

AWS Setup Guidelines

What we will accomplish?

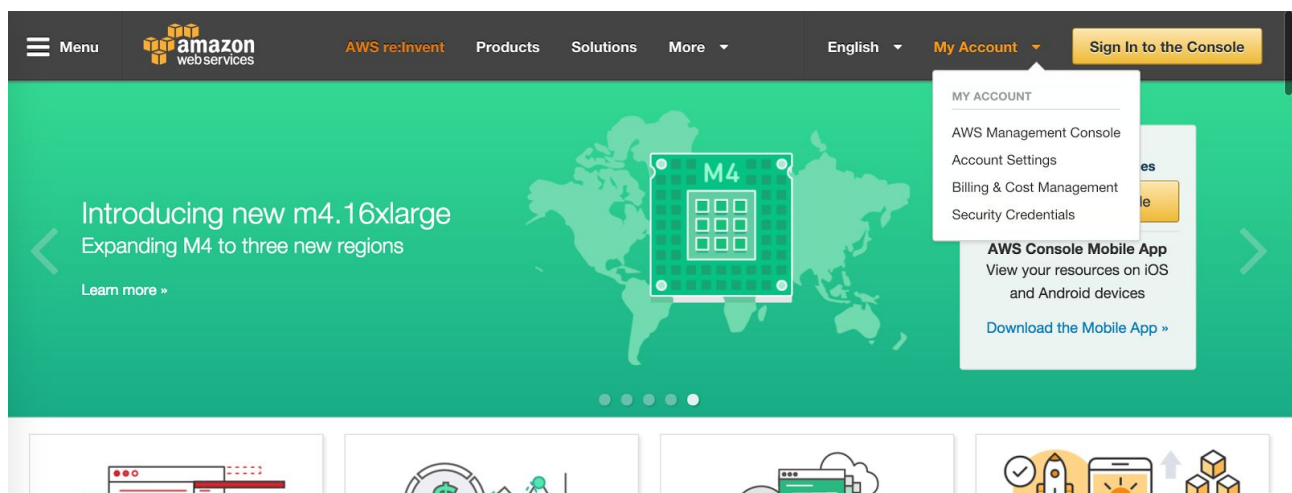
This guideline helps you get set up with the Amazon Web Services (AWS, a “cloud” platform) where you will run large-scale analysis on big data. Here are you will learn to do

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. [Set up a CloudWatch Usage Alert](#)
7. [Familiarize yourself with S3, EC2 and EMR](#) (by doing a sample MapReduce run).
8. [PIG Debugging](#)
9. [Instructions for Windows users](#)

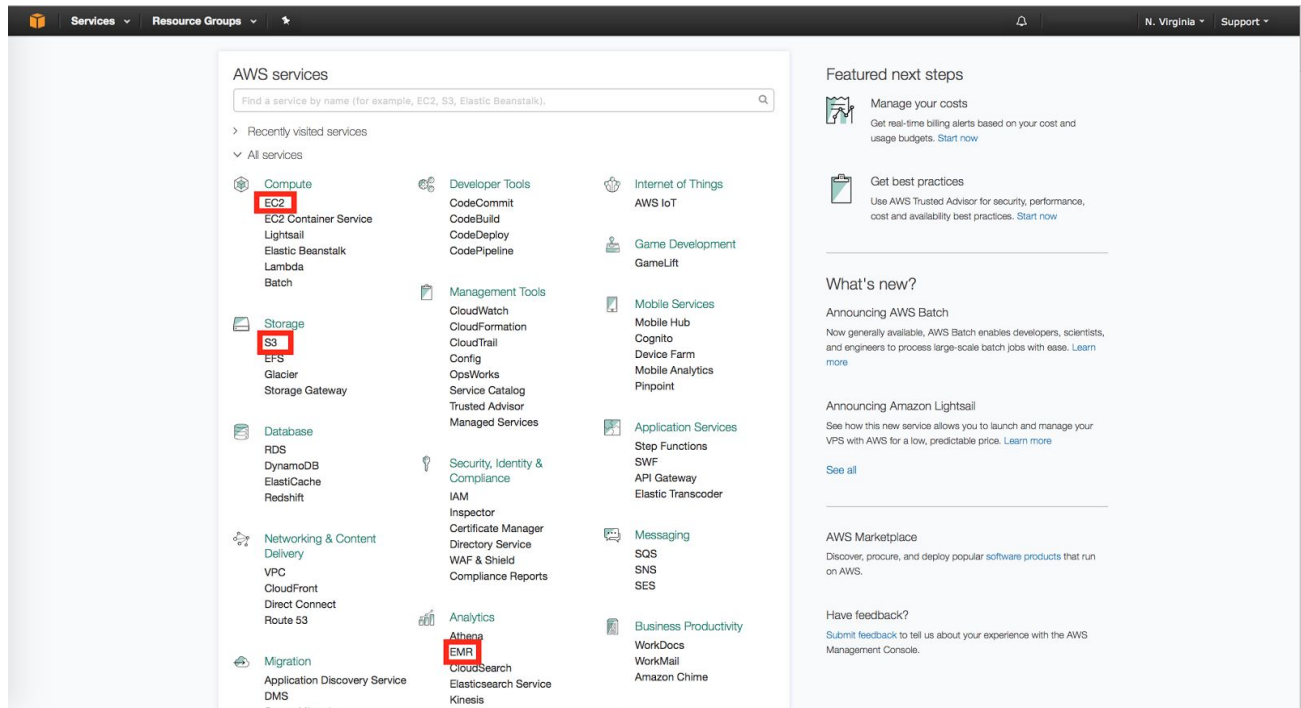
1. Create an AWS account

- Go to [AWS website](#) and sign up for an account.
- For now, please enter the required details, including payment details (you will need a **valid credit card or debit card** to sign up). Please follow Step 5 to redeem the \$100 credits.
- Validate your account with the identity verification through your phone.

