applied only to the Eligible Services.

2. Setting up an EC2 key pair

- Go to EC2 Management Console https://console.aws.amazon.com/ec2/
- Pick region in navigation bar (top right)
- Click on *Key Pairs* and click *Create Key Pair*
- Enter name and click *Create*
- Download of .pem private key
  - lets you access EC2 instance
  - Only time you can download the key

2. Setting up an EC2 key pair (Linux/Mac)

- Change the file permission

```
$ chmod 600 </path/to/saved/keypair/file.pem>
```

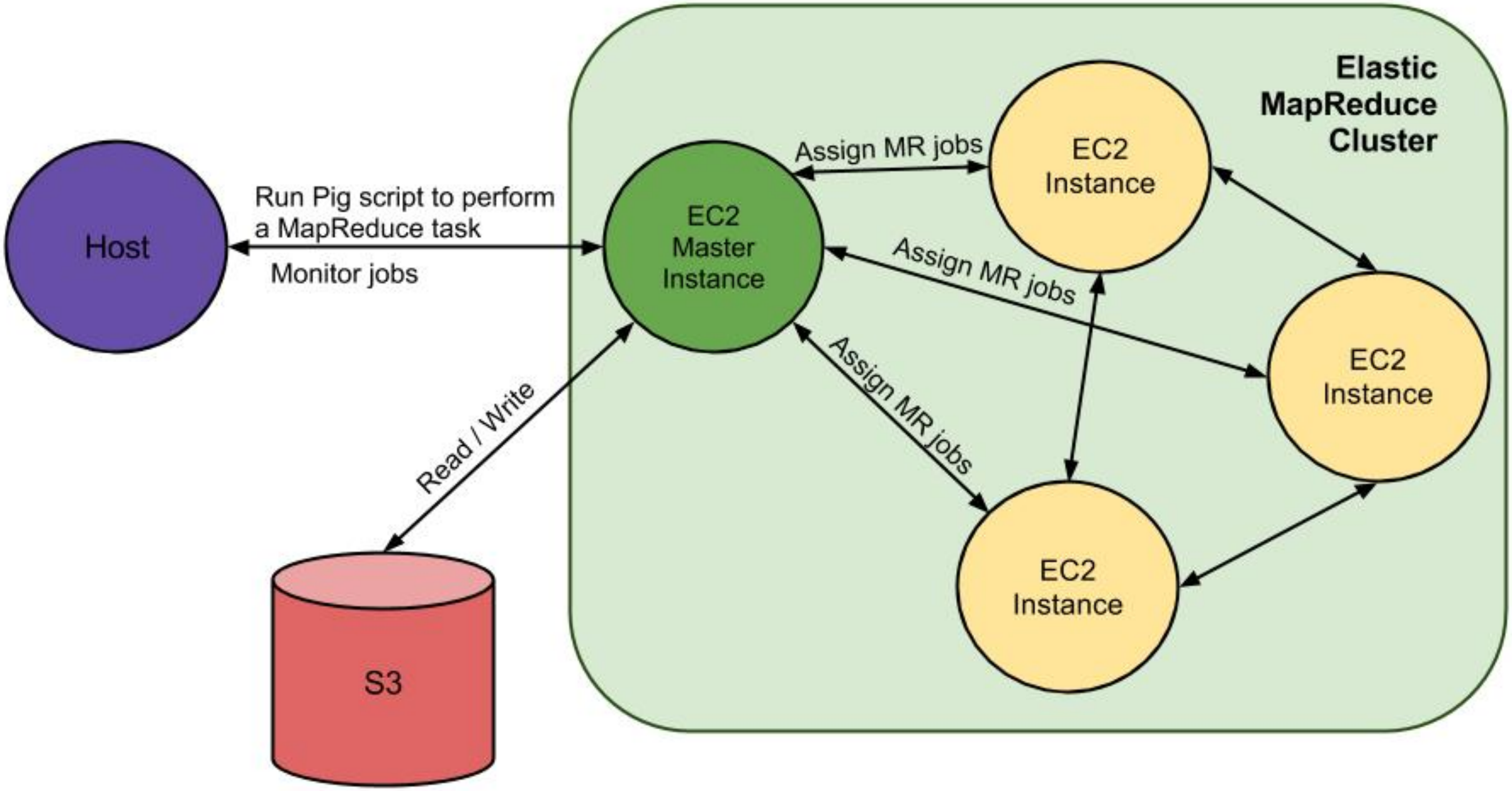
2. Setting up an EC2 key pair (Windows)

