Amazon AWS introduction

CS-543

Prerequisites

Amazon services we need:

Amazon provides a quite a few services. We only need:

- EMR
- S3

EMR is the service that allows you to instantiate a cluster, rented by amazon with all the tools you will need (hadoop - spark-shell).

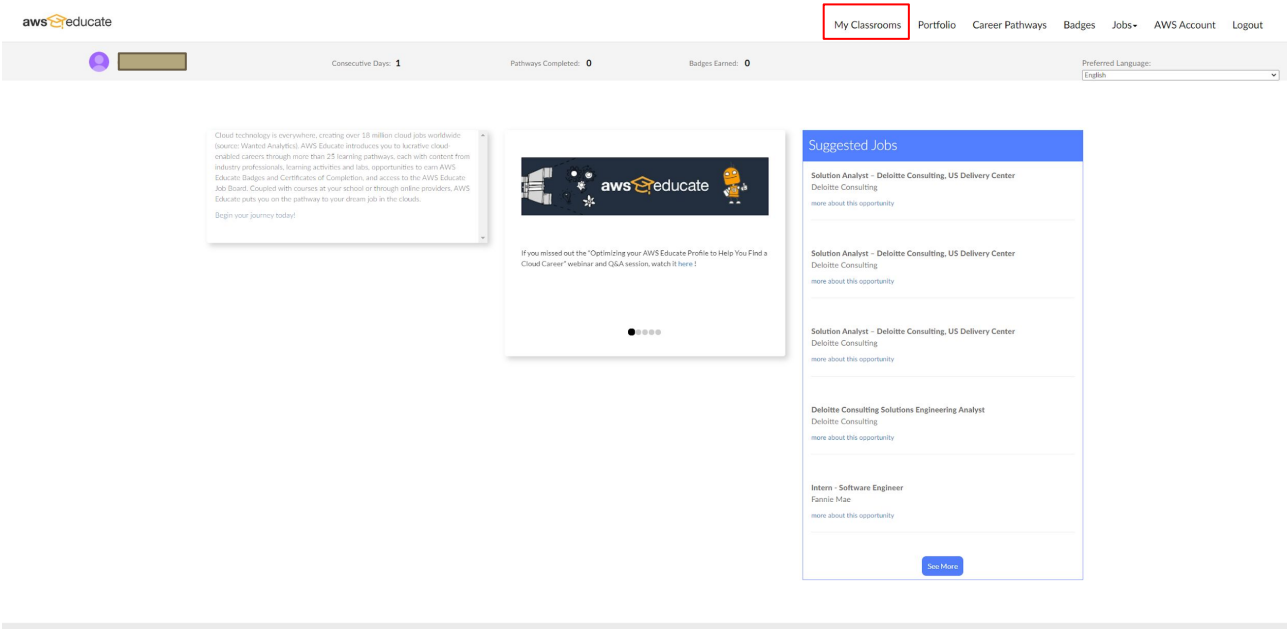
S3 is persistent storage for your files accessible from your cluster machines. When you stop a VM you lose any data loaded to it, so if you want to store it someplace you should use S3.

Setting up the services

Setting up the services

With your newly created amazon educate account head over to https://www.awseducate.com/signin/SiteLogin and login. Once you have logged on press the “AWS account” button on the top right corner of the webpage and the orange “AWS educate starter account” orange button afterwards.

Setting up the services



Setting up the services

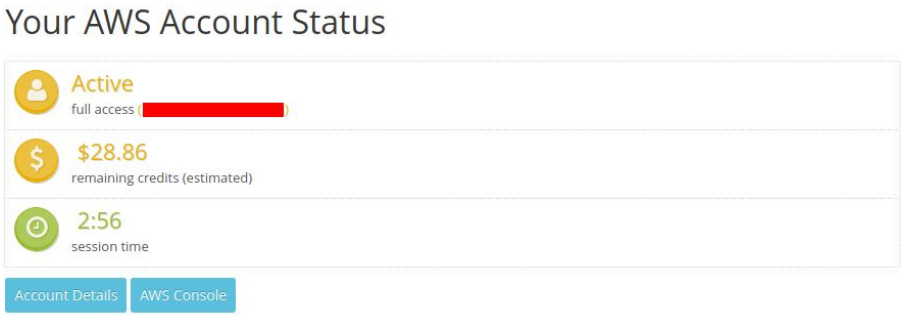
My Classrooms

View your list of Classroom invitations and accept or decline the invitation. Access a Classroom by clicking Go to my classroom.

Course Name ↓↑	Description	Educator ↓↑	Course End Date ↓↑	Credit Allocated Per Student ↓↑	Status
Software Systems and Technologies for Big Data Applications	This is an introductory course to Big Data technologies. The course studies a series of problems, such as, distributed ETL and Machine Learning at scale, distributed data representation, columnar storage, NoSQL databases, eventual consistency, real time analytics (streaming), and cluster management	Christos Kozanitis	06/30/2021	$40	Accepted Go to classroom

