Amazon Web Services (AWS) Setup Guidelines

For CSE6242 HW3, updated version of the guidelines by Diana Maclean
[Estimated time needed: 1 hour]

Note that important steps are highlighted in yellow.

What we will try to accomplish with this?

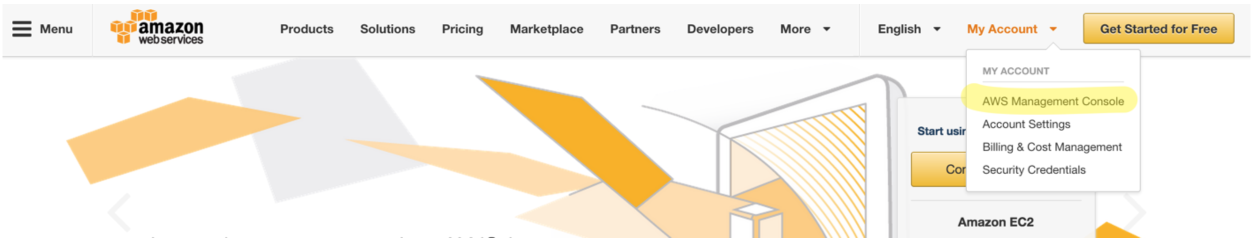
This guideline will help you get set up with the Amazon Web Services (AWS, a "cloud" platform) where you will run large-scale analysis on big data. Here are you will learn to do

1. Create an AWS account (to get access to EC2, Elastic MapReduce and S3 storage).
2. Create storage buckets on S3 (to save outputs and logs of MapReduce jobs).
3. Create a key pair (required for running MapReduce jobs on EC2).
4. Get Access Keys (also required for running jobs on EC2).
5. Redeem your free credit (worth $100).
6. Familiarize yourself with S3, EC2 and EMR (by doing a sample MapReduce run).

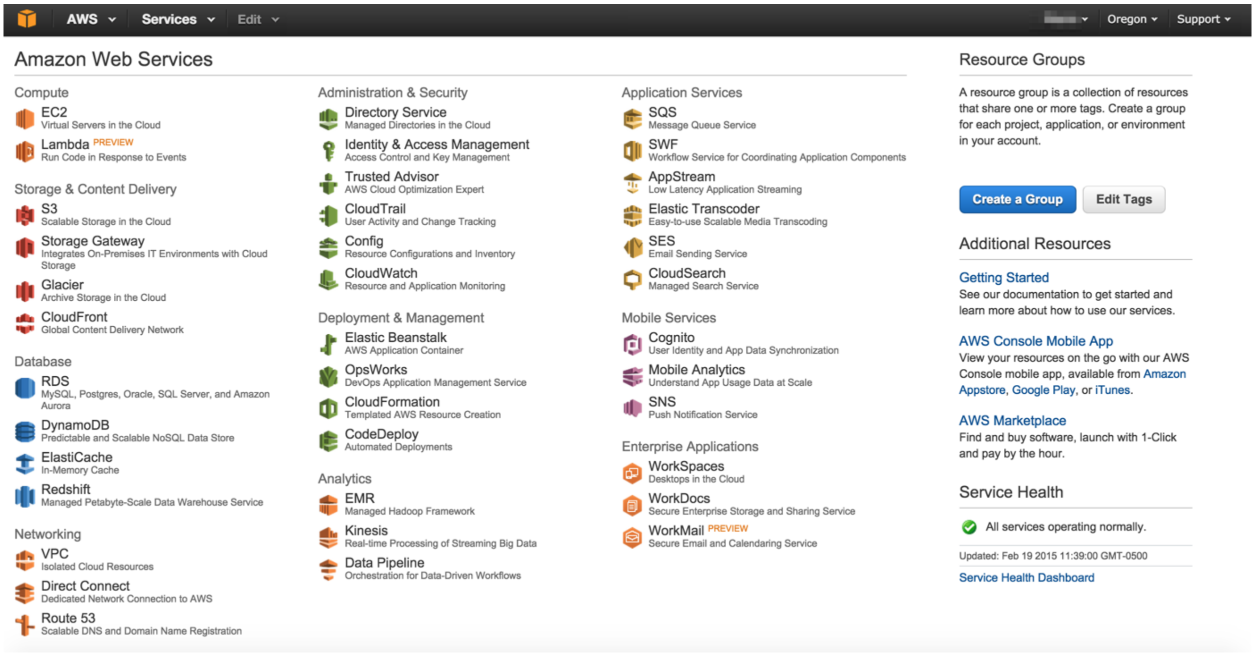
1. Create an AWS account

- Go to http://aws.amazon.com and sign up for an account, if you do not have one already.
- For now, please enter the required details, including payment details (you will need a valid credit card or debit card to sign up). Amazon has generously agreed to provide each student with credit for this class; more on how to redeem this later.
- Validate your account with the identity verification through your phone.

Once your account has been created and your payment method verified, you should have access to the AWS Management Console.



You AWS Management Console should look like this:



Amazon Web Services

Compute

EC2
Virtual Servers in the Cloud

Lambda PREVIEW
Run Code in Response to Events

Storage & Content Delivery

S3
Scalable Storage in the Cloud

Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage

Glacier
Archive Storage in the Cloud

CloudFront
Global Content Delivery Network

Database

RDS
MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora

DynamoDB
Predictable and Scalable NoSQL Data Store

ElastiCache
In-Memory Cache

Redshift
Managed Petabyte-Scale Data Warehouse Service

Networking

VPC
Isolated Cloud Resources

Direct Connect
Dedicated Network Connection to AWS

Route 53
Scalable DNS and Domain Name Registration

Administration & Security

Directory Service
Managed Directories in the Cloud

Identity & Access Management
Access Control and Key Management

Trusted Advisor
AWS Cloud Optimization Expert

CloudTrail
User Activity and Change Tracking

Config
Resource Configurations and Inventory

CloudWatch
Resource and Application Monitoring

Deployment & Management

Elastic Beanstalk
AWS Application Container

OpsWorks
DevOps Application Management Service

CloudFormation
Templated AWS Resource Creation

CodeDeploy
Automated Deployments

Analytics

EMR
Managed Hadoop Framework

Kinesis
Real-time Processing of Streaming Big Data

Data Pipeline
Orchestration for Data-Driven Workflows

Application Services

SQS
Message Queue Service

SWF
Workflow Service for Coordinating Application Components

AppStream
Low Latency Application Streaming

Elastic Transcoder
Easy-to-use Scalable Media Transcoding

SES
Email Sending Service

CloudSearch
Managed Search Service

Mobile Services

Cognito
User Identity and App Data Synchronization

Mobile Analytics
Understand App Usage Data at Scale

SNS
Push Notification Service

Enterprise Applications

WorkSpaces
Desktops in the Cloud

WorkDocs
Secure Enterprise Storage and Sharing Service

WorkMail PREVIEW
Secure Email and Calendaring Service

Resource Groups

A resource group is a collection of resources that share one or more tags. Create a group for each project, application, or environment in your account.

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

Service Health

✓ All services operating normally.

Updated: Feb 19 2015 11:39:00 GMT-0500

Service Health Dashboard

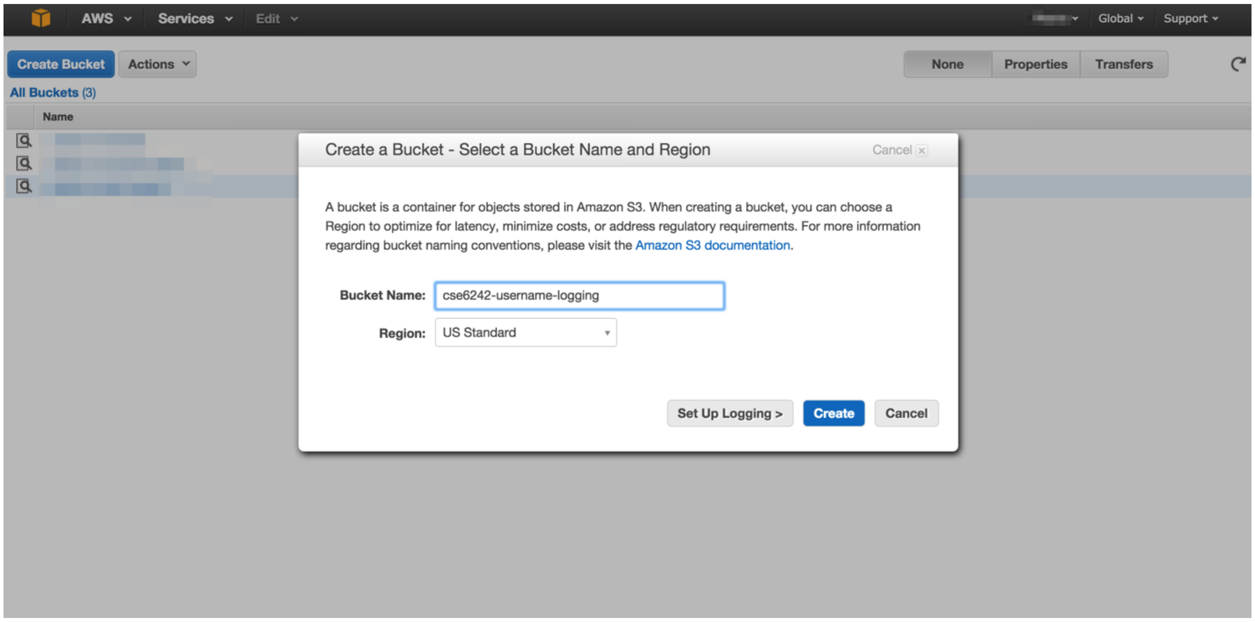
2. Create storage buckets on S3

In the AWS Management Console click on "S3" under **Storage & Content Delivery**. We need S3 for two reasons:
(1) an EMR workflow requires the input data to be on S3;
(2) EMR workflow output is always saved to S3. Data (or objects) in S3 are stored in what we call "buckets". You can think of buckets as folders.

For this assignment, we have put the data you will process in a public bucket called:
cse6242-spring2014-gtcse-data

You will see how to reference this for EMR input later on. In the meanwhile, you will need some buckets of your own to (1) store your EMR output , and (2) store your log files if you wish to debug your EMR runs.
Once you have signed up, we will begin by creating the log bucket first.

i. In the S3 console, click on "Create Bucket".



Create a Bucket - Select a Bucket Name and Region

Cancel

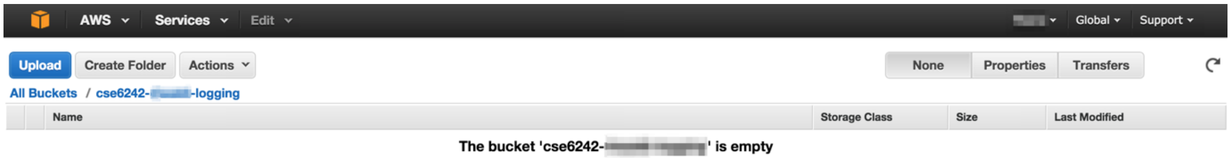
A bucket is a container for objects stored in Amazon S3. When creating a bucket, you can choose a Region to optimize for latency, minimize costs, or address regulatory requirements. For more information regarding bucket naming conventions, please visit the Amazon S3 documentation.