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



You AWS Management Console should look like this:



We will be using the services highlighted in **red**. (S3, EC2 and EMR)

2. Create storage buckets on S3

We need S3 for two reasons:

- (1) An EMR workflow requires the input data to be on S3.
- (2) An 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. All S3 buckets need to have unique names. You will need to create 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 AWS Management Console click on **S3** under **All services** → **Storage**. In the S3 console, click on **Create Bucket**.

The screenshot shows the 'Create bucket' wizard in the AWS Management Console. The title bar is blue with the text 'Create bucket' and a close button (X). Below the title bar are four numbered steps: 1. Name and region (highlighted), 2. Set properties, 3. Set permissions, and 4. Review. The main content area is dark blue and contains the following fields:

- Name and region**
 - Bucket name**: A text input field containing 'cse6242-gtusername-logging'.
 - Region**: A dropdown menu showing 'US East (N. Virginia)'.
- Copy settings from an existing bucket**: A dropdown menu showing 'You have no buckets' and '0 Buckets'.

At the bottom of the form are three buttons: 'Create' (highlighted in blue), 'Cancel', and 'Next'.

- ii. Create a logging bucket: Click on **Create Bucket** and enter for following details.

Bucket name format **cse6242-<gt-username>-logging**

Create bucket ✕

- 1 Name and region
- 2 Set properties
- 3 Set permissions
- 4 Review

Name and region

Bucket name ⓘ

Region

 ▼

Copy settings from an existing bucket

 1 Buckets ▼

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 **Next** and then click on **Logging**.

Click on **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 **Save**. Then click

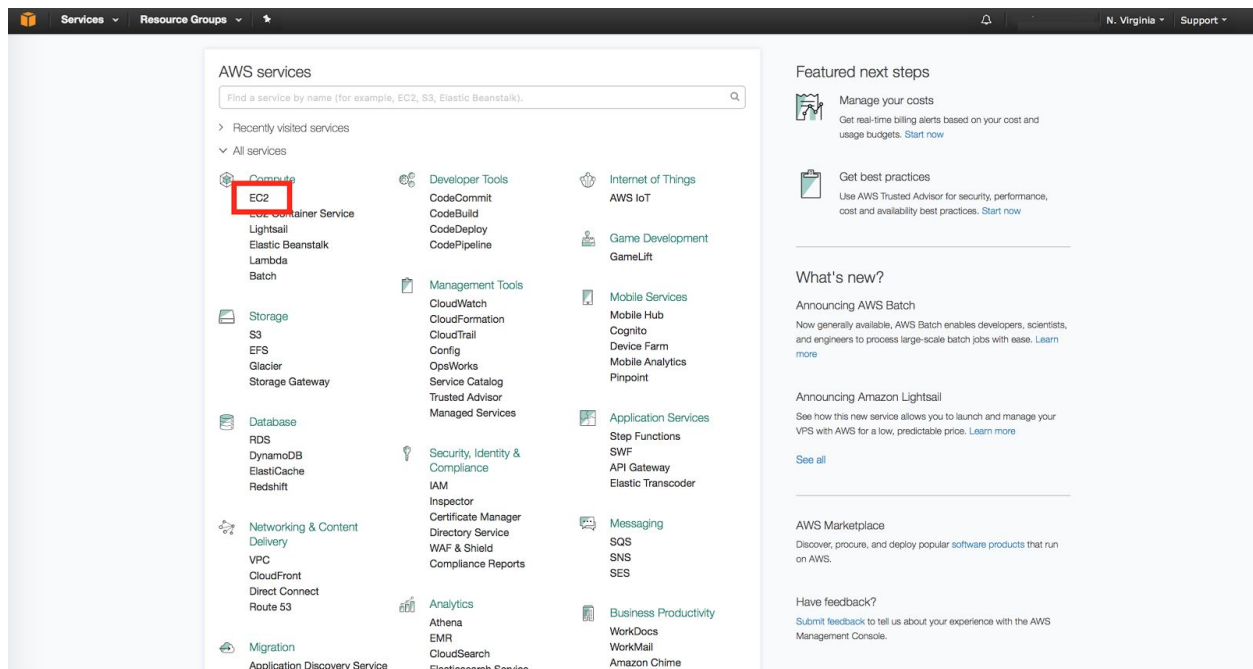
Next twice and click **Create bucket**.

