

Section 9

CSE 344

3/7/2013

Homework 8

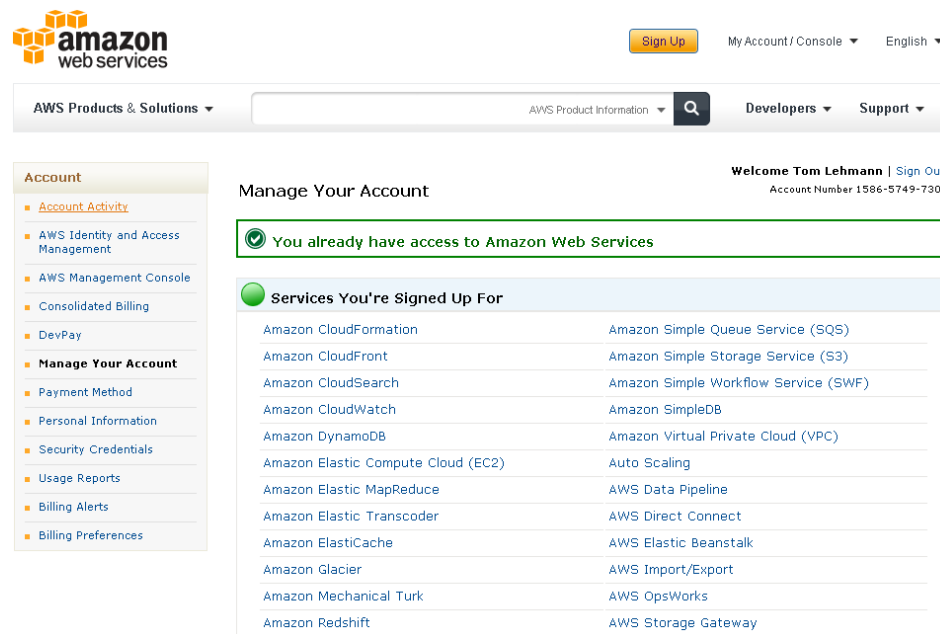
- Final homework, yeah!
- **Important**: no late days allowed!
- Due Friday, 3/15

Connecting to AWS

- <http://aws.amazon.com/>
- Step-by-step guide on the website in the homework spec
- Show you the first steps towards finishing Problem 0

Sign-In

- Make sure that you're signed up for Amazon Elastic MapReduce, Simple Storage Service (S3) and Elastic Compute Cloud (EC2)

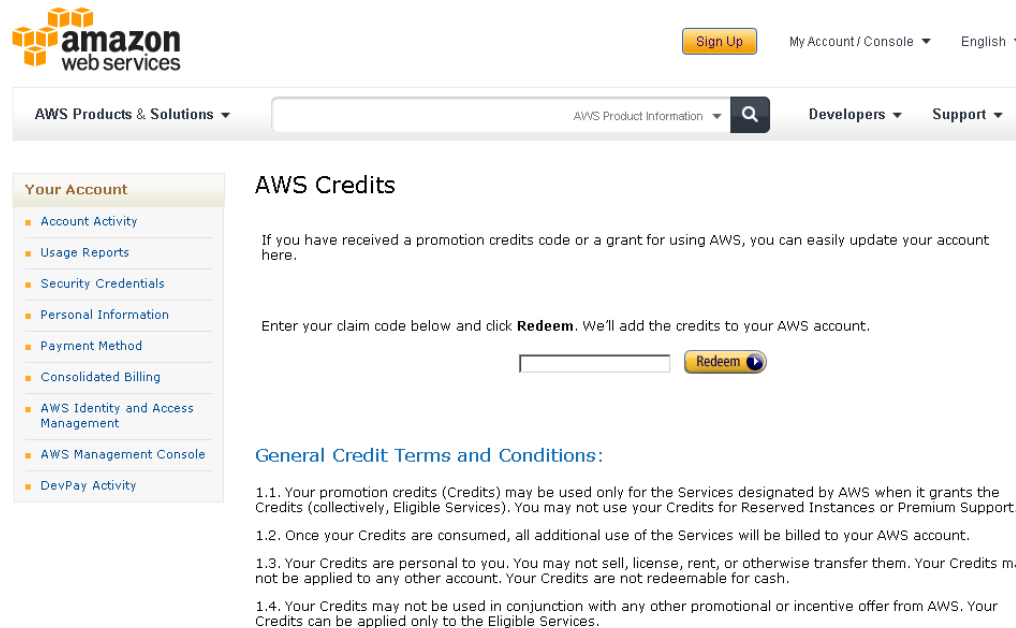


The screenshot displays the Amazon Web Services console interface. At the top left is the Amazon logo and 'amazon web services' text. To the right are links for 'Sign Up', 'My Account / Console', and 'English'. Below this is a navigation bar with 'AWS Products & Solutions', a search bar, 'AWS Product Information', 'Developers', and 'Support'. The main content area is titled 'Manage Your Account' and includes a green confirmation message: 'You already have access to Amazon Web Services'. Below this is a section 'Services You're Signed Up For' containing a table of services.