Bucket Name: cse6242-username-logging

Region: US Standard

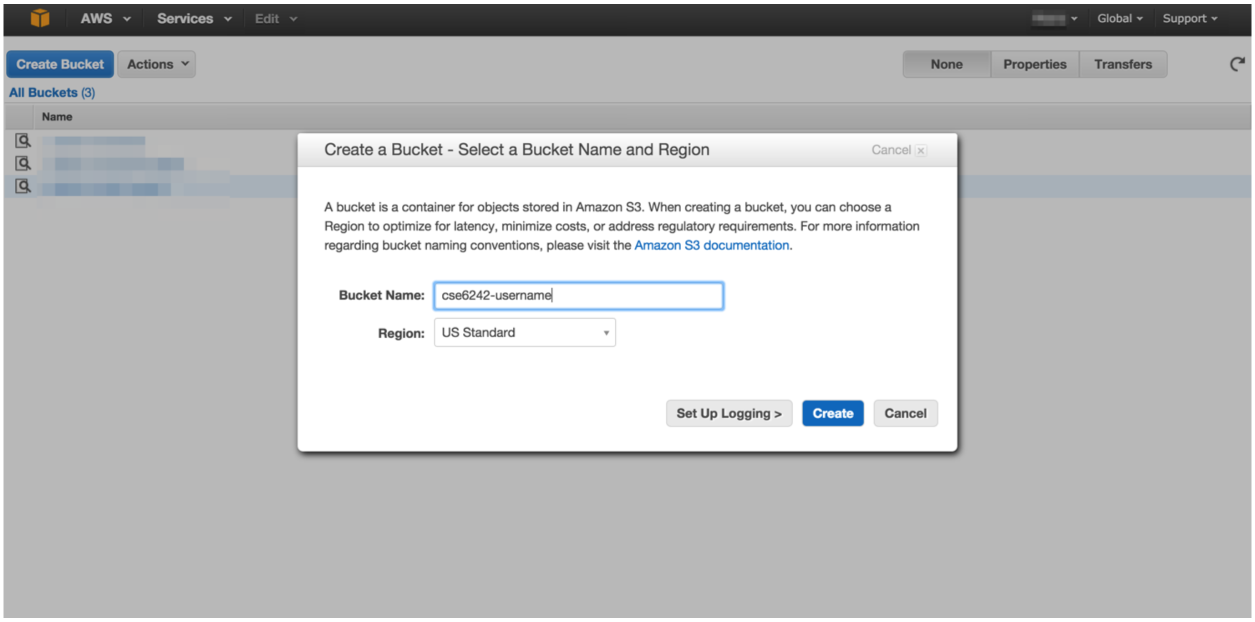
ii. All S3 buckets need to have unique names. You could name the logging bucket *cse6242-<gt-username>-logging*. **Important:** Please select "US Standard" in the Region dropdown. Click on "**Create**" (not on "Set Up Logging >>").

iii. Your new bucket will appear in the S3 console. Clicking on it will show you that it is empty.



The bucket 'cse6242-[redacted]' is empty

iv. Now we will create our main bucket. Go back to the main screen (clicking on "All Buckets"). Again, click on "Create Bucket". Call this one *cse6242-<gt-username>*. Again, pick "US Standard" for the Region dropdown. Since we will link this bucket to our logging bucket, the regions for the two buckets should be the same. We will link our logging bucket to the one we are creating now, so click on "Set Up Logging >".



Create a Bucket - Select a Bucket Name and Region

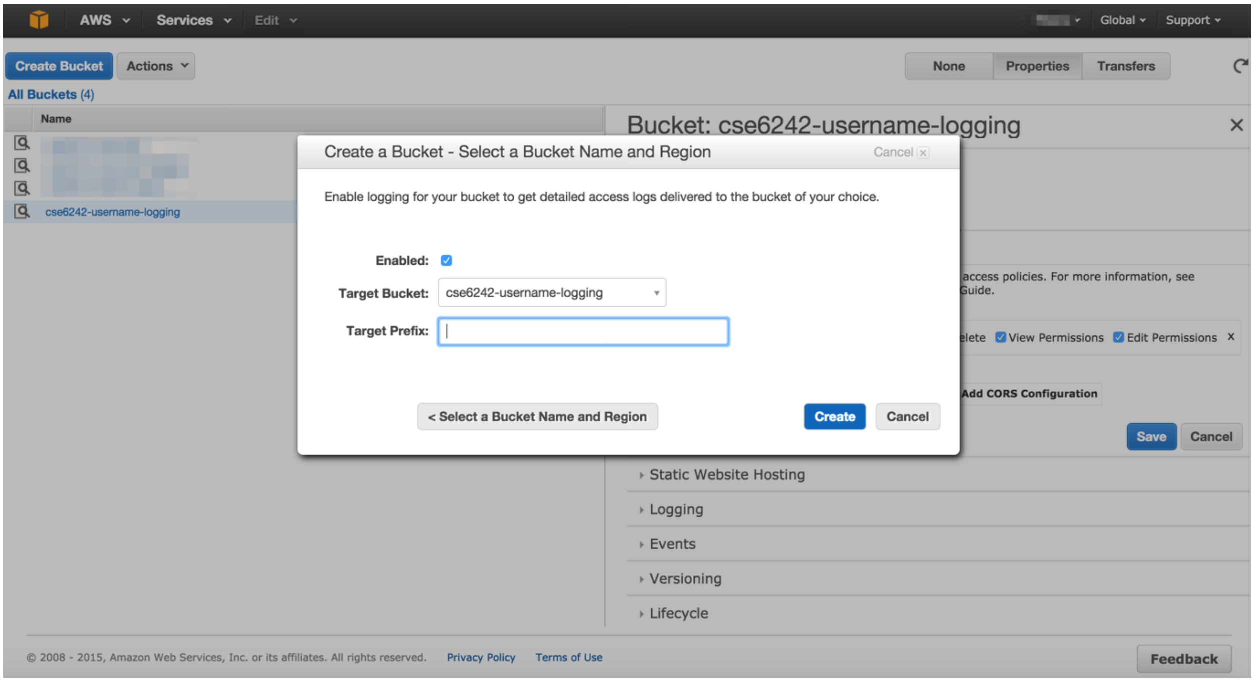
Cancel ☒

A bucket is a container for objects stored in Amazon S3. When creating a bucket, you can choose a Region to optimize for latency, minimize costs, or address regulatory requirements. For more information regarding bucket naming conventions, please visit the Amazon S3 documentation.

Bucket Name: cse6242-username

Region: US Standard

v. Click on "Enabled" to enable logging, and start typing in the name of your logging bucket. It should appear in the drop down menu, select it. Clear the "Target Prefix" field
and click "Create".



Bucket: cse6242-username-logging

Create a Bucket - Select a Bucket Name and Region

Cancel ☒

Enable logging for your bucket to get detailed access logs delivered to the bucket of your choice.

Enabled: ☑

Target Bucket: cse6242-username-logging

Target Prefix:

access policies. For more information, see Guide.

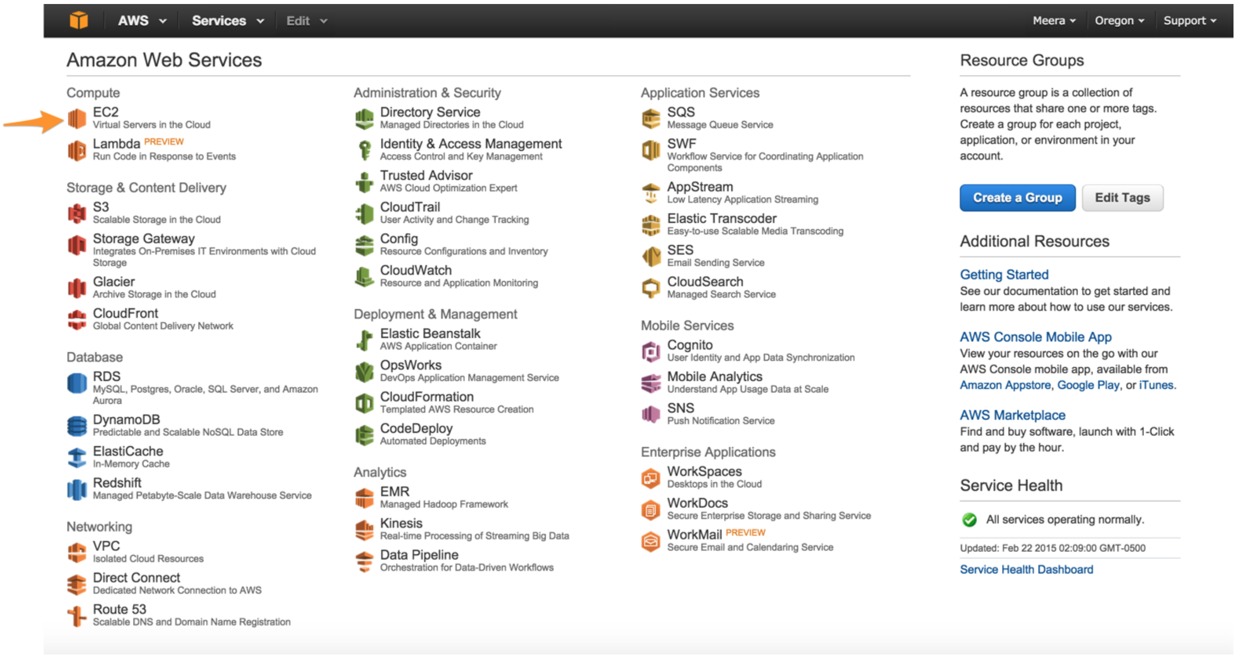
elete ☑ View Permissions ☑ Edit Permissions ×

© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

We are done creating buckets at this point.

3. Create a key pair

When you run jobs on EMR, you will need to have a valid public/private key pair. To create your first key pair, click on “EC2” under **Compute** in the AWS Management Console.



Amazon Web Services

Compute

EC2 Virtual Servers in the Cloud
Lambda PREVIEW Run Code in Response to Events

Storage & Content Delivery

S3 Scalable Storage in the Cloud
Storage Gateway Integrates On-Premises IT Environments with Cloud Storage
Glacier Archive Storage in the Cloud
CloudFront Global Content Delivery Network

Database

RDS MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora
DynamoDB Predictable and Scalable NoSQL Data Store
ElastiCache In-Memory Cache
Redshift Managed Petabyte-Scale Data Warehouse Service

Networking

VPC Isolated Cloud Resources
Direct Connect Dedicated Network Connection to AWS
Route 53 Scalable DNS and Domain Name Registration

Administration & Security

Directory Service Managed Directories in the Cloud
Identity & Access Management Access Control and Key Management
Trusted Advisor AWS Cloud Optimization Expert
CloudTrail User Activity and Change Tracking
Config Resource Configurations and Inventory
CloudWatch Resource and Application Monitoring

Deployment & Management

Elastic Beanstalk AWS Application Container
OpsWorks DevOps Application Management Service
CloudFormation Templated AWS Resource Creation
CodeDeploy Automated Deployments

Analytics

EMR Managed Hadoop Framework
Kinesis Real-time Processing of Streaming Big Data
Data Pipeline Orchestration for Data-Driven Workflows

Application Services

SQS Message Queue Service
SWF Workflow Service for Coordinating Application Components
AppStream Low Latency Application Streaming
Elastic Transcoder Easy-to-use Scalable Media Transcoding
SES Email Sending Service
CloudSearch Managed Search Service

Mobile Services

Cognito User Identity and App Data Synchronization
Mobile Analytics Understand App Usage Data at Scale
SNS Push Notification Service

Enterprise Applications

WorkSpaces Desktops in the Cloud
WorkDocs Secure Enterprise Storage and Sharing Service
WorkMail PREVIEW Secure Email and Calendaring Service

Resource Groups

A resource group is a collection of resources that share one or more tags. Create a group for each project, application, or environment in your account.