The screenshot shows the 'Create bucket' wizard in Google Cloud Storage. The wizard has four steps: 1. Name and region (checked), 2. Set properties (current step), 3. Set permissions, and 4. Review. A 'Learn more' link is visible above a 'Disabled' toggle. A 'Logging' dialog box is open, allowing the user to choose between 'Enable logging' and 'Disable logging'. Under 'Enable logging', there are fields for 'Target bucket' (set to 'cse6242-gtusername-logging') and 'Target prefix' (with a placeholder 'Enter target prefix' and an information icon). 'Cancel' and 'Save' buttons are at the bottom of the dialog. Below the dialog, a 'Tags' section is partially visible. At the bottom right, there are 'Previous' and 'Next' navigation buttons.

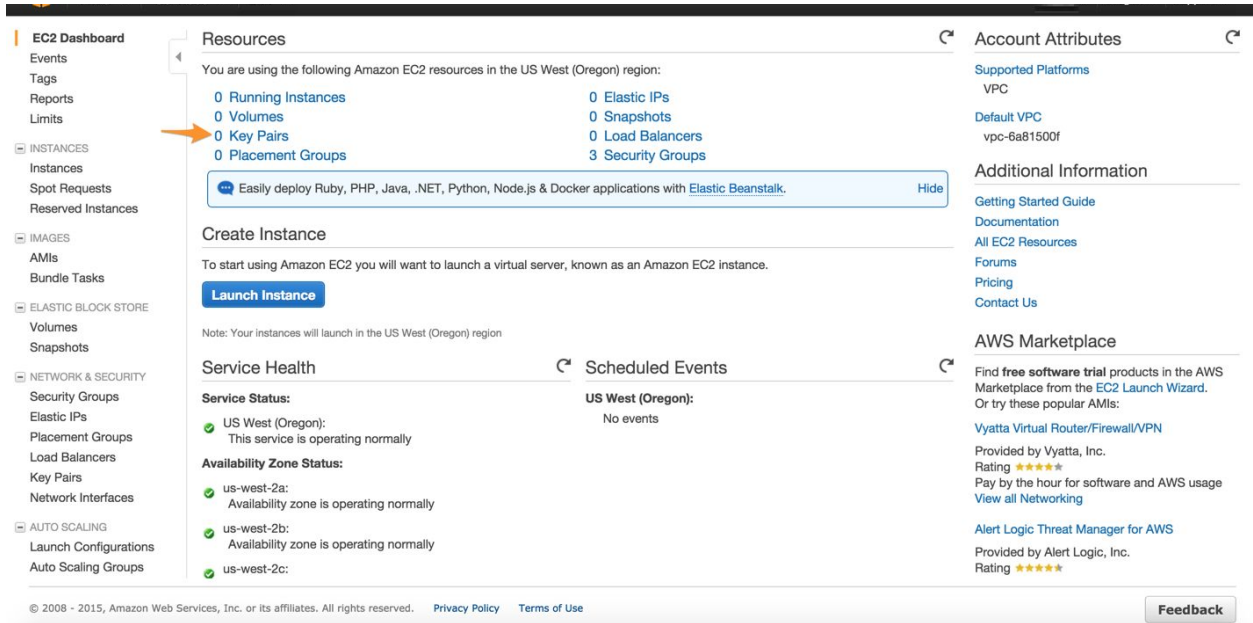
We are done creating the two buckets at this point.

3. Create a key pair

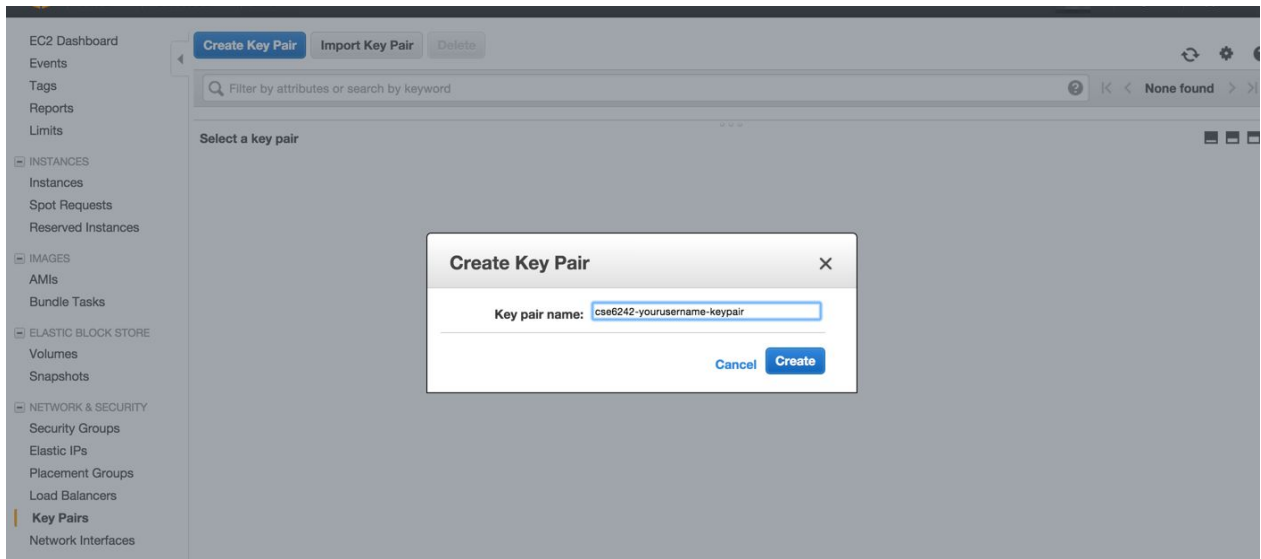
Note: Select the region on the top right as **US East (N. Virginia)** since the data bucket is stored in this region. When you run jobs on EMR, you will need to have a valid public/private key pair. To create your first key pair, click on **EC2** under **Compute** in the AWS Management Console.