- AWS instruction: http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html
- Use PuTTYGen to convert a key pair from .pem to .ppk
- Use PuTTY to establish a connection to EC2 master instance

2. Setting up an EC2 key pair

- Note: Some students were having problem running job flows (next task after setting EC2 key pair) because of no active key found
- If so, go to AWS security credentials page and make sure that you see a key under the access key, if not just click Create a new Access Key.

  https://portal.aws.amazon.com/gp/aws/securityCredentials


Elastic MapReduce Cluster
Host
Run Pig script to perform a MapReduce task
Monitor jobs
EC2 Master Instance
Assign MR jobs
EC2 Instance
Assign MR jobs
EC2 Instance
Assign MR jobs
EC2 Instance
Read / Write
S3

Where is your input file?

- Your input files come from Amazon S3
- You will use three sets, each of different size
  - s3n://uw-cse344-test/cse344-test-file -- 250KB
  - s3n://uw-cse344/btc-2010-chunk-000 -- 2GB
  - s3n://uw-cse344 -- 0.5TB
- See example.pig for how to load the dataset

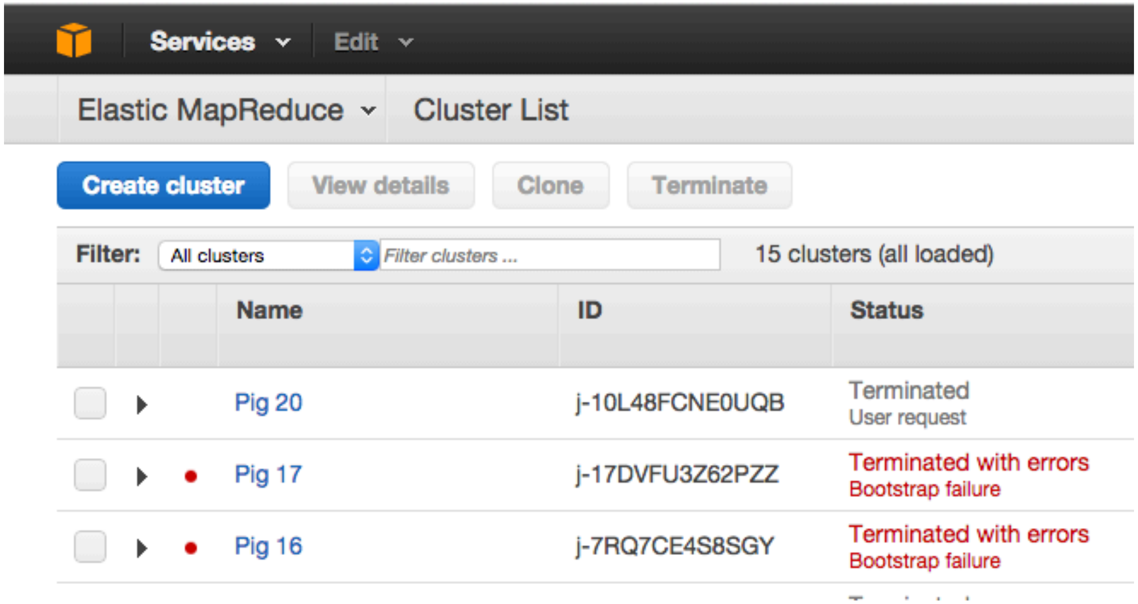
```
raw = LOAD 's3n://uw-cse344-test/cse344-test-file' USING TextLoader as (line:chararray);
```

Instance Types & Pricing

- http://aws.amazon.com/ec2/instance-types/
- http://aws.amazon.com/ec2/pricing/

3. Starting an AWS cluster

- http://console.aws.amazon.com/elasticmapreduce/vnext/home
- Click *Amazon Elastic Map Reduce* Tab
- Click *Create Cluster*



Filter: All clusters | Filter clusters ... | 15 clusters (all loaded)

Name	ID	Status
Pig 20	j-10L48FCNE0UQB	Terminated User request
● Pig 17	j-17DVFU3Z62PZZ	Terminated with errors Bootstrap failure
● Pig 16	j-7RQ7CE4S8SGY	Terminated with errors Bootstrap failure

3. Starting an AWS Cluster

- Enter some "Cluster name”
- Uncheck "Enabled" for "Logging”
- Choose hadoop distribution 2.4.9
- In the "Hardware Configuration" section, change the count of core instances to 1.
- In the "Security and Access" section, select the EC2 key pair you created above.
- Create default roles for both roles under IAM roles.
- Click "Create cluster" at the bottom of the page. You can go back to the cluster list and should see the cluster you just created.