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

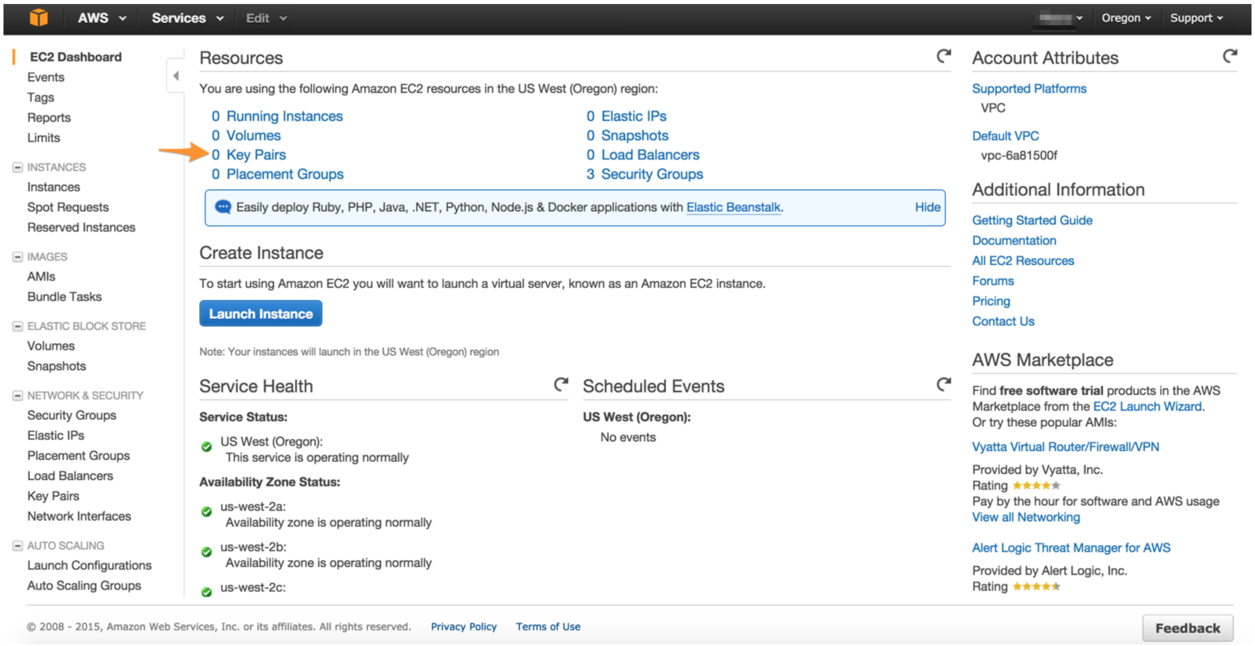
Service Health

All services operating normally.

Updated: Feb 22 2015 02:09:00 GMT-0500

Service Health Dashboard

Select the region on the top right as U.S. West(Oregon) since the data bucket is stored in this region. You should see a link stating “0 Key Pairs” under **Resources**. Click on this.



You are using the following Amazon EC2 resources in the US West (Oregon) region:

Easily deploy Ruby, PHP, Java, .NET, Python, Node.js & Docker applications with Elastic Beanstalk.

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Note: Your instances will launch in the US West (Oregon) region

Service Status:

US West (Oregon):
This service is operating normally

Availability Zone Status:

us-west-2a:
Availability zone is operating normally

us-west-2b:
Availability zone is operating normally

us-west-2c:

US West (Oregon):

No events

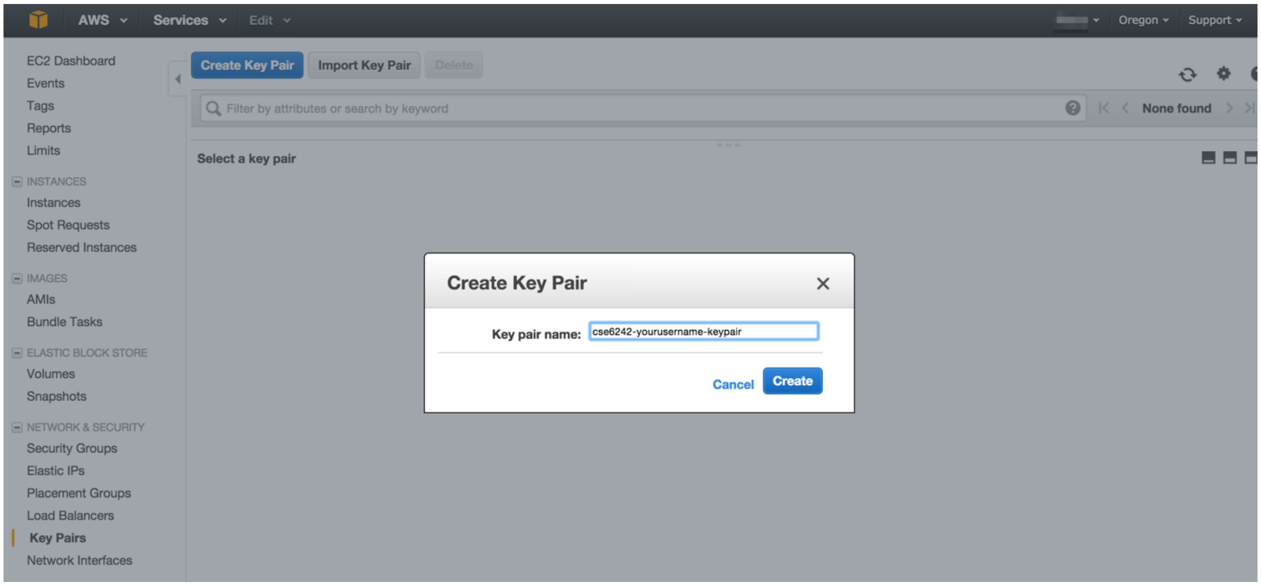
Find **free software trial** products in the AWS Marketplace from the EC2 Launch Wizard.
Or try these popular AMIs:

Provided by Vyatta, Inc.
Rating ★★★★★
Pay by the hour for software and AWS usage

Provided by Alert Logic, Inc.
Rating ★★★★★

© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

You will be given an option to "Create Key Pair". Name your key pair as you wish. Upon providing a name and clicking on "Create", your private key (a .pem file) will automatically download. **Save it in a safe place where you will be able to find it again (IMPORTANT, do not lose this file)**.



Key pair name: cse6242-yourusername-keypair

If you need to access your public key, you will be able to find it in the same place where you found your account credentials. Amazon keeps no record of your private key, and if you lose it, you will need to generate a new set.

If your computer runs **Windows**, use the steps in the following link to convert your .pem file to a .ppk file for use with PuTTY.
Read the section titled Converting Your Private Key Using PuTTYgen in the link below:
http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html

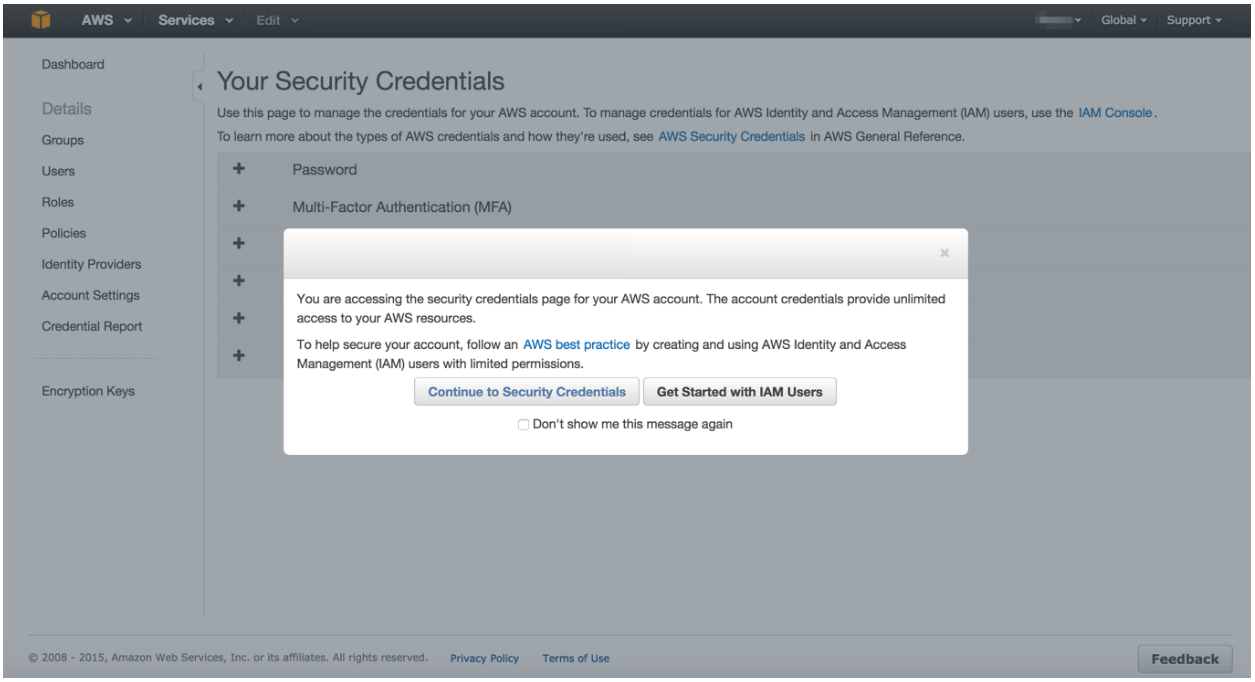
Note: If you use the AWS Management Console, you would typically not be required to access your private key . However, you will be asked to name your access key pair and the private key each time you run an EMR job.

If you wish to log into the master node running your MapReduce job, you will need your `.pem` file (you will need this in case you wish to run an interactive HIVE/PIG job flow). To log on to the master node (you can find the address of the master node from the MapReduce dashboard), you will need to do the following:

```
  $ ssh hadoop@<master-node-address> -i <path-to-pem-file>/<pem-file-name>.pem
```

4. Get Access Keys (new site)

Click on “Security Credentials” under your username (top right). Click on “Continue …”



Your Security Credentials

Use this page to manage the credentials for your AWS account. To manage credentials for AWS Identity and Access Management (IAM) users, use the IAM Console.

To learn more about the types of AWS credentials and how they're used, see AWS Security Credentials in AWS General Reference.

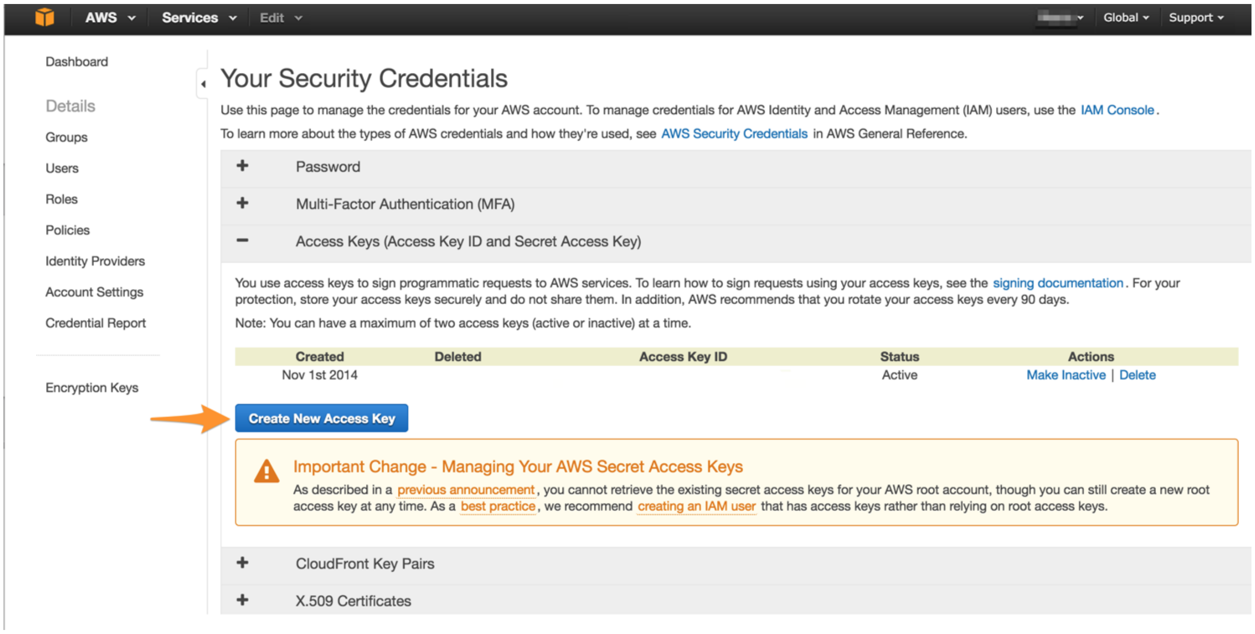
You are accessing the security credentials page for your AWS account. The account credentials provide unlimited access to your AWS resources.