You should see a link stating **0 Key Pairs** under Resources. Click on this.



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 be automatically downloaded. **Save it in a safe place where you will be able to find it again (IMPORTANT, do not lose this file).**



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](#).

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.

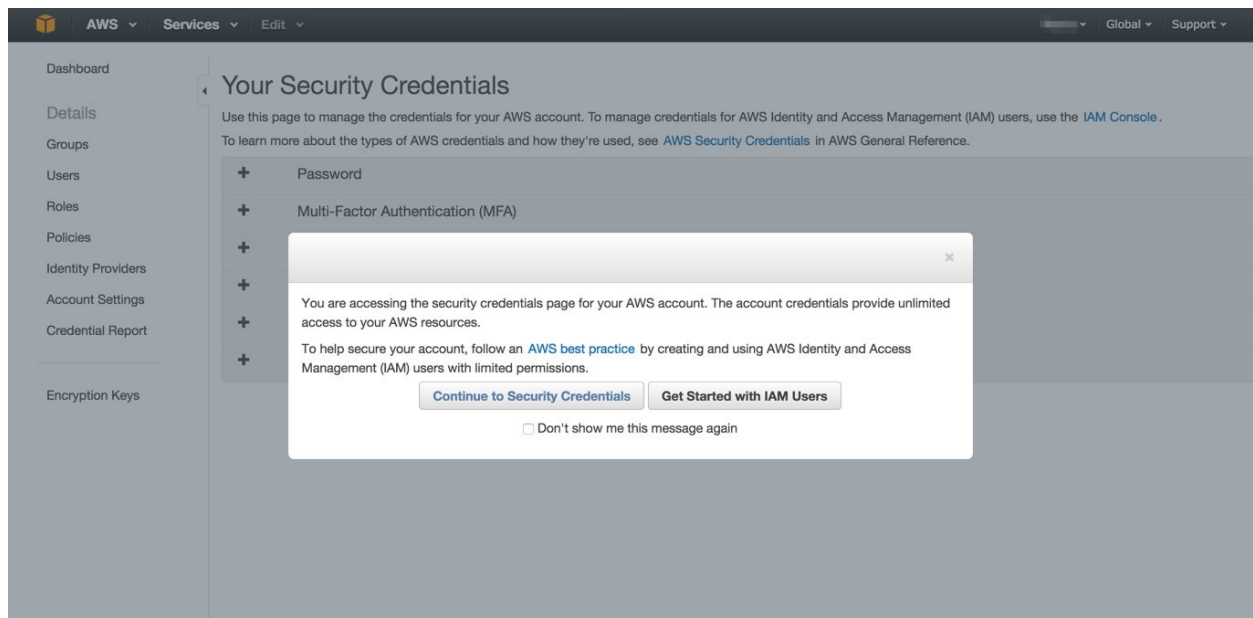
If you wish to log into the master node running your MapReduce job, you will need your .pem file (you will need this in case you wish to run an interactive PIG job flow). To log on to the master node (you can find the address of the master node from the MapReduce dashboard), you will need to do the following:

Note: Do not copy paste the command from this pdf as your command may fail due to the presence of special characters.

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

4. Get Access Keys

Click on **My Security Credentials** under your username (top right). Click on **Continue to Security Credentials**.



Click on **Access Keys** → **Create a new Access Key**. Download the Access Key file (**do not lose this file**). Now you are ready to run a MapReduce job.

Dashboard

Details

Groups

Users

Roles

Policies

Identity Providers

Account Settings

Credential Report

Encryption Keys

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.

- + Password
- + Multi-Factor Authentication (MFA)
- Access Keys (Access Key ID and Secret Access Key)

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

[Create New Access Key](#)

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.

- + CloudFront Key Pairs
- + X.509 Certificates

5. Redeem your free credit

To add the credit to your account, you will need your unique Credit Code obtained after applying for the AWS Educate program for Students (follow the steps at the start of HW3). Once you have your code, Click on your name (top right corner) → Click on **My Account**.

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

The screenshot shows the AWS Billing Dashboard 'Credits' page. The left sidebar contains a navigation menu with items: Dashboard, Bills, Cost Explorer, Payment Methods, Payment History, Consolidated Billing, Account Settings, Reports, Preferences, Credits (highlighted with an orange arrow), Tax Settings, and DevPay. The main content area is titled 'Credits' and includes a 'Redeem' button and a table of redeemed credits. The table has columns for Expiration Date, Credit Name, Credits Used, Credits Remaining, and Applicable Products. Below the table, there is a 'Total Amount of Credits Remaining' field. At the bottom of the page, there is a 'Choose language' dropdown set to 'English' and a 'Feedback' button.