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

Free Credits!

- <http://aws.amazon.com/awscredits/>
- Should have received code from Lee Lee on/around March 4



The screenshot shows the AWS Credits page in the AWS Management Console. At the top, there is the Amazon Web Services logo, a 'Sign Up' button, and links for 'My Account / Console' and 'English'. Below the logo is a navigation bar with 'AWS Products & Solutions', a search bar, and links for 'AWS Product Information', 'Developers', and 'Support'. On the left side, there is a 'Your Account' menu with links for Account Activity, Usage Reports, Security Credentials, Personal Information, Payment Method, Consolidated Billing, AWS Identity and Access Management, AWS Management Console, and DevPay Activity. The main content area is titled 'AWS Credits' and contains the following text: 'If you have received a promotion credits code or a grant for using AWS, you can easily update your account here.' Below this is a form with a text input field and a 'Redeem' button. At the bottom, there is a section for 'General Credit Terms and Conditions:' with four numbered items: 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.

More about Credits

- Amazon charges 10 cents/node/hour
- \$100 worth of credits should be enough
- DON'T forget to terminate your job flows!

Have AWS create a key pair for you

- Go to EC2 Management Console
- <https://console.aws.amazon.com/ec2/>
- Pick region in navigation bar (top right)
- Click on *Key Pairs*
- Click *Create Key Pair*
- Enter name, click *Create*
- Download of .pem file begins which is needed to access any of your instances

Have AWS create a key pair for you

- People using Windows need to set up PuTTY
- <http://docs.aws.amazon.com/gettingstarted/latest/wah-linux/getting-started-deploy-app-connect.html>
- Everybody else just uses the command
- `$ chmod 600 </path/to/saved/keypair/file.pem>`

Starting an AWS cluster

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

Create a New Job Flow Cancel

DEFINE JOB FLOW SPECIFY PARAMETERS CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Name your job flow and select its type. If you don't have an application to run, use one of our samples to get started.

Job Flow Name*:

Choose a descriptive name for the job flow. It does not have to be unique.

Hadoop Version*:

Create a Job Flow*: Run your own application
 Run a sample application

Run your own application: Select the type of application to run Hive, Custom JAR, Streaming, Pig or HBase.

Run a sample application: Select the sample application to run.

Continue * Required field

Starting an AWS Cluster

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

Starting an AWS Cluster

- Select *Start an Interactive Pig Session*
- CONTINUE

Create a New Job Flow Cancel X

DEFINE JOB FLOW **SPECIFY PARAMETERS** CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Choose between either executing an existing Pig script or starting an interactive Pig session.

Execute a Pig Script

Run a Pig script which has been uploaded to S3. With this option the job flow starts, automatically executes the script, then terminates the job flow automatically when the script has completed.

Script Location*:
The location of your Pig script in Amazon S3.

Input Location:
The URL of the Amazon S3 Bucket that contains the input files.

Output Location:
The URL of the Amazon S3 Bucket to store output files. Should be unique.

Extra Args:

Start an Interactive Pig Session

Start a job flow with Pig setup for interactive use. Interactive use requires you to have an SSH client to access the master host via the user "hadoop". When you are finished your session, manually terminate the job flow from the list of running jobs.

[< Back](#) [Continue](#) * Required field

Starting an AWS Cluster

- Select only 1 core instance
- CONTINUE

- Set your previously created Key Pair to be the Amazon EC2 Key Pair
- CONTINUE

Starting an AWS Cluster

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

Configure your Bootstrap Actions

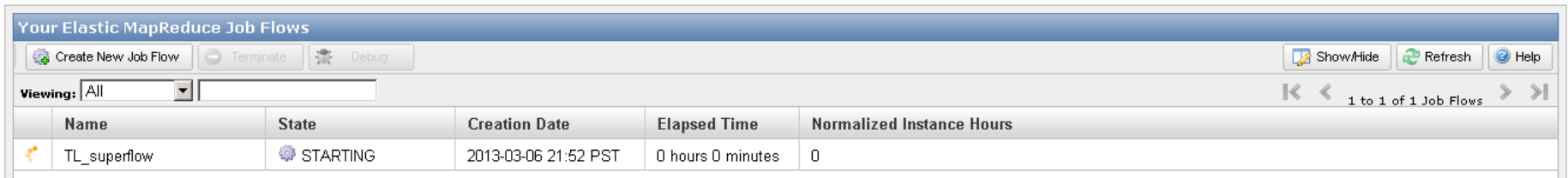
Use the table below to define the name, location and optional arguments for any Bootstrap Actions you want associated with this Job Flow.

