CSE 344 Introduction to Data Management

Section 9: AWS, Hadoop, Pig Latin

TA: Daseul Lee (dslee@cs)

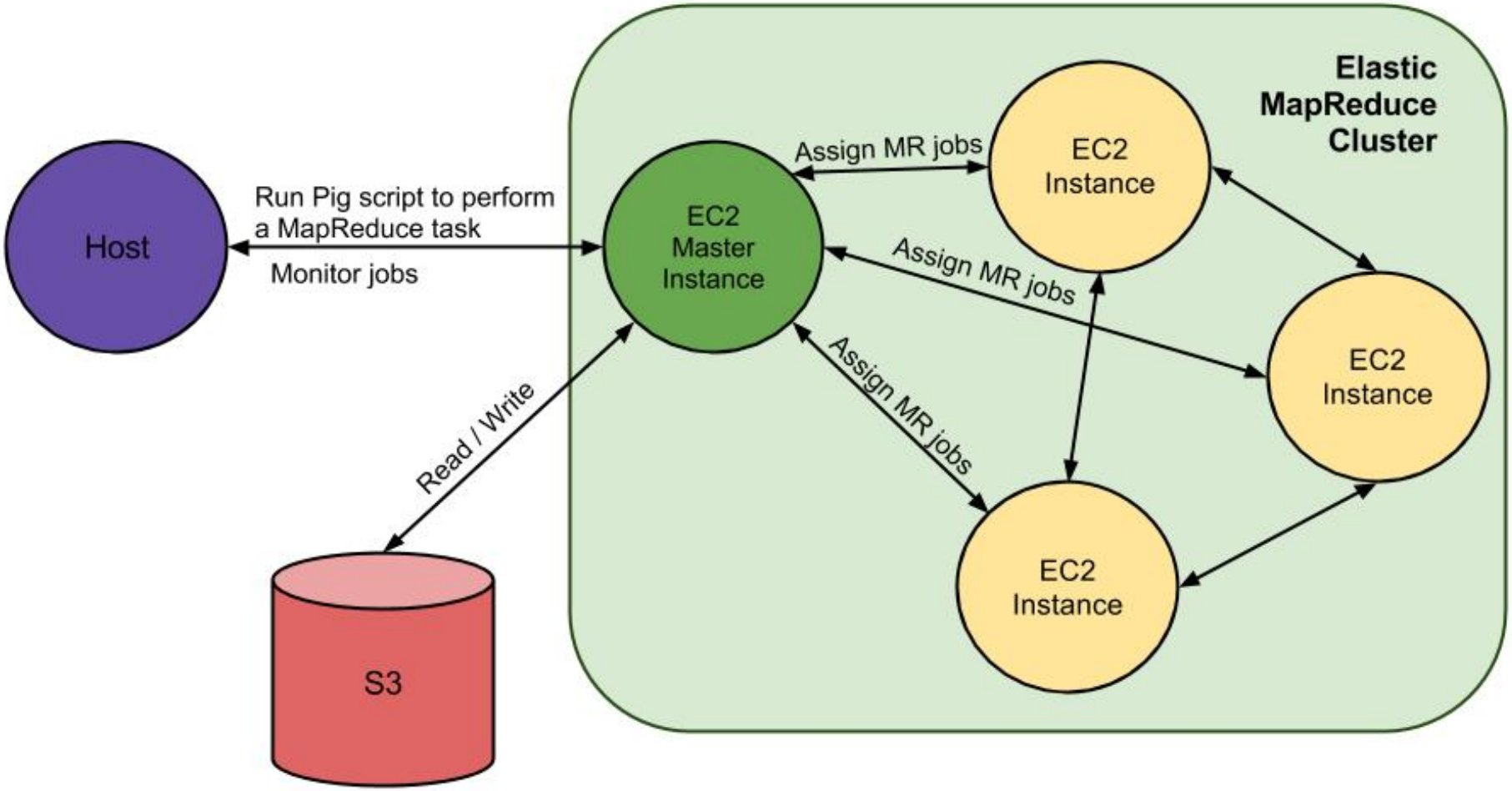
Homework 8

- Big Data analysis on billion triple dataset using Amazon Web Service (AWS)
  - Billion Triple Set: contains web information, obtained by a crawler

    (subject, predicate, object)
  - Working with up to 0.5 TB of data
- You will write pig queries for each task and use MapReduce to perform data analysis.