To help secure your account, follow an AWS best practice by creating and using AWS Identity and Access Management (IAM) users with limited permissions.

□ Don't show me this message again

© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Click on the **Create a new Access Key** link (under Access Keys), and download the Access Key file (**do not lose this file**). Now you are ready to run a MapReduce job.



Your Security Credentials

Use this page to manage the credentials for your AWS account. To manage credentials for AWS Identity and Access Management (IAM) users, use the IAM Console.

To learn more about the types of AWS credentials and how they're used, see AWS Security Credentials in AWS General Reference.

You use access keys to sign programmatic requests to AWS services. To learn how to sign requests using your access keys, see the signing documentation. For your protection, store your access keys securely and do not share them. In addition, AWS recommends that you rotate your access keys every 90 days.

Note: You can have a maximum of two access keys (active or inactive) at a time.

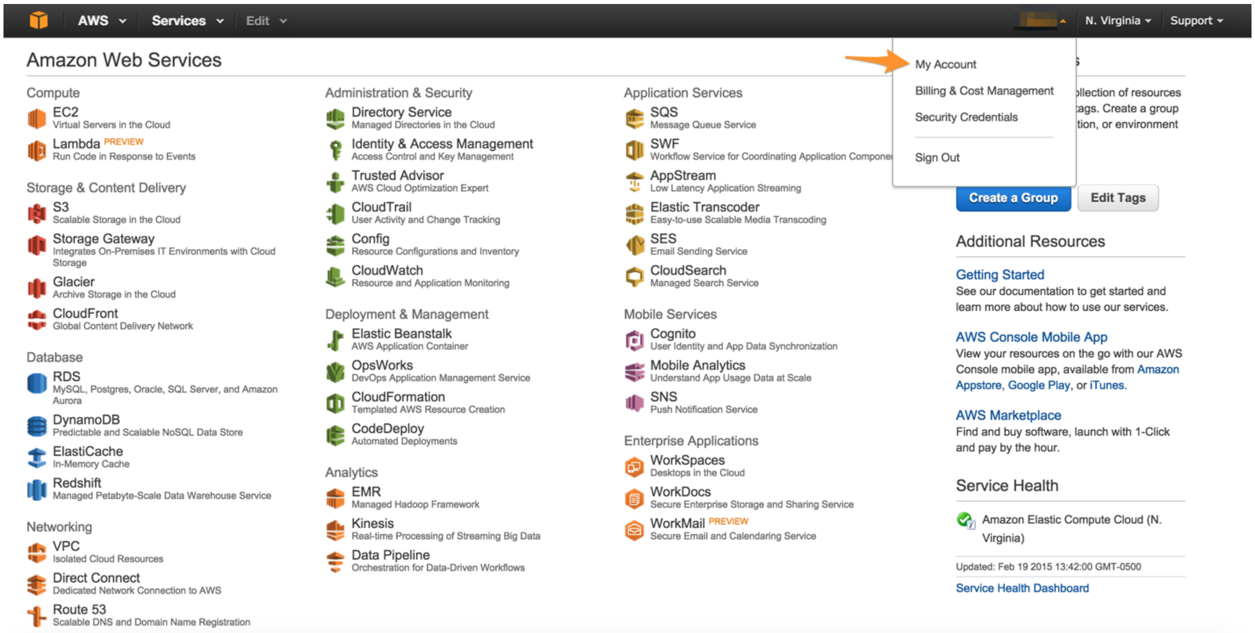
Created	Deleted	Access Key ID	Status	Actions
Nov 1st 2014			Active	Make Inactive | Delete

Important Change - Managing Your AWS Secret Access Keys

As described in a previous announcement, you cannot retrieve the existing secret access keys for your AWS root account, though you can still create a new root access key at any time. As a best practice, we recommend creating an IAM user that has access keys rather than relying on root access keys.

5. Redeem your free credit

In order to add the credit to your account, you will need your unique Credit Coupon Code. If you have not received this yet, please write a private post on Piazza. Once you have your code, go to your account page (http://aws.amazon.com/account)


AWS
Services
Edit
[redacted]
N. Virginia
Support
Amazon Web Services
Compute
EC2
Virtual Servers in the Cloud
Lambda PREVIEW
Run Code in Response to Events
Storage & Content Delivery
S3
Scalable Storage in the Cloud
Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage
Glacier
Archive Storage in the Cloud
CloudFront
Global Content Delivery Network
Database
RDS
MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora
DynamoDB
Predictable and Scalable NoSQL Data Store
ElastiCache
In-Memory Cache
Redshift
Managed Petabyte-Scale Data Warehouse Service
Networking
VPC
Isolated Cloud Resources
Direct Connect
Dedicated Network Connection to AWS
Route 53
Scalable DNS and Domain Name Registration
Administration & Security
Directory Service
Managed Directories in the Cloud
Identity & Access Management
Access Control and Key Management
Trusted Advisor
AWS Cloud Optimization Expert
CloudTrail
User Activity and Change Tracking
Config
Resource Configurations and Inventory
CloudWatch
Resource and Application Monitoring
Deployment & Management
Elastic Beanstalk
AWS Application Container
OpsWorks
DevOps Application Management Service
CloudFormation
Templated AWS Resource Creation
CodeDeploy
Automated Deployments
Analytics
EMR
Managed Hadoop Framework
Kinesis
Real-time Processing of Streaming Big Data
Data Pipeline
Orchestration for Data-Driven Workflows
Application Services
SQS
Message Queue Service
SWF
Workflow Service for Coordinating Application Compone
AppStream
Low Latency Application Streaming
Elastic Transcoder
Easy-to-use Scalable Media Transcoding
SES
Email Sending Service
CloudSearch
Managed Search Service
Mobile Services
Cognito
User Identity and App Data Synchronization
Mobile Analytics
Understand App Usage Data at Scale
SNS
Push Notification Service
Enterprise Applications
WorkSpaces
Desktops in the Cloud
WorkDocs
Secure Enterprise Storage and Sharing Service
WorkMail PREVIEW
Secure Email and Calendaring Service
My Account
Billing & Cost Management
Security Credentials
Sign Out
Create a Group
Edit Tags

llection of resources ags. Create a group tion, or environment

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

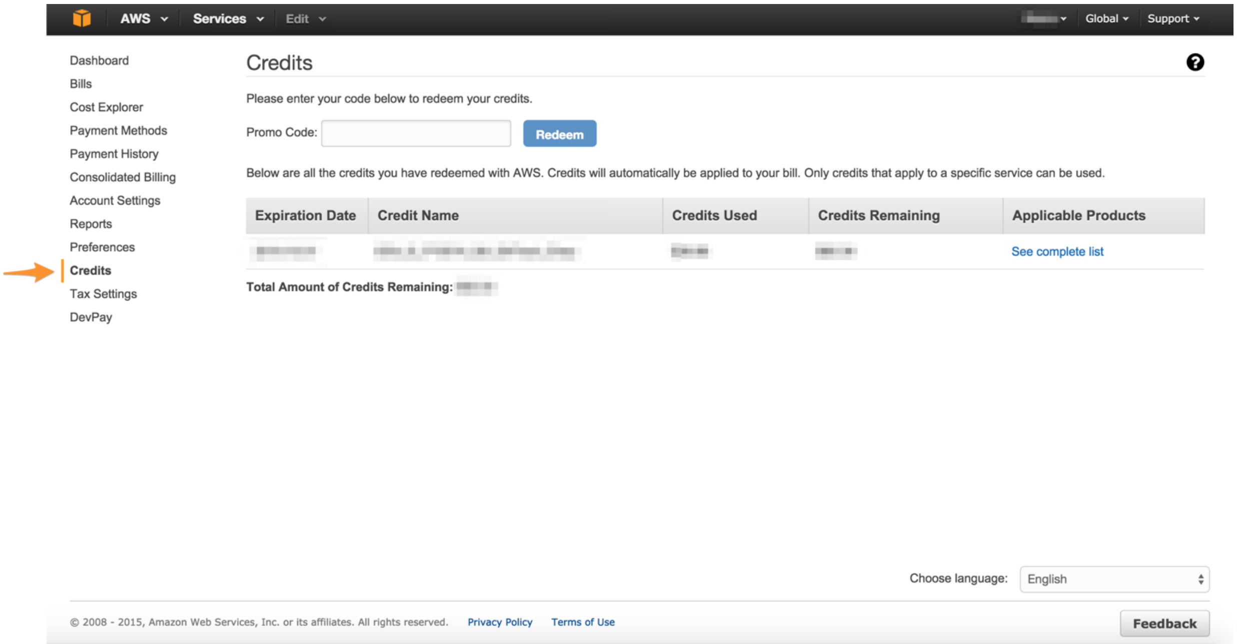
Service Health

Amazon Elastic Compute Cloud (N. Virginia)

Updated: Feb 19 2015 13:42:00 GMT-0500

Service Health Dashboard

Click on "Credits". Enter the Code into the Promo Code text box, and click Redeem.



Credits

Please enter your code below to redeem your credits.

Promo Code:

Below are all the credits you have redeemed with AWS. Credits will automatically be applied to your bill. Only credits that apply to a specific service can be used.

Expiration Date	Credit Name	Credits Used	Credits Remaining	Applicable Products
[redacted]	[redacted]	[redacted]	[redacted]	See complete list

Total Amount of Credits Remaining: [redacted]

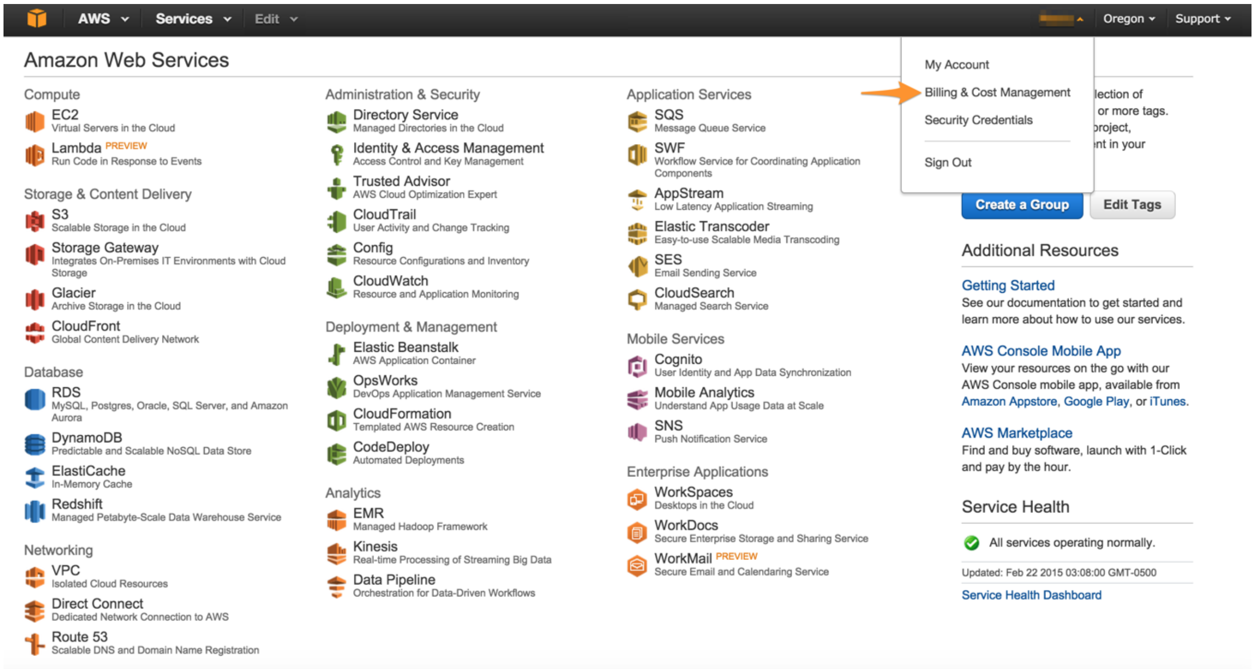
Please email the CSE6242 instructors immediately if this does not work. Unfortunately, we can only give you so much free credit, so don't go too wild! You can check on how much credit you have left by clicking on the "Account Activity" link from your account page or by returning to this page. Sometimes this can take a while to update, so don't be surprised if recent changes are not immediately apparent. We will set up a monitor in the next step which is triggered when you utilize half of the credit.