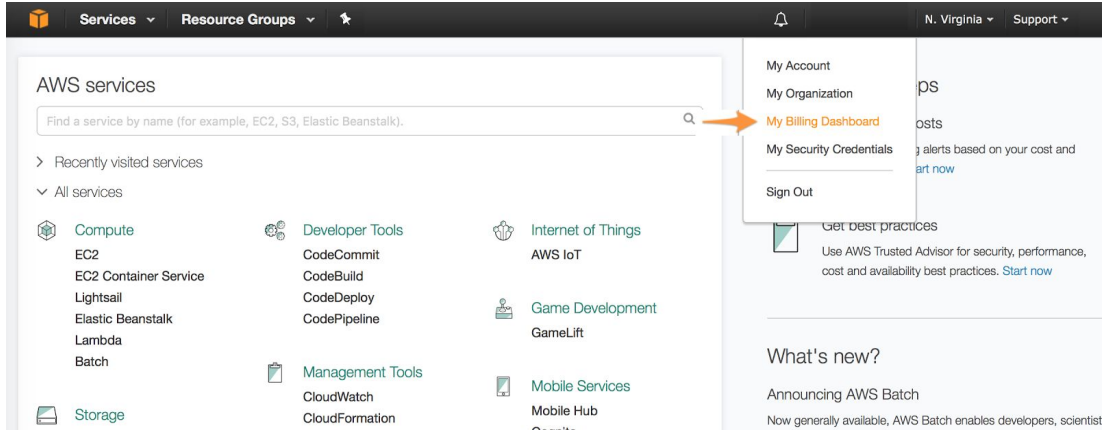
Please contact the CSE 6242 instructors immediately if this does not work. You can check the credit remaining by clicking on the **Bills** from your account page or by returning to this page. Sometimes this can take a while to update, so don't be surprised if recent changes are not immediately apparent. 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

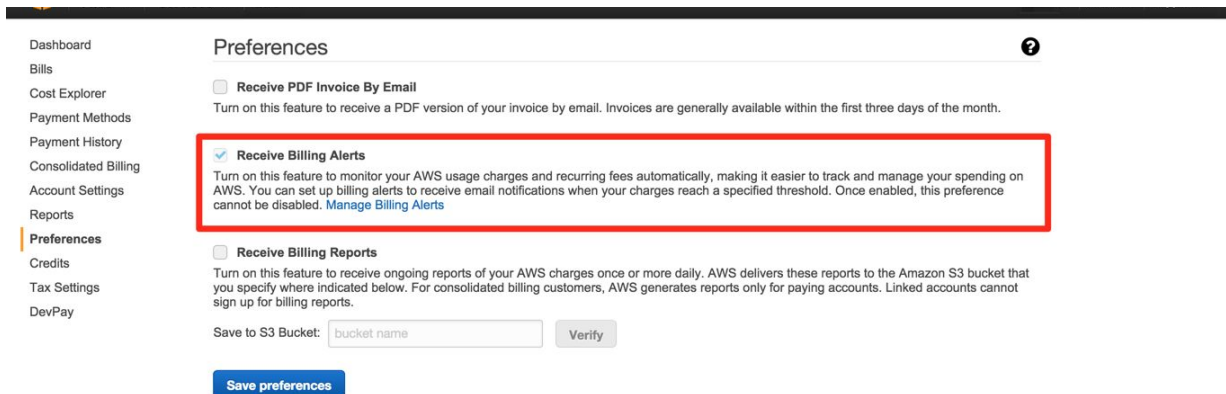
Make sure your region (in the upper right corner of the screen) is set to: US East (N. Virginia). [Test whether this email alert is working before scheduling in practice](#). That is, out of \$100, when your credit balance goes below say \$95, schedule a test alert and make sure it works. Remember this alert works only once. So once you get an alert for \$95, you schedule the next alert for \$70 and the next one for \$60 and so on.

Turn on Alerts

1. Go to My Billing Dashboard page.



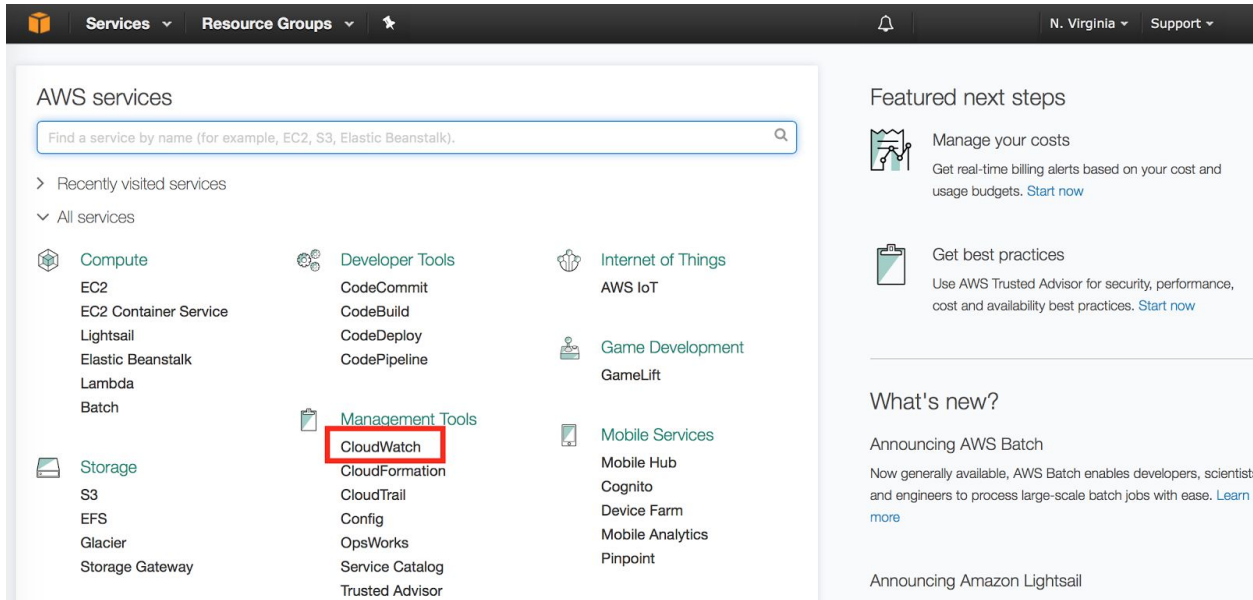
2. Under Preferences, check the box labeled **Receive Billing Alerts**