- Due Friday 12/6
- No late days!

Overview

- AWS offers various cloud computing services. In this assignment, we will use:
  - **Elastic MapReduce**: Managed Hadoop Framework
  - **EC2** (Elastic Computing Cluster): virtual servers in the cloud
  - **S3** (Simple Storage Service): scalable storage in the cloud


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);
```

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?

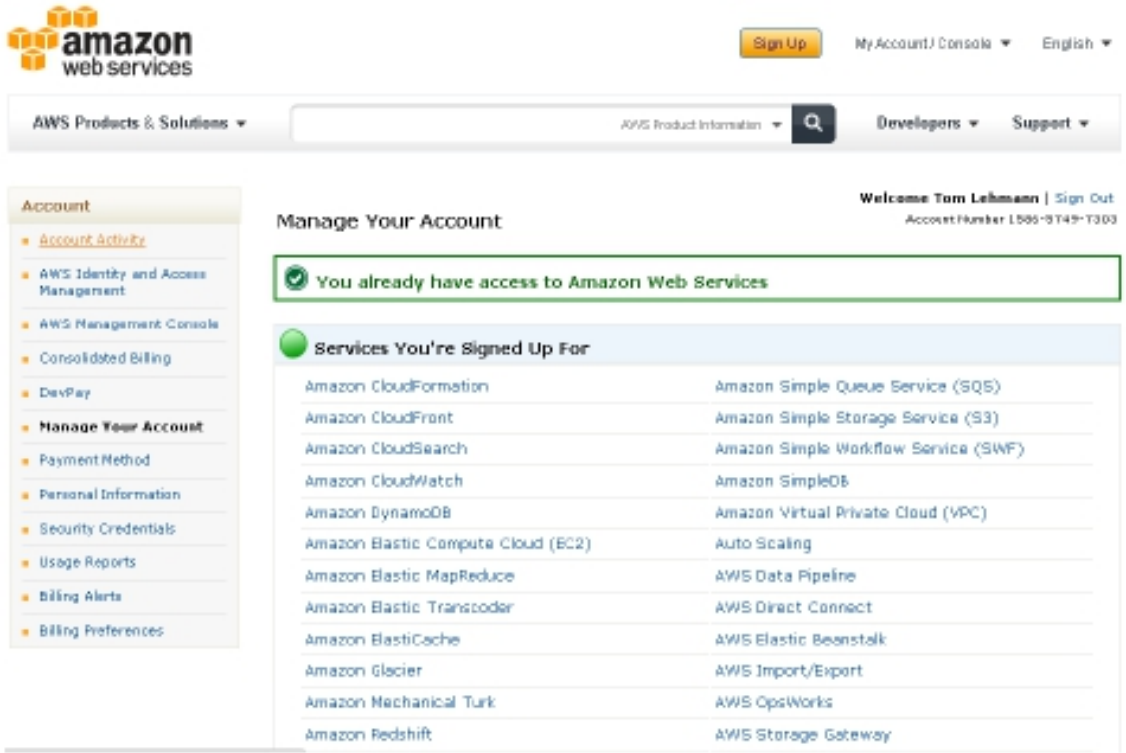
1. Easier and expensive way:
   - Create your own S3 bucket(file system), write the output there
   - Output filenames become s3n://your-bucket/outdir
   - Can download the files via S3 Management Console
   - But S3 does cost money, even when the data isn't going anywhere. DELETE YOUR DATA ONCE YOU'RE DONE!