6. Set up a CloudWatch Usage Alert

Make sure your region (in the upper right corner of the screen) is set to: US West (Oregon).

Now we will turn on alerts.

1. Go to the "Billing and Cost Management" page.



Amazon Web Services

Compute

EC2 Virtual Servers in the Cloud

Lambda PREVIEW Run Code in Response to Events

Storage & Content Delivery

S3 Scalable Storage in the Cloud

Storage Gateway Integrates On-Premises IT Environments with Cloud Storage

Glacier Archive Storage in the Cloud

CloudFront Global Content Delivery Network

Database

RDS MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora

DynamoDB Predictable and Scalable NoSQL Data Store

ElastiCache In-Memory Cache

Redshift Managed Petabyte-Scale Data Warehouse Service

Networking

VPC Isolated Cloud Resources

Direct Connect Dedicated Network Connection to AWS

Route 53 Scalable DNS and Domain Name Registration

Administration & Security

Directory Service Managed Directories in the Cloud

Identity & Access Management Access Control and Key Management

Trusted Advisor AWS Cloud Optimization Expert

CloudTrail User Activity and Change Tracking

Config Resource Configurations and Inventory

CloudWatch Resource and Application Monitoring

Deployment & Management

Elastic Beanstalk AWS Application Container

OpsWorks DevOps Application Management Service

CloudFormation Templated AWS Resource Creation

CodeDeploy Automated Deployments

Analytics

EMR Managed Hadoop Framework

Kinesis Real-time Processing of Streaming Big Data

Data Pipeline Orchestration for Data-Driven Workflows

Application Services

SQS Message Queue Service

SWF Workflow Service for Coordinating Application Components

AppStream Low Latency Application Streaming

Elastic Transcoder Easy-to-use Scalable Media Transcoding

SES Email Sending Service

CloudSearch Managed Search Service

Mobile Services

Cognito User Identity and App Data Synchronization

Mobile Analytics Understand App Usage Data at Scale

SNS Push Notification Service

Enterprise Applications

WorkSpaces Desktops in the Cloud

WorkDocs Secure Enterprise Storage and Sharing Service

WorkMail PREVIEW Secure Email and Calendaring Service

lection of
or more tags.
project,
nt in your

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

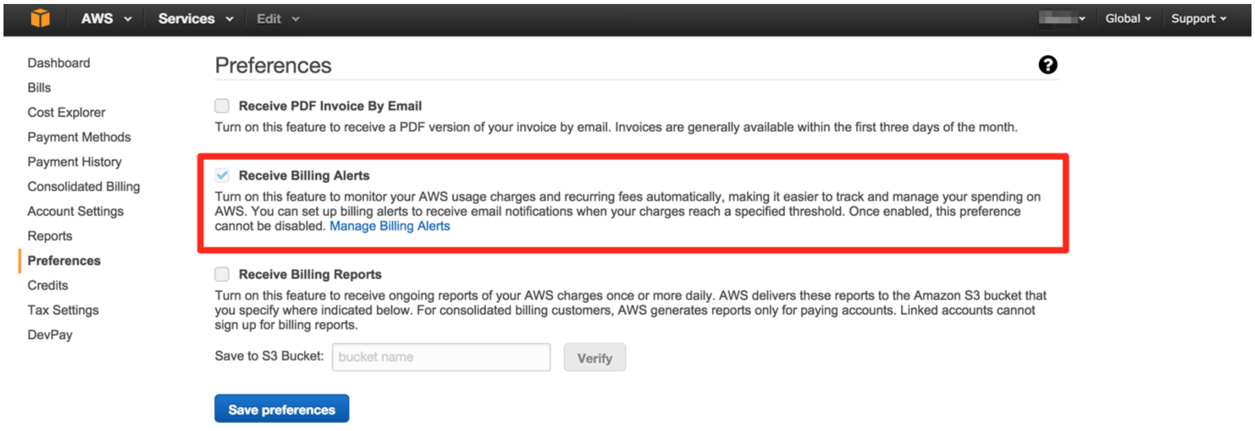
Service Health

All services operating normally.

Updated: Feb 22 2015 03:08:00 GMT-0500

Service Health Dashboard

2. Log In using your AWS credentials if necessary
3. Under **Preferences**, check the box labeled **Receive Billing Alerts**



Preferences

☐ **Receive PDF Invoice By Email**

Turn on this feature to receive a PDF version of your invoice by email. Invoices are generally available within the first three days of the month.

☑ **Receive Billing Alerts**

Turn on this feature to monitor your AWS usage charges and recurring fees automatically, making it easier to track and manage your spending on AWS. You can set up billing alerts to receive email notifications when your charges reach a specified threshold. Once enabled, this preference cannot be disabled. Manage Billing Alerts

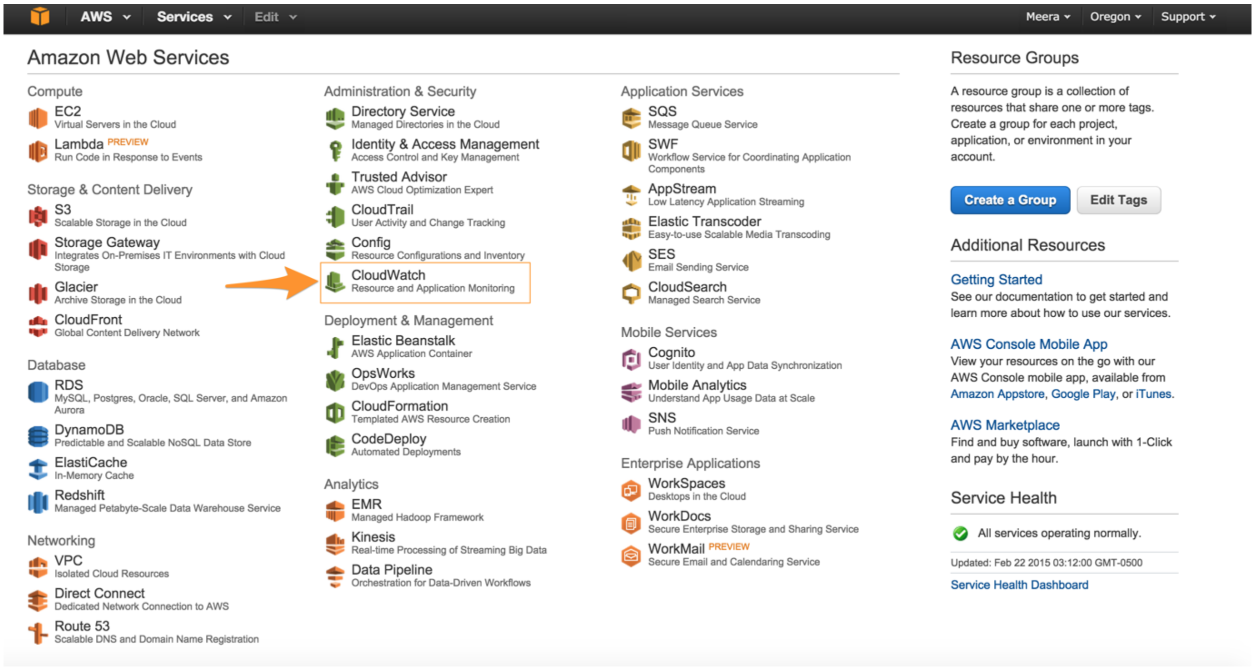
☐ **Receive Billing Reports**

Turn on this feature to receive ongoing reports of your AWS charges once or more daily. AWS delivers these reports to the Amazon S3 bucket that you specify where indicated below. For consolidated billing customers, AWS generates reports only for paying accounts. Linked accounts cannot sign up for billing reports.

Save to S3 Bucket: bucket name

Now we need to create a custom alarm so that it tells you when you have spent money.

1. Open the Amazon CloudWatch console. Click CloudWatch in the AWS Management Console.



AWS | Services | Edit | Meera | Oregon | Support

Amazon Web Services

Compute

EC2 – Virtual Servers in the Cloud

Lambda PREVIEW – Run Code in Response to Events

Storage & Content Delivery

S3 – Scalable Storage in the Cloud

Storage Gateway – Integrates On-Premises IT Environments with Cloud Storage

Glacier – Archive Storage in the Cloud

CloudFront – Global Content Delivery Network

Database

RDS – MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora

DynamoDB – Predictable and Scalable NoSQL Data Store

ElastiCache – In-Memory Cache

Redshift – Managed Petabyte-Scale Data Warehouse Service

Networking

VPC – Isolated Cloud Resources

Direct Connect – Dedicated Network Connection to AWS

Route 53 – Scalable DNS and Domain Name Registration

Administration & Security

Directory Service – Managed Directories in the Cloud

Identity & Access Management – Access Control and Key Management

Trusted Advisor – AWS Cloud Optimization Expert

CloudTrail – User Activity and Change Tracking

Config – Resource Configurations and Inventory

CloudWatch – Resource and Application Monitoring

Deployment & Management

Elastic Beanstalk – AWS Application Container

OpsWorks – DevOps Application Management Service

CloudFormation – Templated AWS Resource Creation

CodeDeploy – Automated Deployments

Analytics

EMR – Managed Hadoop Framework

Kinesis – Real-time Processing of Streaming Big Data

Data Pipeline – Orchestration for Data-Driven Workflows

Application Services

SQS – Message Queue Service

SWF – Workflow Service for Coordinating Application Components

AppStream – Low Latency Application Streaming

Elastic Transcoder – Easy-to-use Scalable Media Transcoding

SES – Email Sending Service

CloudSearch – Managed Search Service

Mobile Services

Cognito – User Identity and App Data Synchronization

Mobile Analytics – Understand App Usage Data at Scale

SNS – Push Notification Service

Enterprise Applications

WorkSpaces – Desktops in the Cloud

WorkDocs – Secure Enterprise Storage and Sharing Service

WorkMail PREVIEW – Secure Email and Calendaring Service

Resource Groups

A resource group is a collection of resources that share one or more tags. Create a group for each project, application, or environment in your account.

Create a Group | Edit Tags

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

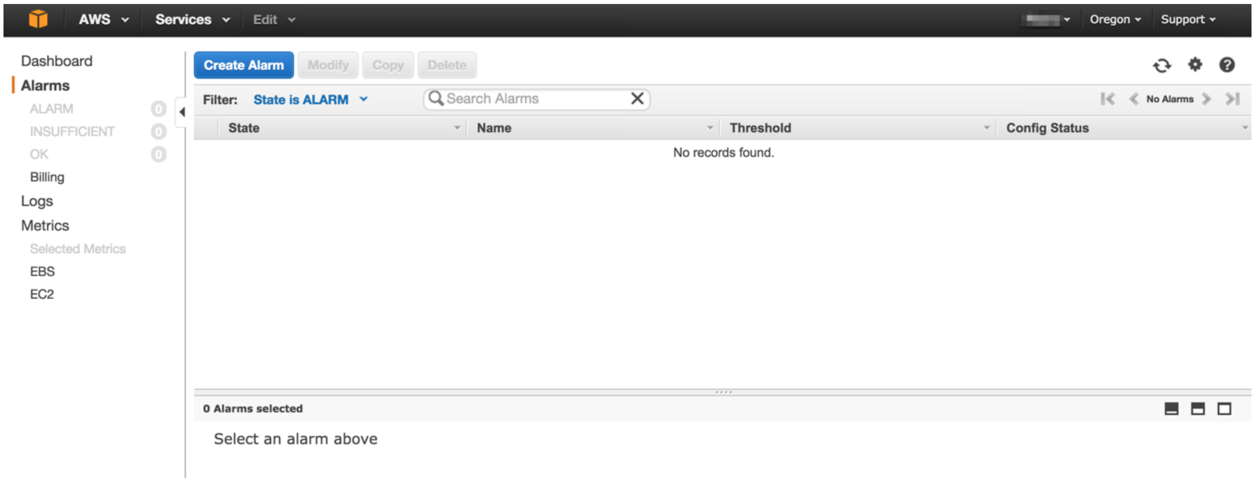
Service Health

✓ All services operating normally.