Turn on Custom Alerts

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

1. Click **CloudWatch** in the AWS Management Console.



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



3. In the **CloudWatch Metrics by Category** pane, under **1. Select Metric**, in the dropdown choose **Billing** and check currency as **USD**. Select **Maximum** and **6 Hours** in the dropdown as shown in the image below. Click **Next**.

4. Fill out the alarm details and click **New List** next to **Send notification to:** and enter your name and email.

Alarm Threshold

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

Name:

Description:

Whenever charges for: EstimatedCharges
 is: USD \$

Actions

Define what actions are taken when your alarm changes state.

Notification Delete

Whenever this alarm:

Send notification to: New list

Actions

Define what actions are taken when your alarm changes state.

Notification Delete

Whenever this alarm:

Send notification to:

Email list:

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

1. Click on the Elastic MapReduce (EMR) link in the Analytics section of the AWS Console.
2. Click on the button **Create cluster**, and then click **Go to advanced options**.
3. Under **Step 1: Software and Steps**, select only **Hadoop** and **Pig** from the options and unselect others in the Software Configuration options menu.
4. In the **Add Steps**, choose Step Type **Pig Program** and then click **Configure**. Check the option “Auto-terminate cluster after the last step is completed” when you’re not debugging the program.

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Vendor Amazon MapR

Release

Hadoop 2.7.2 Zeppelin 0.6.1 Tez 0.8.4

Ganglia 3.7.2 HBase 1.2.2 Pig 0.16.0

Hive 2.1.0 Presto 0.150 ZooKeeper 3.4.8

Sqoop 1.4.6 Mahout 0.12.2 Hue 3.10.0

Phoenix 4.7.0 Oozie 4.2.0 Spark 2.0.0

HCatalog 2.1.0

Edit software settings (optional)

Enter configuration Load JSON from S3

`classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]`

Add steps (optional)

Step type **Configure**

Auto-terminate cluster after the last step is completed

Cancel **Next**

1a. Fill the form with details as provided in the box and image below.

Name:	(any name)
Script S3 Location:	s3://cse6242-<gt-username>/pig-small.txt (upload the script here) OR s3://cse6242-<gt-username>/pig-big.txt
Input S3 Location:	s3://cse6242-2017fall-aws-small/* OR s3://cse6242-2017fall-aws-big/*

Output S3 Location: s3://cse6242-<gt-username>/output-small (don't create the output folder)
OR s3://cse6242-<gt-username>/output-big

Action on failure: Terminate cluster (else you may be charged even if the task fails)

Create an S3 bucket **cse6242-<gt-username>** (you must have already done this in step 2), and upload all pig scripts here. Do not create any output folder as given in the box above as they are created automatically. Each time you run your code, ensure you delete all the new output folders created or have different output folder names for each run (e.g., output1-big, output2-big, etc).

Summary: Upload your script to an S3 location and select the location of your script from the list of items available at **Script S3 Location**. For the S3 output Location you should specify the bucket and an **additional unique folder for each new run**. Now, click **Save**.

The screenshot shows the 'Add step' dialog box with the following configuration:

- Step type:** Pig program
- Name:** ngram-pig
- Script S3 location*:** s3://cse6242-gtusername/pig-small.txt
- Input S3 location:** s3://cse6242-2017fall-aws-small/*
- Output S3 location:** s3://cse6242-gtusername/output-small
- Arguments:** (Empty text area)
- Action on failure:** Terminate cluster

Note:

- Ensure that you first test your code on the **smaller** dataset.
- Each time you run the code, it may take a couple of hours to terminate.
- To test and debug your code step by step, refer to the Debugging section at the end of the document. This is highly recommended if you are not familiar with Pig.

2. For **Step 2: Hardware** configuration, modify the EC2 instances as per your needs and select **Next**. One Master instance and 1-15 Core instances should be sufficient. You may face Bootstrapping errors if you exceed a certain limit of core instances. For EC2 Subnet, select the one in us-east-1b.

Note:

You may sometimes get the following error:

The requested instance type m3.xlarge is not supported in the requested availability zone

If you face this error, repeat the cluster creation process by selecting other subnets of the form **us-east-1x** (where x can be a/b/c/d/e/f), until you no longer face this error.

Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware**
- Step 3: General Cluster Settings
- Step 4: Security

Hardware Configuration

If you need more than 20 EC2 instances, [see this topic](#).

Instance group configuration

- Uniform instance groups**
Specify a single instance type and purchasing option for each node type.
- Instance fleets**
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#)

Network vpc-8da45df5 (172.31.0.0/16) (default) [Create a VPC](#)

EC2 Subnet subnet-9cd9d6c6 | Default in us-east-1b

Root device EBS volume size 10 GiB

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1	m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Maximum bid price: \$	Not available for Master
Core Core - 2	m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Maximum bid price: \$	Not enabled
Task Task - 3	m3.xlarge 8 vCPU, 15 GiB memory, 80 SSD GB storage EBS Storage: none	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Maximum bid price: \$	Not enabled

The costs listed in [pricing](#) are charged on an hourly rate, based on the number and type of nodes in your cluster.

3. For **Step 3: General Cluster Settings**, type a cluster name of your choice, and add the correct path to the logging folder (created in Step 2). Check Logging, Debugging and Termination protection as shown in the image below. Click **Next**.

Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings**
- Step 4: Security

General Options

Cluster name

Logging ⓘ
S3 folder

Debugging ⓘ

Termination protection ⓘ

Scale down behavior

Tags ⓘ

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

EMRFS consistent view ⓘ

Custom AMI ID

▶ Bootstrap Actions

[Cancel](#) [Previous](#) [Next](#)

4. For **Step 4: Security**, select your key pair and click **Create Cluster** to run the application.

Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security**

Security Options

EC2 key pair

Cluster visible to all IAM users in account ⓘ

Permissions ⓘ

Default Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) ⓘ

EC2 Instance profile [EMR_EC2_DefaultRole](#) ⓘ

▶ Encryption Options

▶ EC2 Security Groups

[Cancel](#) [Previous](#) [Create cluster](#)

5. The cluster must start running as follows:

You now can view the status of your application in this “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.

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.

8. Debugging

A very important part of running Pig Scripts on AWS is the ability to also run your code directly on the master node. You can run your script step by step and identify the exact step where an error occurred. The steps to debug are given below.

1. You must repeat all the steps in Section 7, except with three modifications:
 - a. Ensure that you verify the script location, its input and output path. Do this each time you create/clone a cluster (many students make a mistake here).
 - b. Modify the action on failure option to “Continue”.
 - c. Uncheck the “Auto-terminate cluster after....” option.

Note: You must revert back these changes after debugging else you may leave the clusters

running forever and you will be charged for this!

Add step

Step type Pig program

Name Pig program

Script S3 location* s3://cse6242-gtusername/pig-small.txt
S3 location of your Pig script.
s3://<bucket-name>/<path-to-file>

Input S3 location s3://cse6242-2017fall-aws-small/*
S3 location of your Pig input files.
s3://<bucket-name>/<folder>/

Output S3 location s3://cse6242-gtusername/output-small
S3 location of your Pig output files.
s3://<bucket-name>/<folder>/

Arguments
Specify optional arguments for your script.

Action on failure Continue What to do if the step fails.

Cancel Save

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release emr-5.9.0

- | | | |
|--|--|---|
| <input checked="" type="checkbox"/> Hadoop 2.7.3 | <input type="checkbox"/> Zeppelin 0.7.2 | <input type="checkbox"/> Livy 0.4.0 |
| <input type="checkbox"/> Tez 0.8.4 | <input type="checkbox"/> Flink 1.3.2 | <input type="checkbox"/> Ganglia 3.7.2 |
| <input type="checkbox"/> HBase 1.3.1 | <input checked="" type="checkbox"/> Pig 0.17.0 | <input type="checkbox"/> Hive 2.3.0 |
| <input type="checkbox"/> Presto 0.184 | <input type="checkbox"/> ZooKeeper 3.4.10 | <input type="checkbox"/> Sqoop 1.4.6 |
| <input type="checkbox"/> Mahout 0.13.0 | <input type="checkbox"/> Hue 4.0.1 | <input type="checkbox"/> Phoenix 4.11.0 |
| <input type="checkbox"/> Oozie 4.3.0 | <input type="checkbox"/> Spark 2.2.0 | <input type="checkbox"/> HCatalog 2.3.0 |

Edit software settings (optional)

Enter configuration Load JSON from S3

classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]

Add steps (optional)

Name	Action on failure	JAR location	Arguments
Pig program	Continue	command-runner.jar	pig-script --run-pig-script --pig-versions 0.17.0 --args -f s3://cse6242-gtusername/pig-small.txt -p INPUT=s3://cse6242-2017fall-aws-small/* -p OUTPUT=s3://cse6242-gtusername/output-small

Step type Select a step

Configure

Auto-terminate cluster after the last step is completed

2. Once the cluster is running, you can open the TCP Port of your Master node to allow SSH

connections. Click on the security group of your master node.