2. Harder and cheapskate way:
   - Write to cluster's HDFS
   - Output directory name is /user/hadoop/outdir. You'll need to create /user/hadoop
   - Need to double download
     1. from HDFS to master node's filesystem with *hadoop dfs –copyToLocal*
     2. from master node to local machine with scp

Set-up

(Disclaimer: Important details are found in the spec)

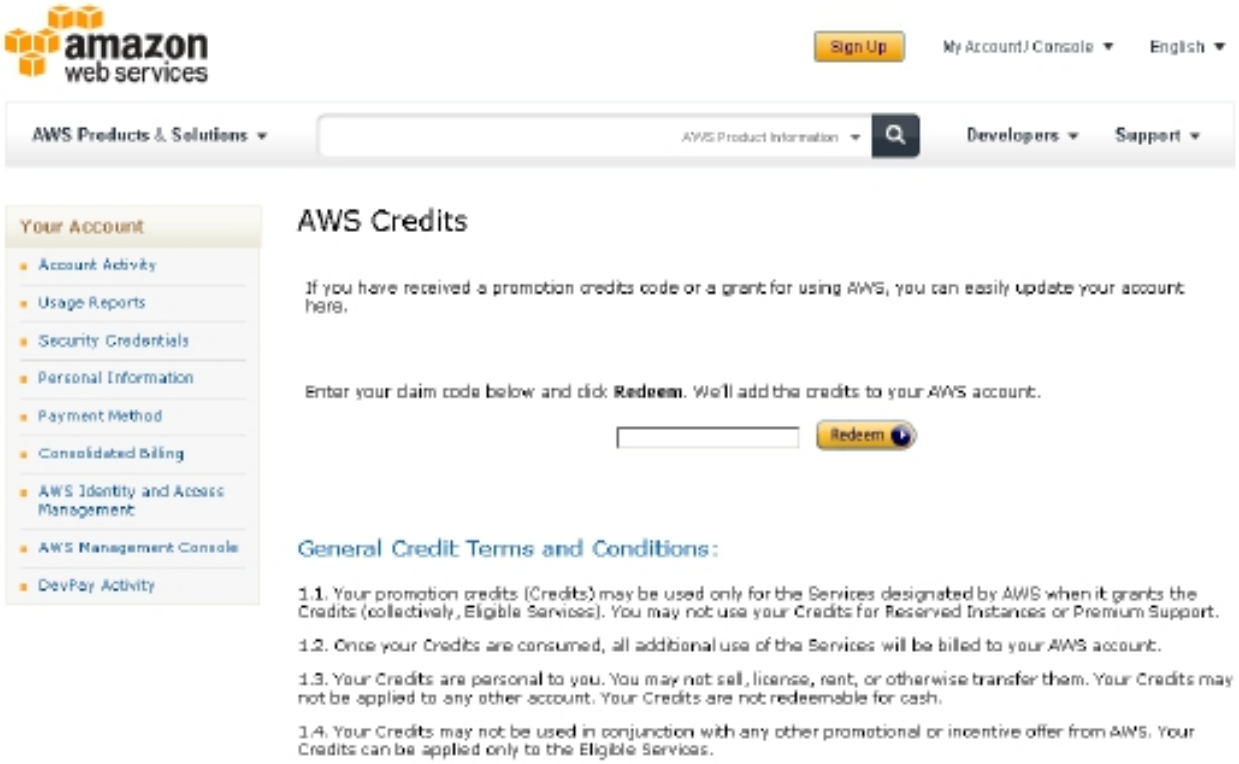
1. Setting up AWS account

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



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 job flows to avoid extra charges!