Updated: Feb 22 2015 03:12:00 GMT-0500

Service Health Dashboard

2. In the navigation pane on the left, click **Alarms**, and then in the **Alarms** pane, click **Create Alarm**.



AWS | Services | Edit | [redacted] | Oregon | Support

Dashboard
Alarms
ALARM 0
INSUFFICIENT 0
OK 0
Billing
Logs
Metrics
Selected Metrics
EBS
EC2

Create Alarm | Modify | Copy | Delete

Filter: State is ALARM | Search Alarms | No Alarms

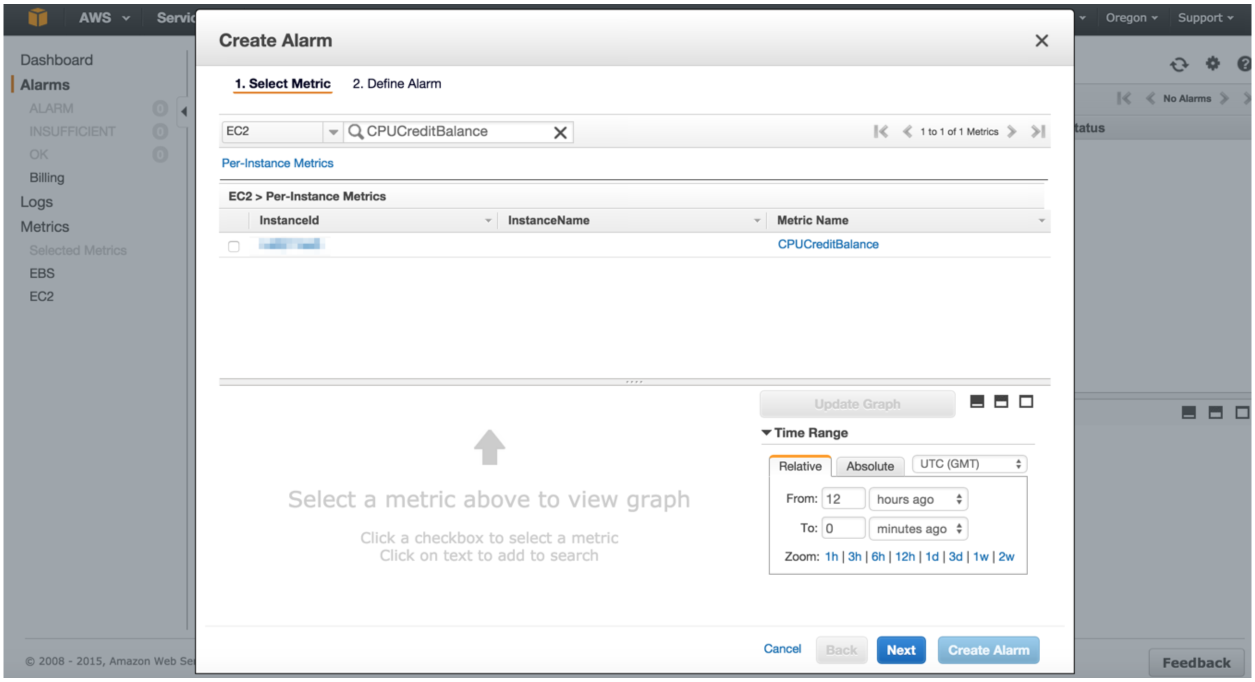
State	Name	Threshold	Config Status

No records found.

0 Alarms selected

Select an alarm above

3. In the **CloudWatch Metrics by Category** pane, under **Select Metrics**, select **EC2,** followed by the **CPUCreditBalance metric.**



EC2 | CPUCreditBalance

1 to 1 of 1 Metrics

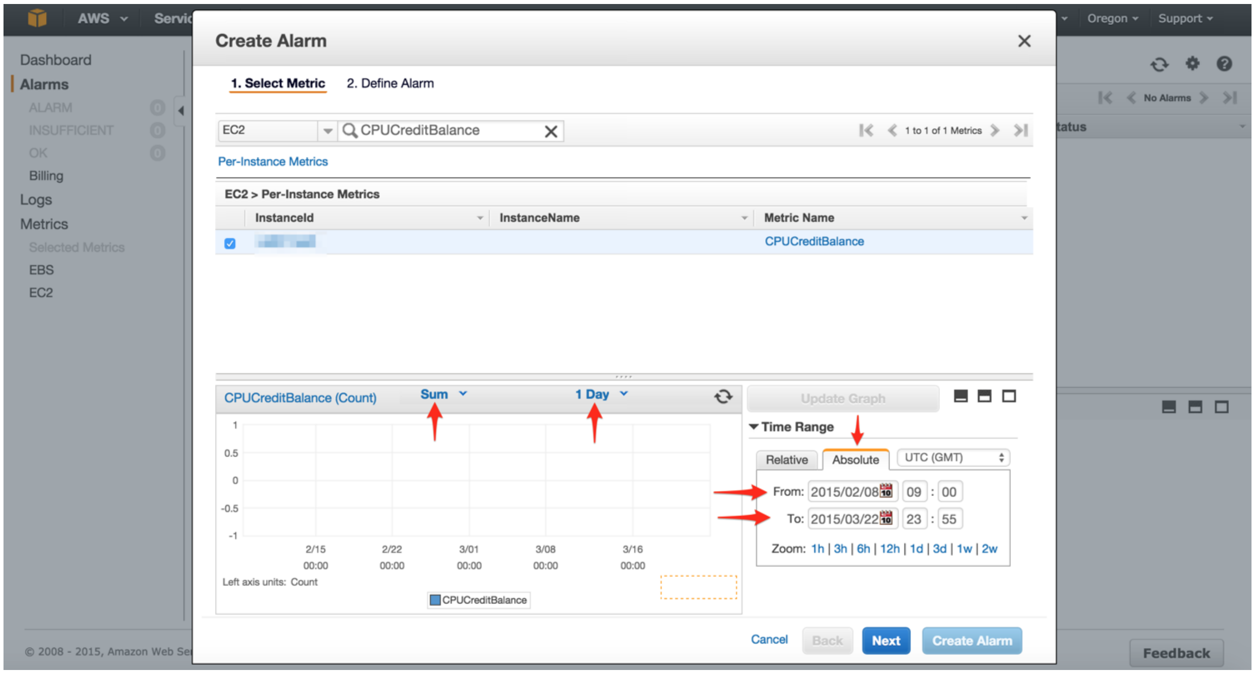
☐ CPUCreditBalance

UTC (GMT)

From: 12 hours ago

To: 0 minutes ago

4. Select USD then click the Absolute tab in the lower right. Enter the date dates below, select Sum, and “1 Day”. Then click next.



EC2 | CPUCreditBalance

1 to 1 of 1 Metrics

☑ CPUCreditBalance

Sum

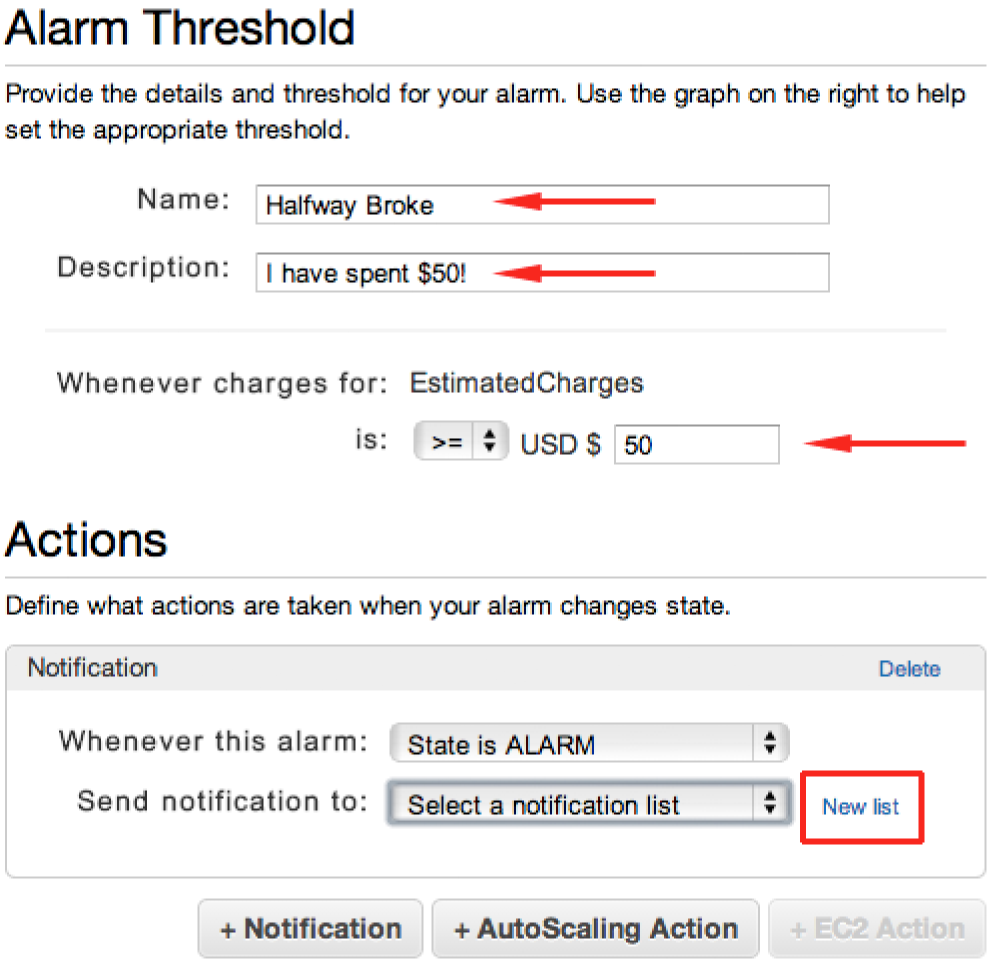
1 Day

UTC (GMT)

From: 2015/02/08 09 : 00

To: 2015/03/22 23 : 55

5. Fill out the alarm details and click **New List** next to “Send notification to:”:



Alarm Threshold

Provide the details and threshold for your alarm. Use the graph on the right to help set the appropriate threshold.

Name: Halfway Broke

Description: I have spent $50!