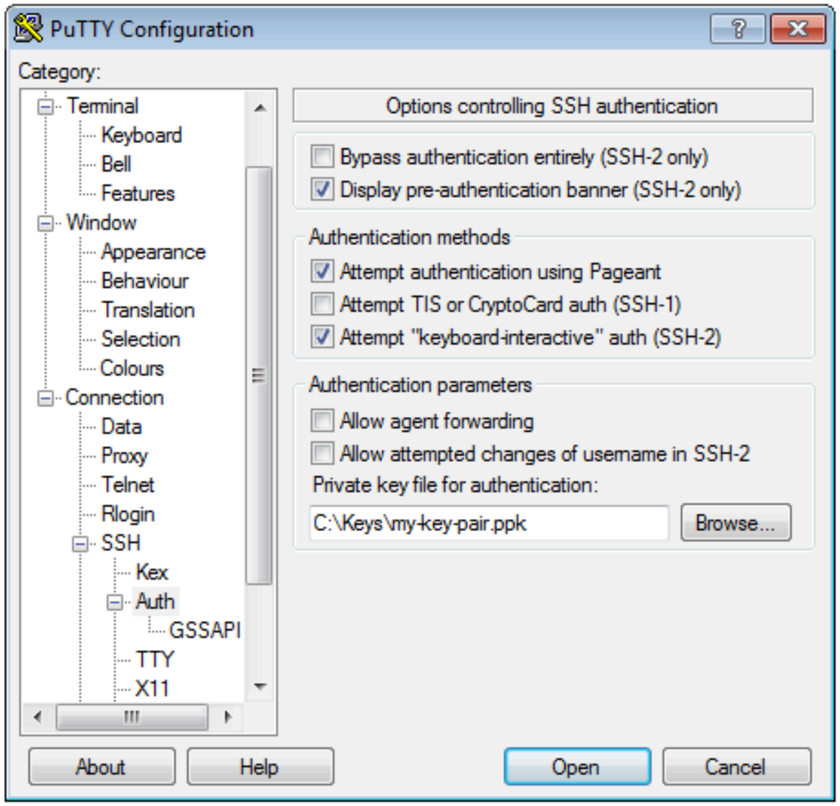
Connecting to the master

- Click on cluster name. You will find the Master Public DNS at the top.

-
  ```
  $ ssh -o "ServerAliveInterval 10"
      -L 9100:localhost:9100
      -i </path/to/saved/keypair/file.pem>
       hadoop@<master.public-dns-name.amazonaws.com>
  ```

Connecting to the master in Windows

- http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html



Options controlling SSH authentication

□ Bypass authentication entirely (SSH-2 only)
☑ Display pre-authentication banner (SSH-2 only)

Authentication methods

☑ Attempt authentication using Pageant
□ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

□ Allow agent forwarding
□ Allow attempted changes of username in SSH-2

Private key file for authentication:
C:\Keys\my-key-pair.ppk

For tunneling (to monitor jobs)

1. Choose Tunnels
2. Put source port as 9100
3. Put destination as
   localhost:9100

4. Press Add (Don't forget this)

4. Running Pig interactively

- Once you successfully made a connection to EC2 cluster, type pig, and it will show

```
grunt>
```

- Time to write some pig queries!
- To run a pig script – use $pig example.pig

Lets run example.pig

```
register s3n://uw-cse344-code/myudfs.jar

raw = LOAD 's3n://uw-cse344-test/cse344-test-file' USING TextLoader as
(line:chararray);

ntriples = foreach raw generate FLATTEN(myudfs.RDFSplit3(line)) as
(subject:chararray,predicate:chararray,object:chararray);

objects = group ntriples by (object) PARALLEL 50;

count_by_object = foreach objects generate flatten($0), COUNT($1) as count
PARALLEL 50;

count_by_object_ordered = order count_by_object by (count)  PARALLEL 50;

store count_by_object_ordered into '/user/hadoop/example-results8' using
PigStorage();
```

OR

```
store count_by_object_ordered into ’s3://mybucket/myfile’;
```

5. Monitoring Hadoop jobs

Possible options are:

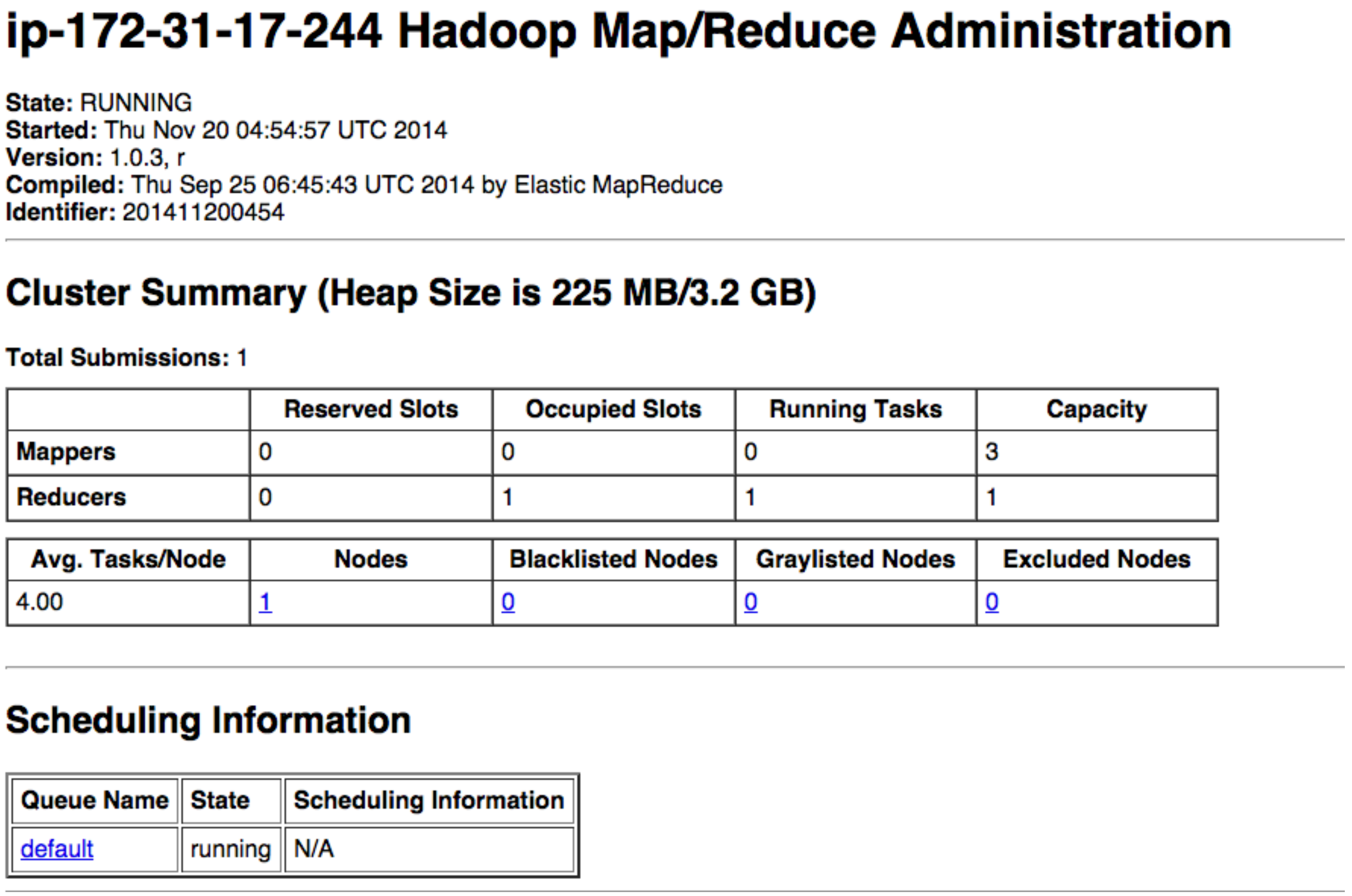
1. Using ssh tunneling (recommended)

```
ssh -L 9100:localhost:9100 -o "ServerAliveInterval 10"
-i </path/to/saved/keypair/file.pem>
hadoop@<master.public-dns-name.amazonaws.com>
```

2. Using LYNX

```
lynx http://localhost:9100/
```

3. Using SOCKS proxy



ip-172-31-17-244 Hadoop Map/Reduce Administration

State: RUNNING
Started: Thu Nov 20 04:54:57 UTC 2014
Version: 1.0.3, r
Compiled: Thu Sep 25 06:45:43 UTC 2014 by Elastic MapReduce
Identifier: 201411200454

Cluster Summary (Heap Size is 225 MB/3.2 GB)

Total Submissions: 1

	Reserved Slots	Occupied Slots	Running Tasks	Capacity
Mappers	0	0	0	3
Reducers	0	1	1	1

Avg. Tasks/Node	Nodes	Blacklisted Nodes	Graylisted Nodes	Excluded Nodes
4.00	1	0	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Where is your output stored?