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/gettingstarted/latest/computebasics-linux/getting-started-deploy-app-connect.html
- Use PuTTYGen to convert a key pair from .pem to .ppk (part 1 – 2)
- Use PuTTY to establish a connection to EC2 master instance (part 3 – 6)

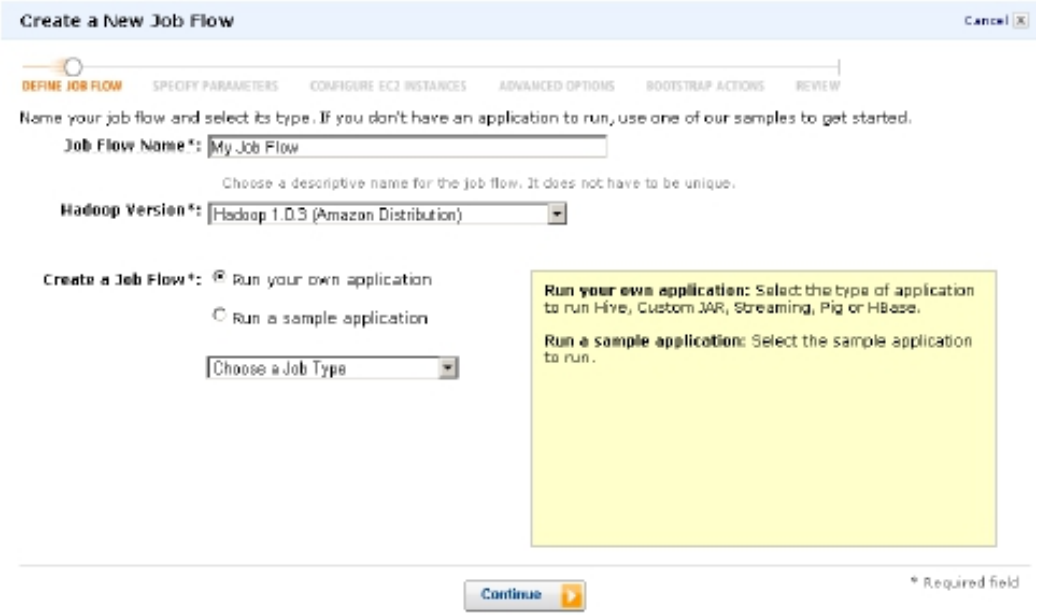
2. Setting up an EC2 key pair

- 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

3. Starting an AWS cluster

- http://console.aws.amazon.com/elasticmapreduce/home
- Click *Amazon Elastic Map Reduce* Tab
- Click *Create New Job Flow*

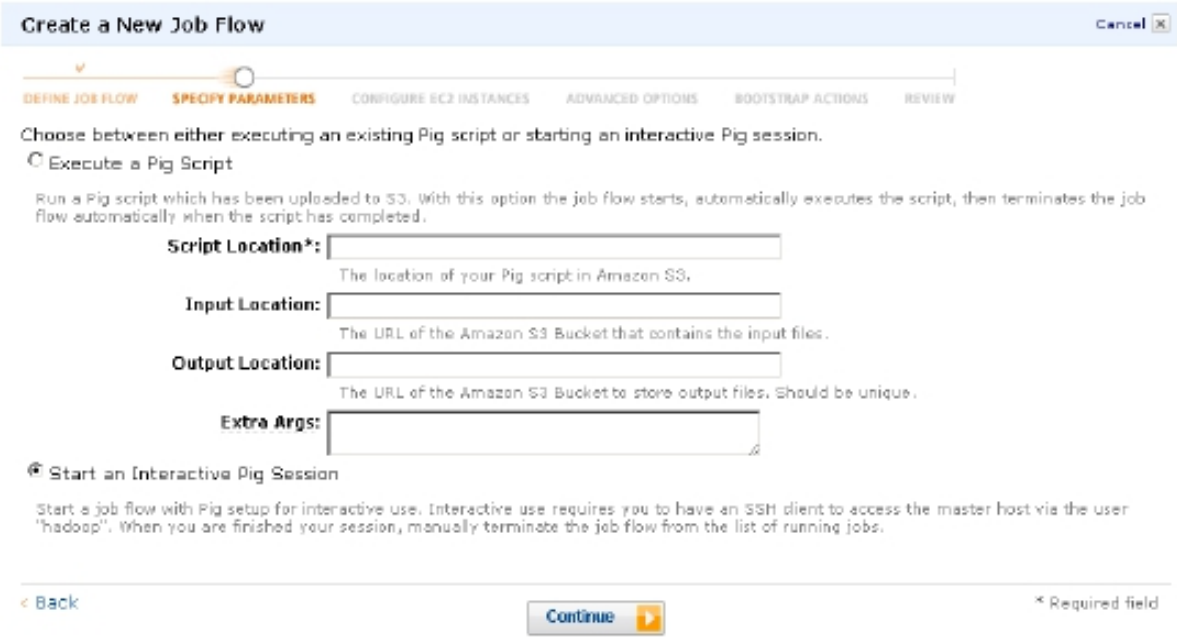


3. Starting an AWS Cluster

- Name the Job Flow
- Select Run your own application
- Select Pig Program as Job Type
- CONTINUE

3. Starting an AWS Cluster