Whenever charges for: EstimatedCharges

is: >= USD $ 50

Actions

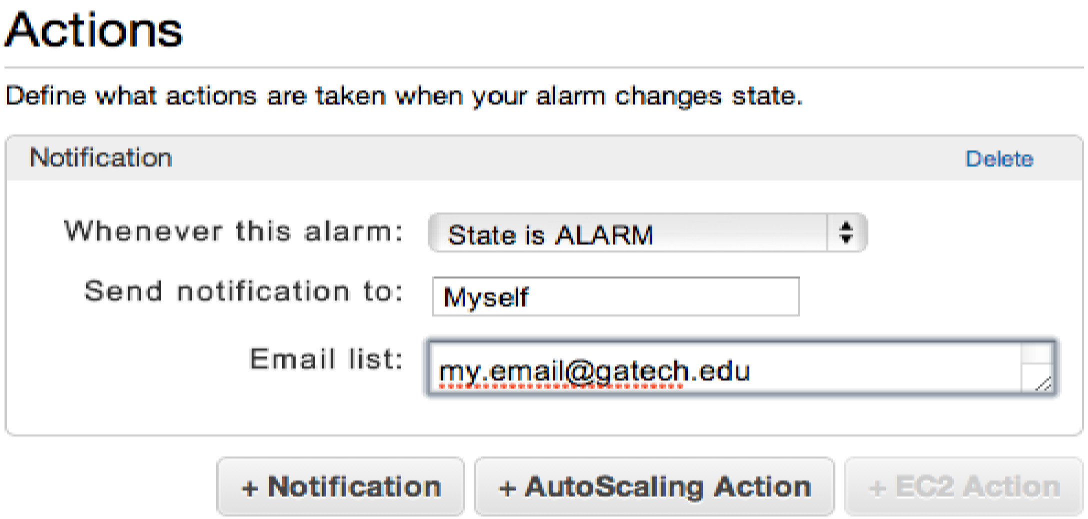
Define what actions are taken when your alarm changes state.

Notification

Whenever this alarm: State is ALARM

Send notification to: Select a notification list

Enter your name and email.



Actions

Define what actions are taken when your alarm changes state.

Notification

Whenever this alarm: State is ALARM

Send notification to: Myself

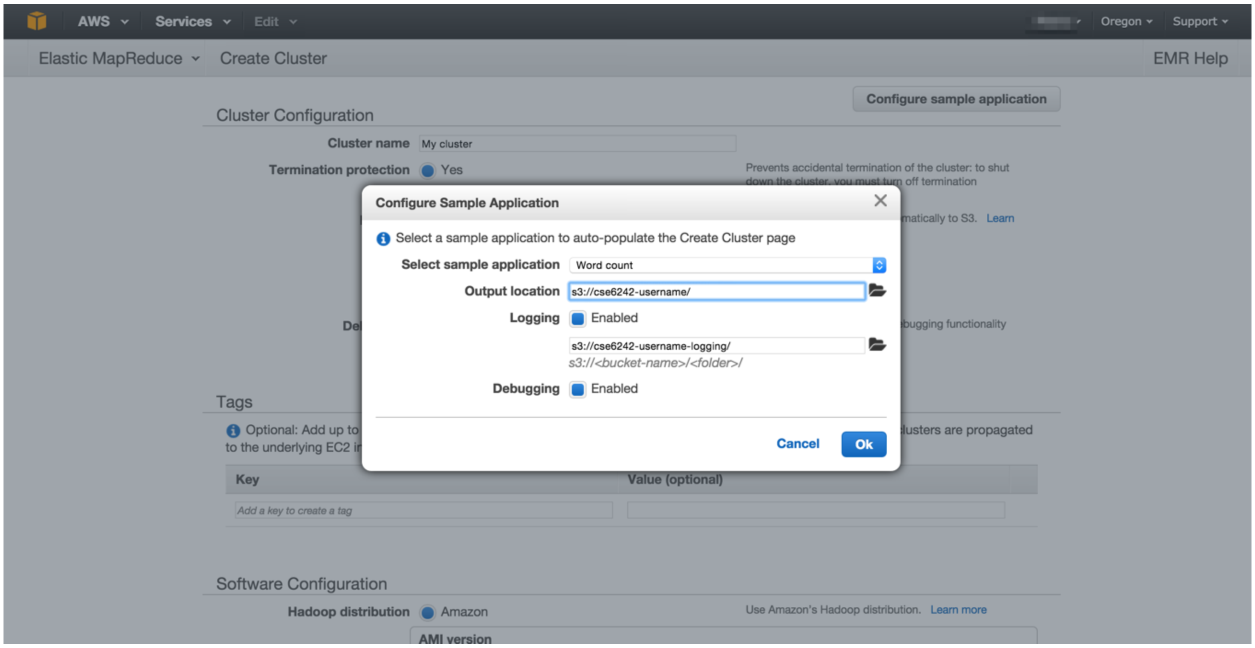
Email list: my.email@gatech.edu

You have now created an alert that will bother you when you pass $50. Consider making another alert which is activated when you use up $90 so that you do not get charged!

7. Familiarize yourself with S3, EC2 and EMR

We will now attempt to run a sample application of word count that comes with AWS. We will begin by clicking on the Elastic MapReduce(**EMR**) link in the Analytics section of the AWS Management Console. This will take you to the EMR Job Flows page. Click on the **“Create Cluster”** link.

Click on “**Configure sample application**”. Choose application “**Word count**”, then choose the output and logging bucket as the buckets you created previously. Click “OK” to configure the sample application.



Cluster Configuration

Cluster name: My cluster

Termination protection: ● Yes — Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination

Configure Sample Application

Select a sample application to auto-populate the Create Cluster page

Select sample application: Word count

Output location: s3://cse6242-username/

Logging: ☑ Enabled

s3://cse6242-username-logging/

s3://<bucket-name>/<folder>/

Debugging: ☑ Enabled

Tags

Optional: Add up to … to the underlying EC2 i… …lusters are propagated

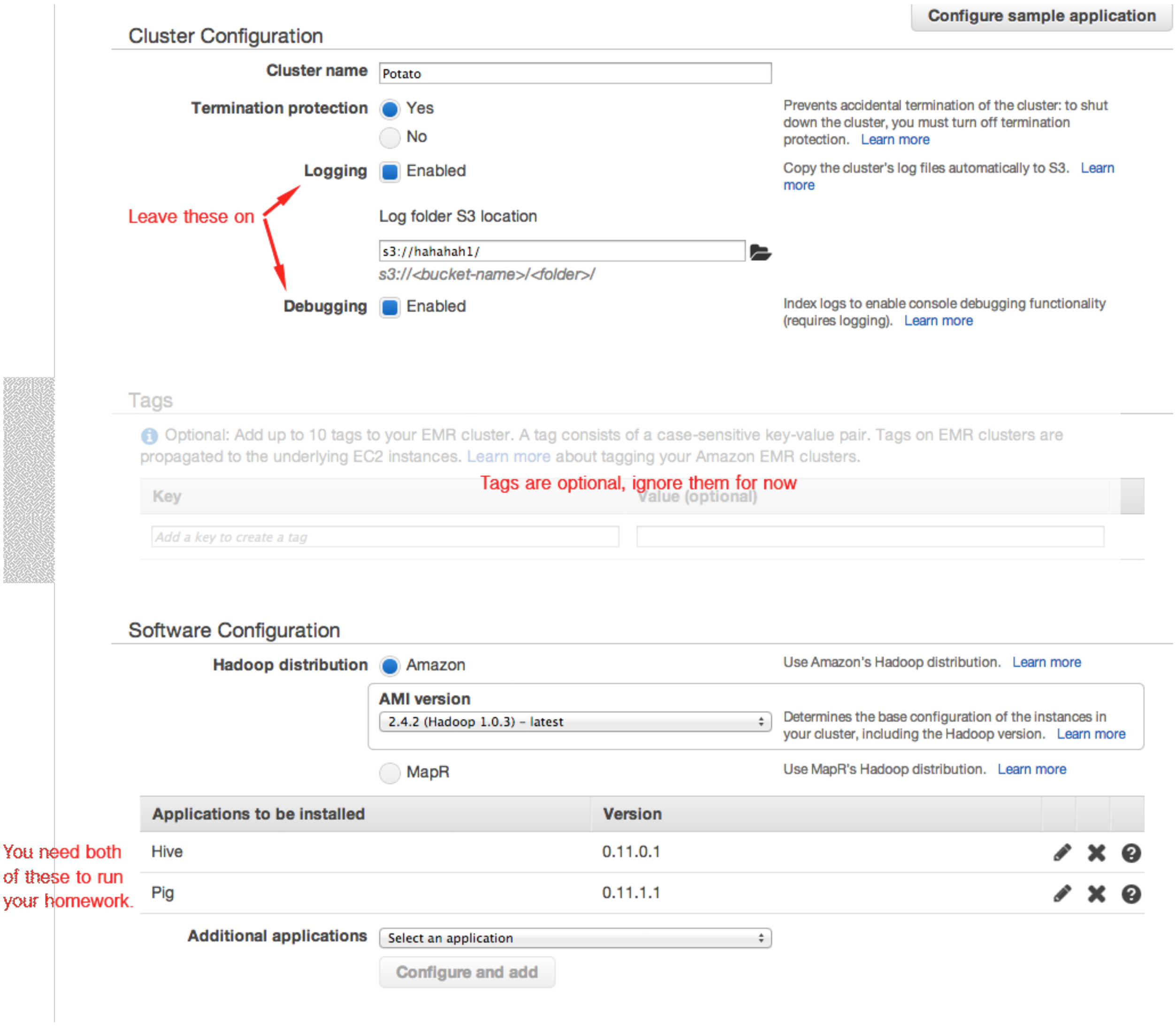
Key	Value (optional)
Add a key to create a tag	

Software Configuration

Hadoop distribution: ● Amazon — Use Amazon's Hadoop distribution. Learn more

AMI version

The configure screen has many options, selecting a sample application will configure the cluster for you. We will show you how to do it manually, since you will need to do so for your assignment.



Cluster Configuration

Cluster name: Potato

Termination protection: ● Yes ○ No

Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination protection. Learn more

Logging: ☑ Enabled

Copy the cluster's log files automatically to S3. Learn more

Leave these on

Log folder S3 location

s3://hahahah1/

s3://<bucket-name>/<folder>/

Debugging: ☑ Enabled

Index logs to enable console debugging functionality (requires logging). Learn more

Tags

Optional: Add up to 10 tags to your EMR cluster. A tag consists of a case-sensitive key-value pair. Tags on EMR clusters are propagated to the underlying EC2 instances. Learn more about tagging your Amazon EMR clusters.

Tags are optional, ignore them for now

Key	Value (optional)
Add a key to create a tag	

Software Configuration

Hadoop distribution: ● Amazon

Use Amazon's Hadoop distribution. Learn more

AMI version: 2.4.2 (Hadoop 1.0.3) – latest

Determines the base configuration of the instances in your cluster, including the Hadoop version. Learn more