- Two options
  1. Hadoop File System

     The AWS Hadoop cluster maintains its own HDFS instance, which dies with the cluster -- this fact is not inherent in HDFS. Don’t forget to copy them to your local machine before terminating the job.

  2. S3

     S3 is persistent storage. But S3 costs money while it stores data. Don’t forget to delete them once you are done.

- It will output a set of files stored under a directory. Each file is generated by a reduce worker to avoid contention on a single output file.

How can you get the output files?

1. Easier and expensive way:
   - Create your own S3 bucket(file system), write the output there
   - Output filenames become s3n://your-bucket/outdir
   - Can download the files via S3 Management Console
   - But S3 does cost money, even when the data isn't going anywhere. DELETE YOUR DATA ONCE YOU'RE DONE!

How can you get the output files?

1. Harder and cheapskate way:
   - Write to cluster's HDFS (see example.pig)
   - Output directory name is /user/hadoop/outdir.
   - Need to double download
     1. from HDFS to master node's filesystem with

        hadoop fs –copyToLocal

        eg. hadoop fs -copyToLocal /user/hadoop/example-results ./res

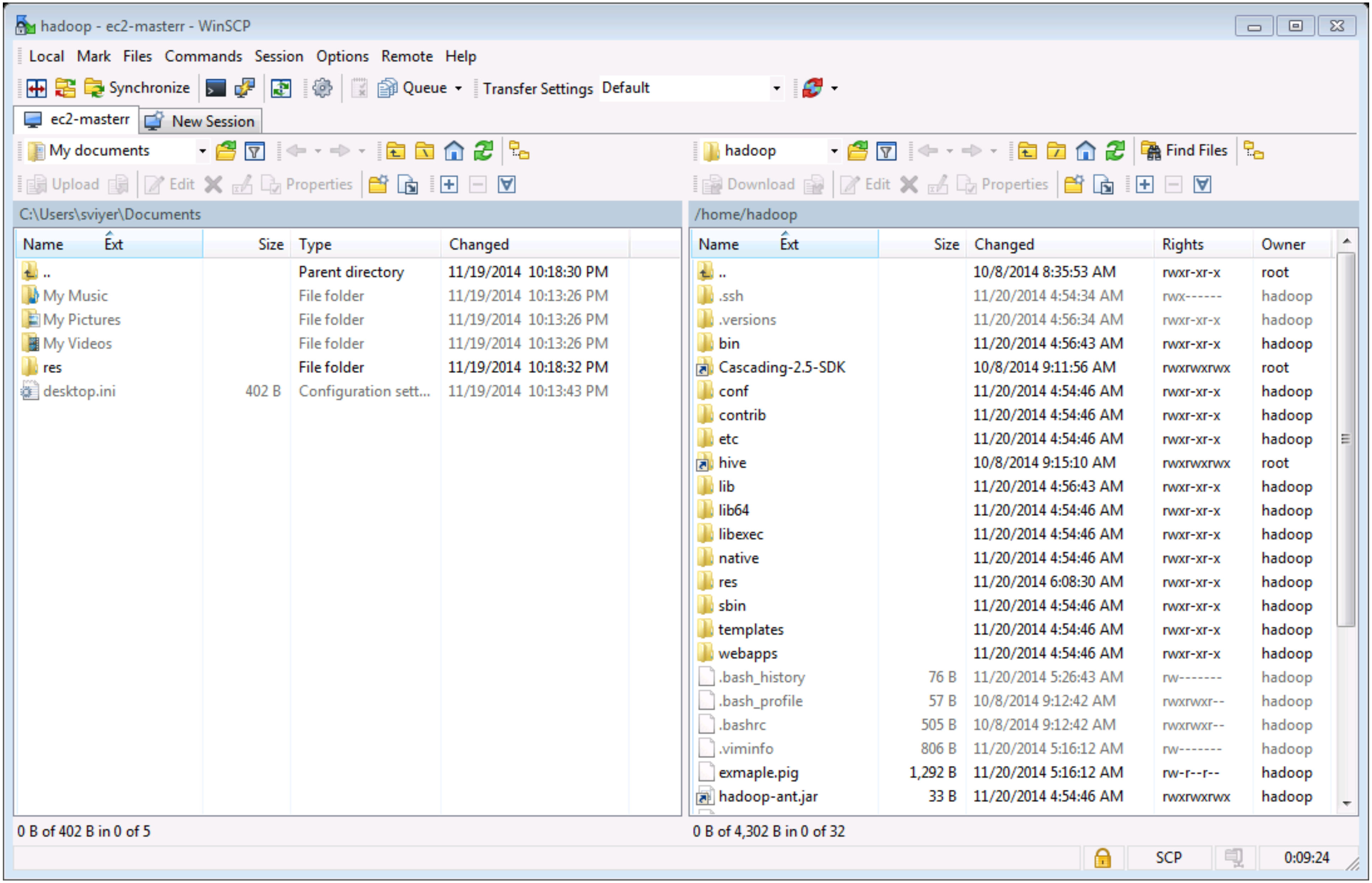
     2. from master node to local machine with scp