The screenshot shows the Amazon EMR console for a cluster named 'ngram' in a 'Terminated' state. The left sidebar contains navigation options: Amazon EMR, Cluster list, Security configurations, VPC subnets, and Help. The main content area is divided into three columns: Summary, Configuration Details, and Network and Hardware. The Summary column shows the cluster ID (j-1ATUWNEV0TS4F), creation and end dates, and elapsed time. The Configuration Details column lists the release label (emr-5.0.0), Hadoop distribution (Amazon 2.7.2), applications (Pig 0.16.0), log URI, and EMRFS settings. The Network and Hardware column shows availability zone (us-east-1b), subnet ID (subnet-42f30634), and master/core tasks. A 'Security and Access' section is highlighted with a red box, showing the key name (cse6242-mnatraj3-keypair), EC2 instance profile (EMR_EC2_DefaultRole), and a list of security groups: sg-ea868890 (ElasticMapReduce-Master: master) and sg-eb868891 (ElasticMapReduce-slave).

Add an entry for SSH in the inbound tab of your master node with the exact details as follows.

The screenshot shows the Amazon Security Groups console. The left sidebar lists navigation options: EC2 Dashboard, Events, Tags, Reports, Limits, INSTANCES, IMAGES, ELASTIC BLOCK STORE, NETWORK & SECURITY, LOAD BALANCING, AUTO SCALING, and COMMANDS. The main content area shows the 'Create Security Group' page for 'sg-ea868890'. A table lists security groups, with 'sg-ea868890' highlighted. Below, the 'Inbound' tab is selected, showing a table of rules. A red box highlights the SSH rule configuration: Type: SSH, Protocol: TCP, Port Range: 22, Source: 0.0.0.0/0.

Name	Group ID	Group Name	VPC ID	Description
sg-ea868890	sg-ea868890	ElasticMapReduce-master	vpc-bd62d0d9	Master group for Elastic...
sg-eb868891	sg-eb868891	ElasticMapReduce-slave	vpc-bd62d0d9	Slave group for Elastic...

Type	Protocol	Port Range	Source
All TCP	TCP	0 - 65535	sg-ea868890 (ElasticMapReduce-master)
All TCP	TCP	0 - 65535	sg-eb868891 (ElasticMapReduce-slave)
SSH	TCP	22	0.0.0.0/0
Custom TCP Rule	TCP	8443	207.171.112.0/32
Custom TCP Rule	TCP	8443	54.239.98.0/24
Custom TCP Rule	TCP	8443	54.240.217.8/29
Custom TCP Rule	TCP	8443	207.171.167.26/32
Custom TCP Rule	TCP	8443	72.21.198.64/29
Custom TCP Rule	TCP	8443	207.171.167.101/32
Custom TCP Rule	TCP	8443	72.21.196.64/29
Custom TCP Rule	TCP	8443	54.240.217.80/29
Custom TCP Rule	TCP	8443	54.240.217.16/29
Custom TCP Rule	TCP	8443	207.171.167.25/32
Custom TCP Rule	TCP	8443	72.21.217.0/24
Custom TCP Rule	TCP	8443	54.240.217.64/28
All UDP	UDP	0 - 65535	sg-ea868890 (ElasticMapReduce-master)
All UDP	UDP	0 - 65535	sg-eb868891 (ElasticMapReduce-slave)
All ICMP	All	N/A	sg-ea868890 (ElasticMapReduce-master)
All ICMP	All	N/A	sg-eb868891 (ElasticMapReduce-slave)

You can now SSH into your master node.

3. To SSH, first copy the command as follows.

The image shows the Amazon EMR console interface. On the left is a navigation menu with 'Cluster list' selected. The main area displays details for a cluster named 'ngram' which is 'Terminated' and has 'Steps completed'. At the top of the cluster details are buttons for 'Add step', 'Resize', 'Clone', 'Terminate', and 'AWS CLI export'. The 'Master public DNS' field is highlighted with a red box and contains the text 'ec2-54-208-56-158.compute-1.amazonaws.com' followed by an 'SSH' link, also highlighted with a red box. Below this are sections for 'Summary', 'Configuration Details', and 'Network and Hardware'. The 'Summary' section includes ID, creation and end dates, elapsed time, and termination protection. 'Configuration Details' lists the release label, Hadoop distribution, applications, and log URI. 'Network and Hardware' shows availability zone, subnet ID, and master/core instances. A 'Security and Access' section is also present. Overlaid on the bottom of the console is a window titled 'SSH' with the heading 'Connect to the Master Node Using SSH'. It provides instructions for connecting via SSH and shows a terminal command: `ssh -i ~/cse6242-mnatraj3-keypair.pem hadoop@ec2-54-208-56-158.compute-1.amazonaws.com`, which is highlighted with a red box. The window also has 'Windows' and 'Mac / Linux' tabs and a 'Close' button.

Modify the path to your `.pem` file and run the command on your terminal. **Ensure that the file permissions of your `.pem` file is set to 400.**

4. You will now be logged into the master node. Type `pig` to be able to run commands on the pig shell.

5. Run your code line by line and spot the errors! Use commands such as `illustrate`, `describe` and `dump` to debug your code.

9. Instructions for Windows Users

1. Open PuTTY on your computer. On the cluster's page you'll see the term **Master DNS**. Under host name in PuTTY, type `hadoop@DNS`, where DNS is the aforementioned master DNS.
2. On the left toolbar, go to connection → SSH → Auth and upload the private key you should have generated earlier (the .ppk file).
3. Click Open, then Yes to the popup.
4. You should see a hadoop shell. Type `pig` and press enter to get started.