○ MapR

Use MapR's Hadoop distribution. Learn more

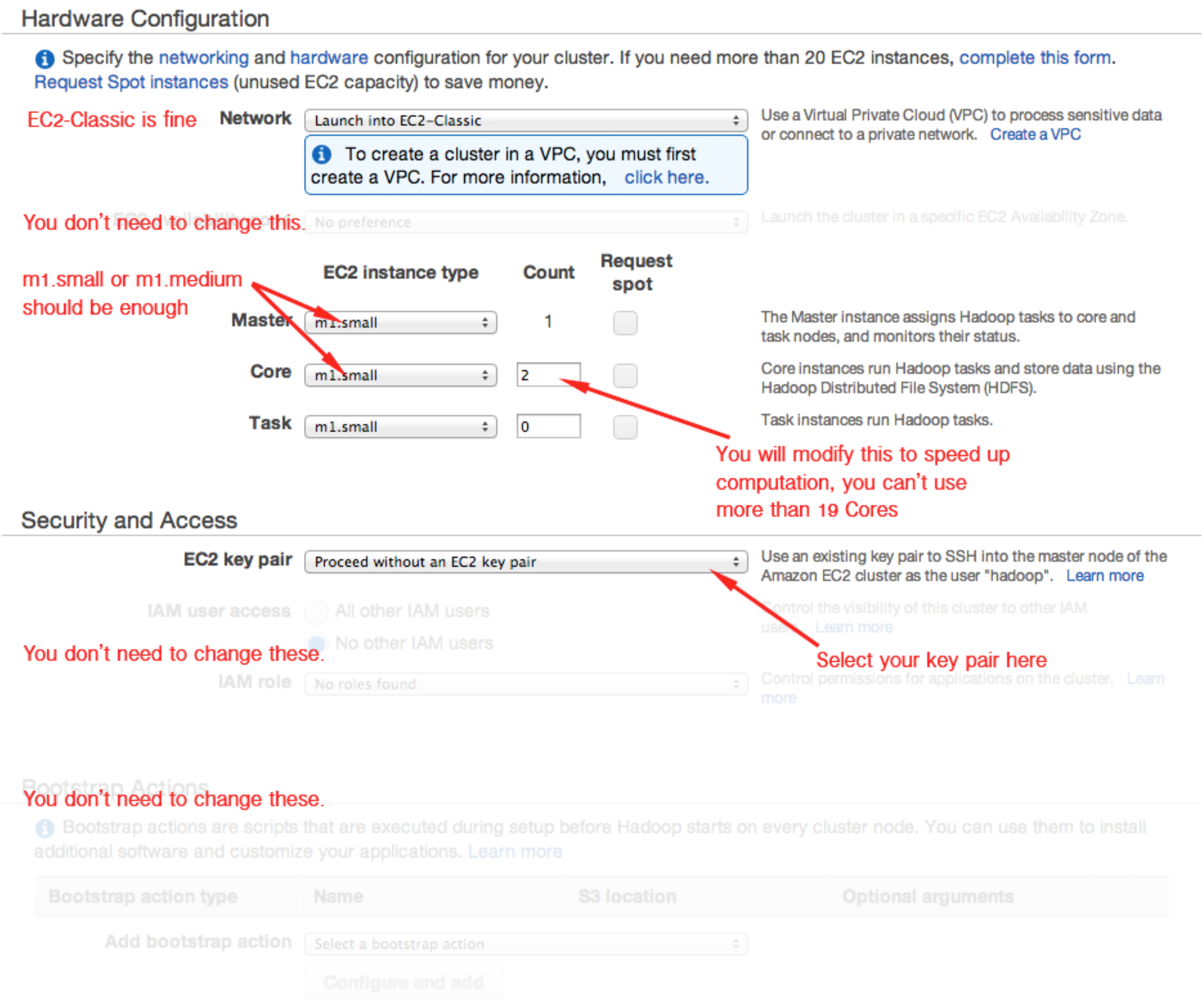
Applications to be installed	Version
Hive	0.11.0.1
Pig	0.11.1.1

You need both of these to run your homework.

Additional applications: Select an application

*Note the versions of the applications and the distribution may be different.

Make sure to select your logging bucket you made earlier under the log folder location.



Hardware Configuration

Specify the networking and hardware configuration for your cluster. If you need more than 20 EC2 instances, complete this form. Request Spot instances (unused EC2 capacity) to save money.

EC2-Classic is fine

Network: Launch into EC2-Classic

To create a cluster in a VPC, you must first create a VPC. For more information, click here.

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Create a VPC

You don't need to change this.

EC2 availability zone: No preference

Launch the cluster in a specific EC2 Availability Zone.

m1.small or m1.medium should be enough

	EC2 instance type	Count	Request spot	
Master	m1.small	1	□	The Master instance assigns Hadoop tasks to core and task nodes, and monitors their status.
Core	m1.small	2	□	Core instances run Hadoop tasks and store data using the Hadoop Distributed File System (HDFS).
Task	m1.small	0	□	Task instances run Hadoop tasks.

You will modify this to speed up computation, you can't use more than 19 Cores

Security and Access

EC2 key pair: Proceed without an EC2 key pair

Use an existing key pair to SSH into the master node of the Amazon EC2 cluster as the user "hadoop". Learn more

Select your key pair here

You don't need to change these.

IAM user access: ○ All other IAM users ● No other IAM users

Control the visibility of this cluster to other IAM users. Learn more

IAM role: No roles found

Control permissions for applications on the cluster. Learn more

Bootstrap Actions

You don't need to change these.

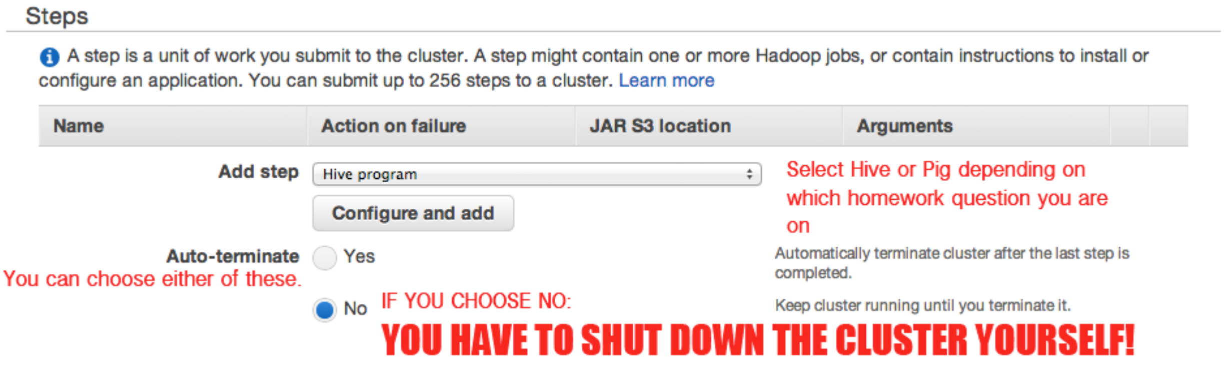
Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. Learn more

Bootstrap action type	Name	S3 location	Optional arguments

Add bootstrap action: Select a bootstrap action

Note: If your account supports only EC2-VPC, you can select the default VPC from the Network list i.e. you will not see "EC2-Classic".

The costs listed(http://aws.amazon.com/ec2/pricing/) are charged on an hourly rate, based on the number and type of nodes in your cluster.



Steps

A step is a unit of work you submit to the cluster. A step might contain one or more Hadoop jobs, or contain instructions to install or configure an application. You can submit up to 256 steps to a cluster. Learn more

Name	Action on failure	JAR S3 location	Arguments

Add step: Hive program

Configure and add

Select Hive or Pig depending on which homework question you are on

Auto-terminate: ○ Yes — Automatically terminate cluster after the last step is completed.

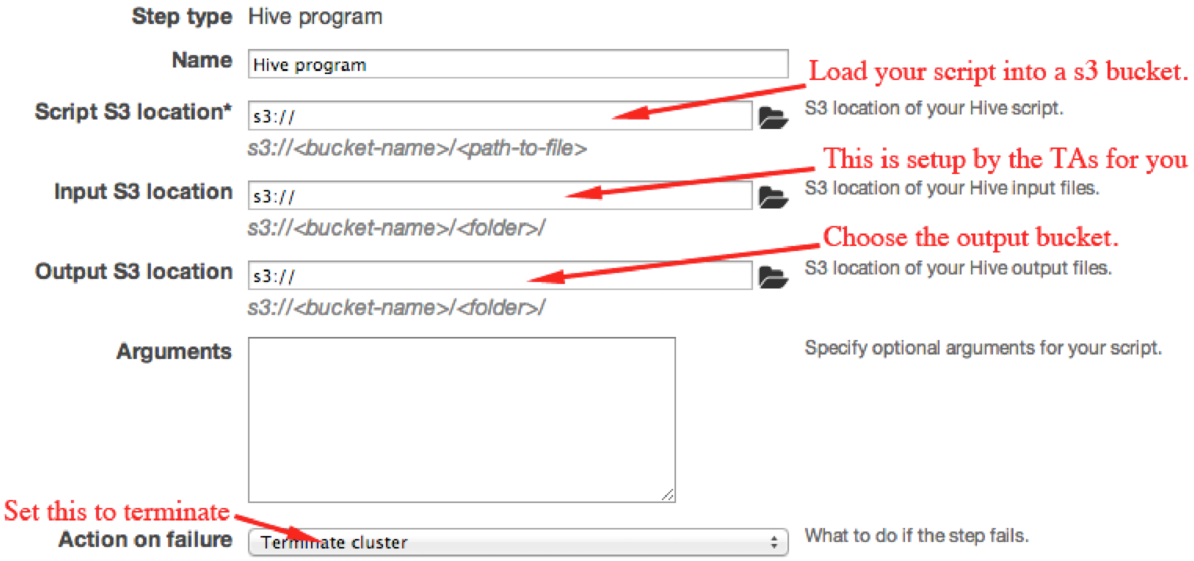
● No — Keep cluster running until you terminate it.

You can choose either of these.

IF YOU CHOOSE NO:

YOU HAVE TO SHUT DOWN THE CLUSTER YOURSELF!

Click “Configure and add” to add steps.



Step type: Hive program

Name: Hive program

Script S3 location*: s3://

s3://<bucket-name>/<path-to-file>

S3 location of your Hive script.

Load your script into a s3 bucket.

Input S3 location: s3://

s3://<bucket-name>/<folder>/

S3 location of your Hive input files.

This is setup by the TAs for you

Output S3 location: s3://

s3://<bucket-name>/<folder>/

S3 location of your Hive output files.

Choose the output bucket.

Arguments: — Specify optional arguments for your script.

Set this to terminate

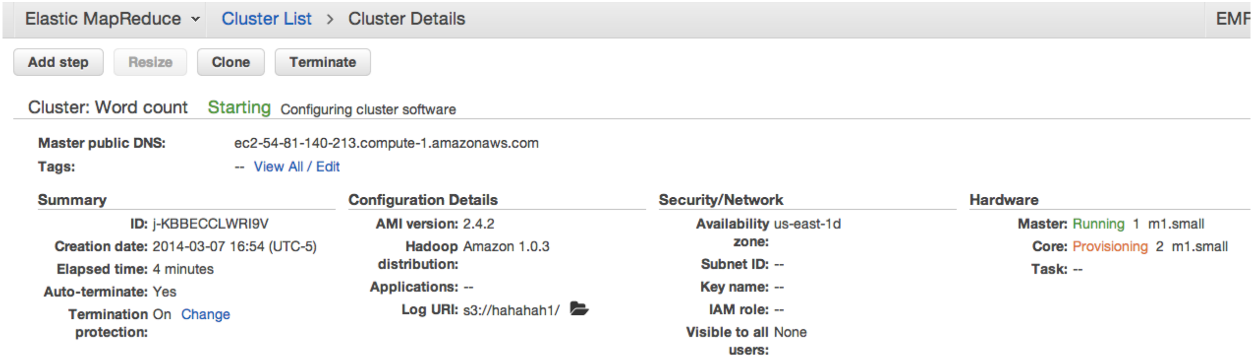
Action on failure: Terminate cluster — What to do if the step fails.

For the S3 output Location you should specify the bucket and an additional unique folder for each new run. It will help with organization.

Remember for the word count sample application it sets everything up for you. Scroll down to the end of the page, click on “**Create cluster**” to run the application.

You now can view the status of your application in “Cluster Details” screen. It takes several minutes for the whole process to run.
Provisioning - Amazon locates resources for your application
Bootstrapping - Amazon sets up and configures the nodes to run your application
Running - Runs and writes to your output bucket.
Terminating - Amazon deconstructs the setups you used for the application

You can track its progress once it’s been created.



Cluster: Word count Starting Configuring cluster software

Master public DNS: ec2-54-81-140-213.compute-1.amazonaws.com
Tags: -- View All / Edit

Summary
ID: j-KBBECCLWRI9V
Creation date: 2014-03-07 16:54 (UTC-5)
Elapsed time: 4 minutes
Auto-terminate: Yes
Termination protection: On Change

Configuration Details
AMI version: 2.4.2
Hadoop distribution: Amazon 1.0.3
Applications: --
Log URI: s3://hahahah1/

Security/Network
Availability zone: us-east-1d
Subnet ID: --
Key name: --
IAM role: --
Visible to all users: None

Hardware
Master: Running 1 m1.small
Core: Provisioning 2 m1.small
Task: --

After the application terminates, you could go back to the S3 output bucket you chose. The results will be written to the output folder. You should have several partxxxx files in the output folder. These are texts of the output! You have just successfully completed a MapReduce job flow on AWS and are ready for large scale data analytics!