Bootstrap Action	
Action Type Choose Bootstrap Action <input type="button" value="▼"/> Learn More	Optional Arguments <div style="border: 1px solid gray; height: 100px;"></div>
Name <input type="text"/>	
Amazon S3 Location <input type="text"/>	

[+ Add another Bootstrap Action](#)

Starting an AWS Cluster

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

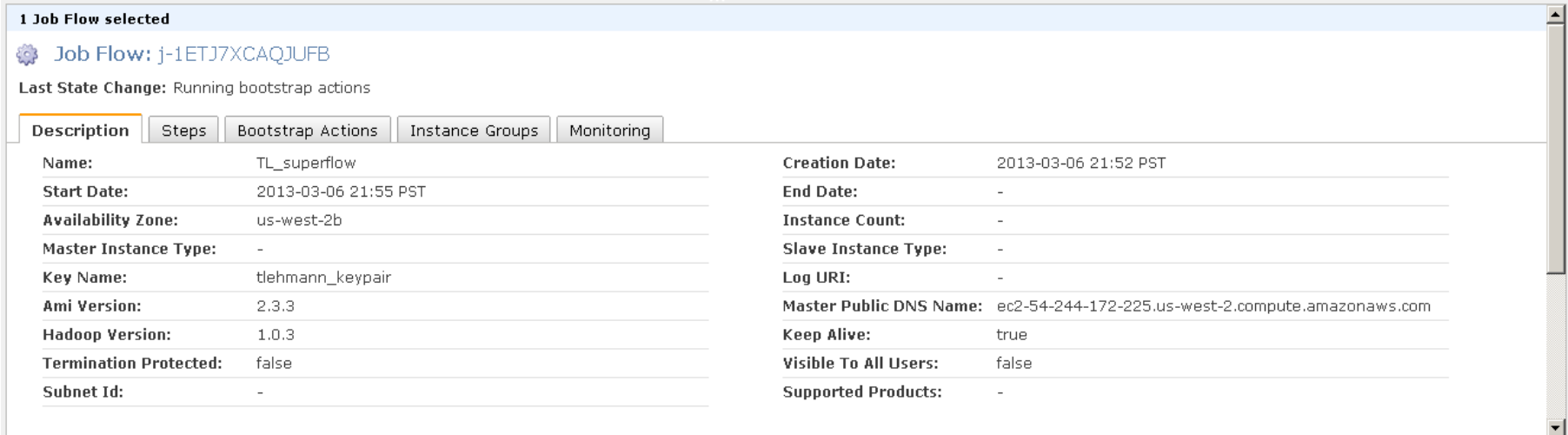


The screenshot shows the AWS Elastic MapReduce console interface. At the top, there's a header "Your Elastic MapReduce Job Flows". Below the header, there are several buttons: "Create New Job Flow", "Terminate", and "Debug". On the right side, there are "Show/Hide", "Refresh", and "Help" buttons. Below these buttons, there's a "Viewing:" dropdown menu set to "All" and a search input field. The main content is a table with the following columns: "Name", "State", "Creation Date", "Elapsed Time", and "Normalized Instance Hours". The table contains one row with the following data: "TL_superflow", "STARTING", "2013-03-06 21:52 PST", "0 hours 0 minutes", and "0".

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
TL_superflow	STARTING	2013-03-06 21:52 PST	0 hours 0 minutes	0

Starting an AWS Cluster

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



1 Job Flow selected

Job Flow: j-1ETJ7XCAQJUF8

Last State Change: Running bootstrap actions

Description Steps Bootstrap Actions Instance Groups Monitoring

Name:	TL_superflow	Creation Date:	2013-03-06 21:52 PST
Start Date:	2013-03-06 21:55 PST	End Date:	-
Availability Zone:	us-west-2b	Instance Count:	-
Master Instance Type:	-	Slave Instance Type:	-
Key Name:	tlehmann_keypair	Log URI:	-
Ami Version:	2.3.3	Master Public DNS Name:	ec2-54-244-172-225.us-west-2.compute.amazonaws.com
Hadoop Version:	1.0.3	Keep Alive:	true
Termination Protected:	false	Visible To All Users:	false
Subnet Id:	-	Supported Products:	-

Starting an AWS Cluster

- Windows users use PuTTY to connect to cluster
- Everybody else runs
`ssh -o "ServerAliveInterval 10" -i </path/to/saved/keypair/file.pem>
hadoop@<master.public-dns-name.amazonaws.com>`
from command line

Starting an AWS Cluster

- Type *pig*
- Result → grunt>
- Time to write some pig queries!



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

Additional Tasks

- Monitoring Hadoop jobs
 1. Using ssh tunneling
OR
 2. Using LYNX
OR
 3. Using SOCKS proxy

Additional Tasks

- Terminating Cluster
- Go to Management Console
- Select Job Flow
- Click *Terminate*
- Wait a couple minutes....
- Eventually status should be

 TERMINATED

Final Comment

- Start early!
- Running jobs will take several hours
- **GOOD LUCK!**