- Select Start an Interactive Pig Session
- CONTINUE

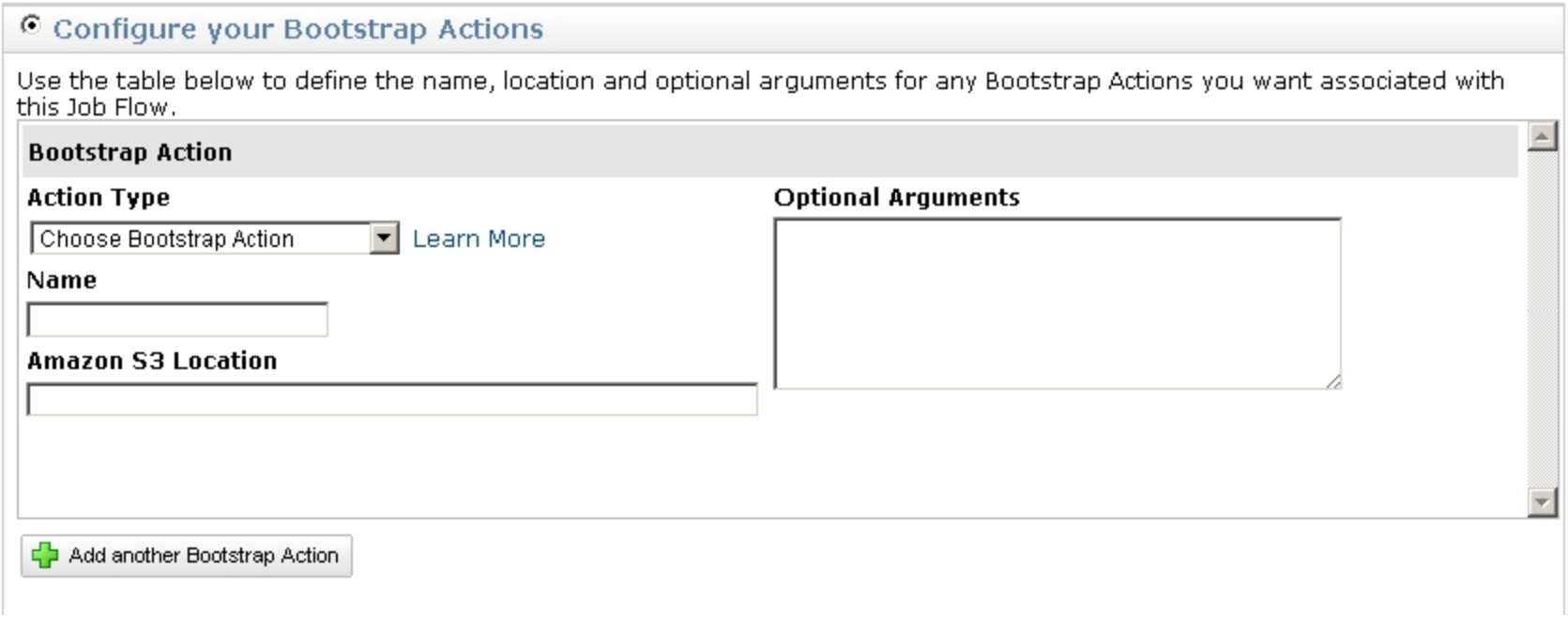


3. Starting an AWS Cluster

- Select only 1 core instance
- CONTINUE
- Set your previously created Key Pair to be the Amazon EC2 Key Pair
- CONTINUE

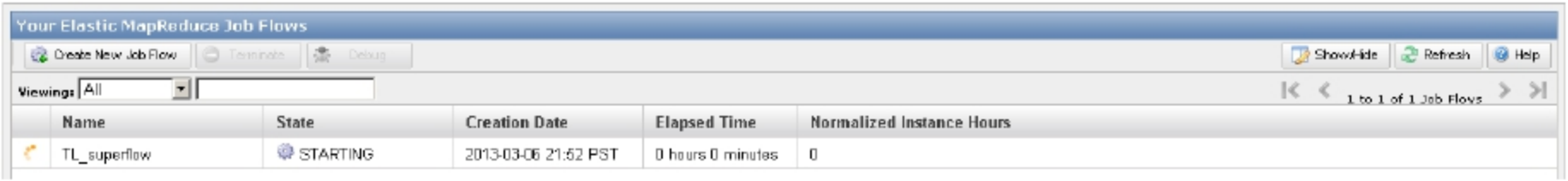
3. Starting an AWS Cluster

- Configure your Bootstrap Actions
- Action Type: Memory Intensive Configuration



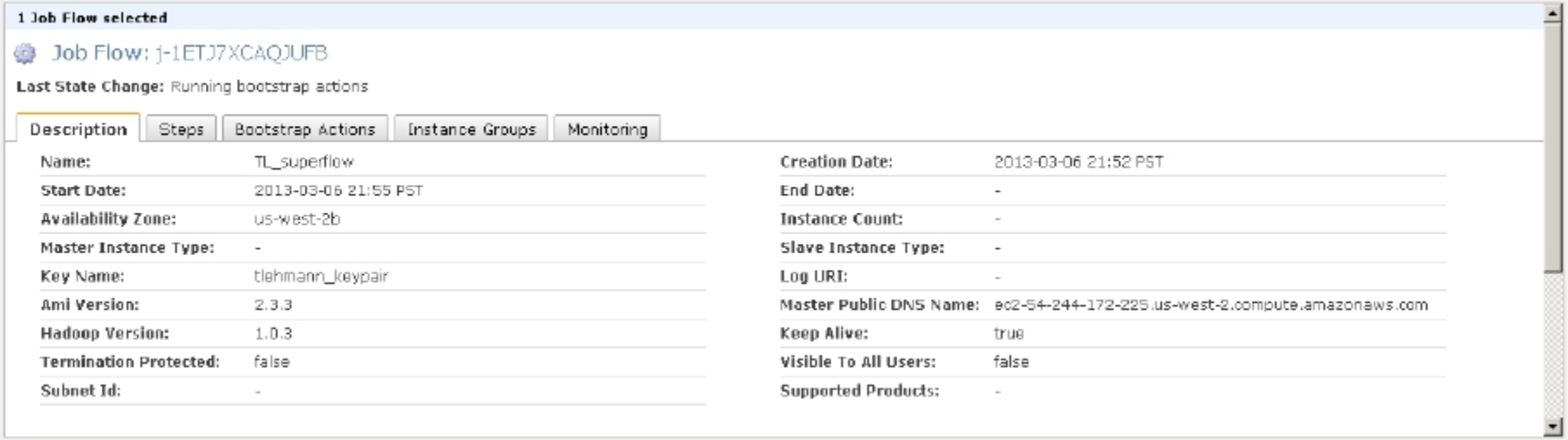
3. Starting an AWS Cluster

- CONTINUE
- Create Job Flow
- Refresh page to see your job flow (might take a few minutes...)



3. Starting an AWS Cluster

- Click on your Job Flow
- Retrieve the Master Public DNS Name



3. Starting an AWS Cluster

- Windows users use PuTTY to connect to cluster
- Everybody else runs this from command line

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

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!

4. Running Pig interactively

example.pig

- Found in the project archive
- Loads and parses billion triple dataset:

  Triples (subject, predicate, object)
- Group object by attribute, sort in descending order based on count of tuple
- Check out the README for more information

5. Monitoring Hadoop jobs

Possible options are:

1. Using ssh tunneling (recommended)
2. Using LYNX
3. Using SOCKS proxy

6. Terminating Cluster

- Go to Management Console
- Select Job Flow
- Click Terminate
- Wait a few minutes ...
- Eventually status should be

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
  - Extra credit problem takes about ~4 hours.