        Linux: scp -r -i /path/to/key

        hadoop@ec2-54-148-11-252.us-west-2.compute.amazonaws.com:res <local_folder>

Transfer the files using Windows

- Launch WinSCP
- Set File Protocol to SCP
- Enter master public dns name
- Set the port as 22
- Set the username as hadoop
- Choose Advanced
- Choose >SSH>Authentication (left menu)
- Uncheck all boxes
- Then check all boxes under GSSAPI
- Load your private key file (which you created using puttygen) .. Press OK
- Save the connection and double click on the entry



Transfer Settings Default

My documents

C:\Users\sviyer\Documents

Name	Ext	Size	Type	Changed
..			Parent directory	11/19/2014 10:18:30 PM
My Music			File folder	11/19/2014 10:13:26 PM
My Pictures			File folder	11/19/2014 10:13:26 PM
My Videos			File folder	11/19/2014 10:13:26 PM
res			File folder	11/19/2014 10:18:32 PM
desktop.ini		402 B	Configuration sett...	11/19/2014 10:13:43 PM

0 B of 402 B in 0 of 5

hadoop

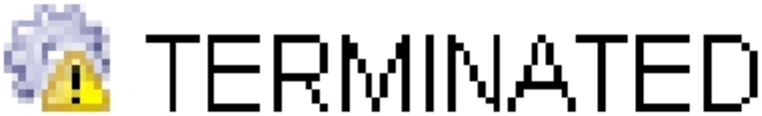
/home/hadoop

Name	Ext	Size	Changed	Rights	Owner
..			10/8/2014 8:35:53 AM	rwxr-xr-x	root
.ssh			11/20/2014 4:54:34 AM	rwx------	hadoop
.versions			11/20/2014 4:56:34 AM	rwxr-xr-x	hadoop
bin			11/20/2014 4:56:43 AM	rwxr-xr-x	hadoop
Cascading-2.5-SDK			10/8/2014 9:11:56 AM	rwxrwxrwx	root
conf			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
contrib			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
etc			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
hive			10/8/2014 9:15:10 AM	rwxrwxrwx	root
lib			11/20/2014 4:56:43 AM	rwxr-xr-x	hadoop
lib64			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
libexec			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
native			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
res			11/20/2014 6:08:30 AM	rwxr-xr-x	hadoop
sbin			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
templates			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
webapps			11/20/2014 4:54:46 AM	rwxr-xr-x	hadoop
.bash_history		76 B	11/20/2014 5:26:43 AM	rw-------	hadoop
.bash_profile		57 B	10/8/2014 9:12:42 AM	rwxrwxr--	hadoop
.bashrc		505 B	10/8/2014 9:12:42 AM	rwxrwxr--	hadoop
.viminfo		806 B	11/20/2014 5:16:12 AM	rw-------	hadoop
exmaple.pig		1,292 B	11/20/2014 5:16:12 AM	rw-r--r--	hadoop
hadoop-ant.jar		33 B	11/20/2014 4:54:46 AM	rwxrwxrwx	hadoop

0 B of 4,302 B in 0 of 32

6. Terminating Cluster

- Go to Management Console > EMR
- Select Cluster List
- Click on your cluster
- Press Terminate
- Wait a few minutes ...
- Eventually status should be



TERMINATED

Final Comment

- Start early
- Important: read the spec carefully!

  If you get stuck or have an unexpected outcome, it is likely that you miss some step or there may be important directions/notes in the spec.
- Running jobs may take up to several hours
  - Last problem takes about ~4 hours.