Setting up the services

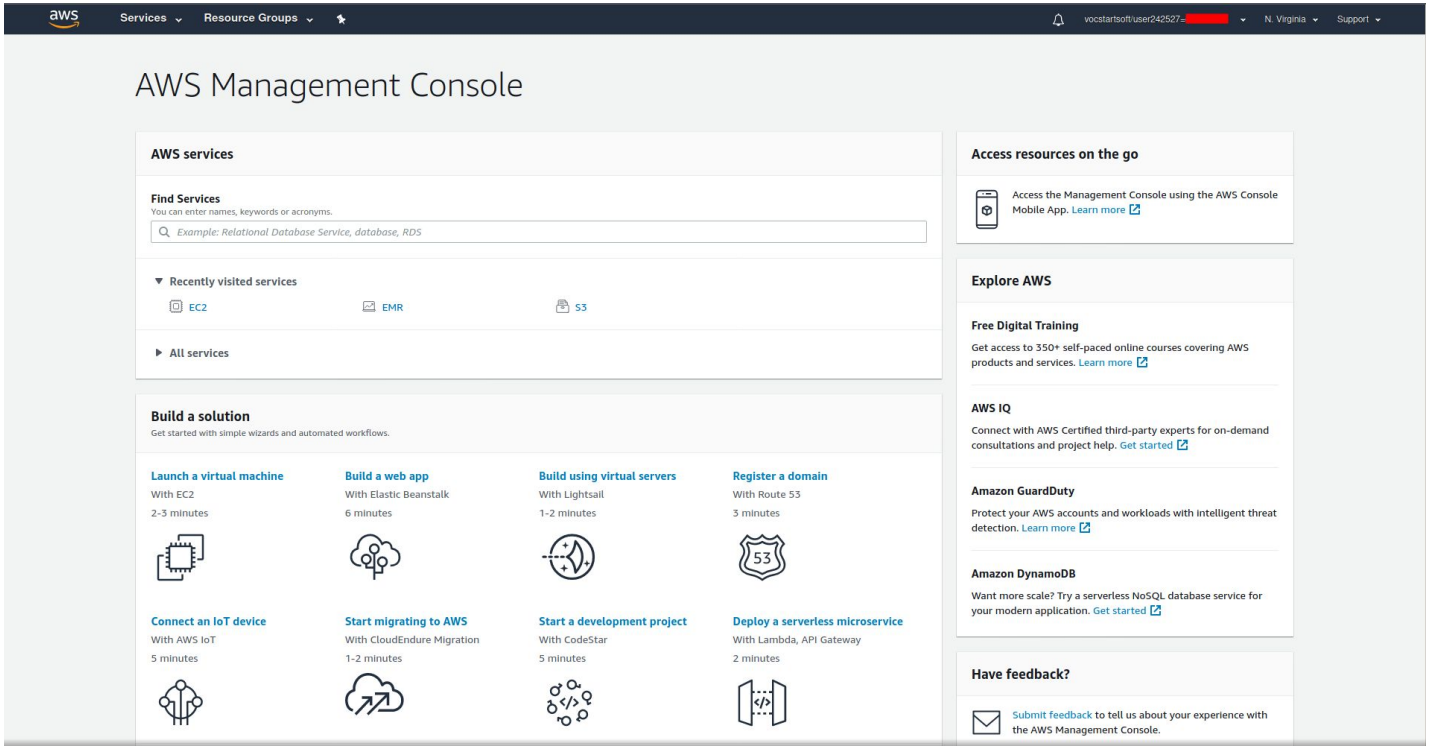
You should see a screen containing the following:



Pressing the “Account Details” > “show” next to Amazon CLI will show you your aws credentials. We will need them later. Keep in mind these change upon **LOGIN**.

Pressing “AWS Console” will redirect you to the Amazon dashboard where you access all the services amazon provides.

Setting up the services



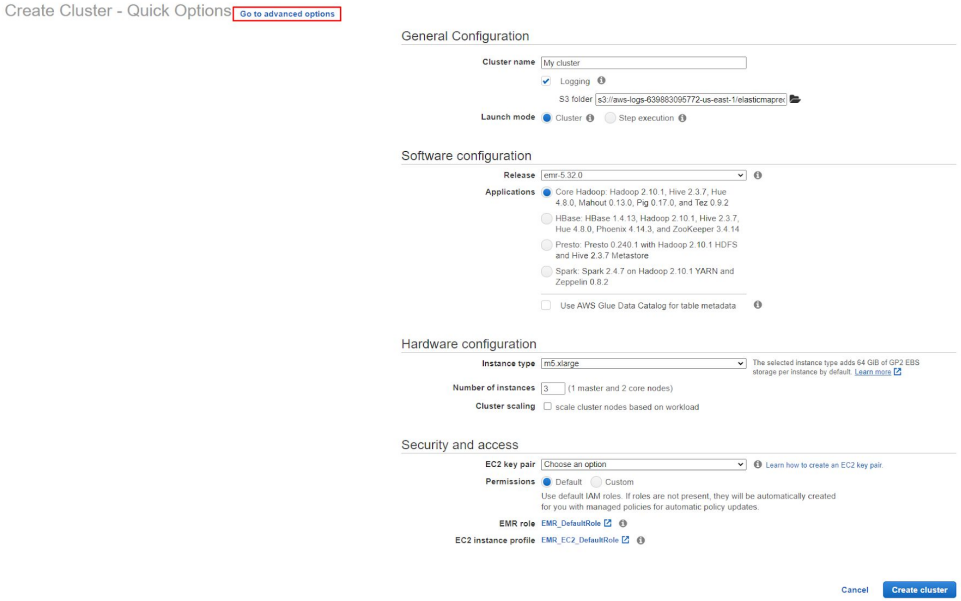
Setting up a key pair

1. Go to AWS management console, and select EC2
2. Click on key-pairs
3. On top right corner press “Create key pair”
4. Enter a name for it. Feel free to use either ppk or pem, but i would suggest using pem if you have linux.
5. When you click “Create key pair” a new key will be downloaded automagically.

- In order to use that key it needs to have 700 permissions. (chmod 700 key.pem)

Elastic Map Reduce - EMR

- Go to AWS Management Console and select EMR.
- Press Create Cluster > advanced options.



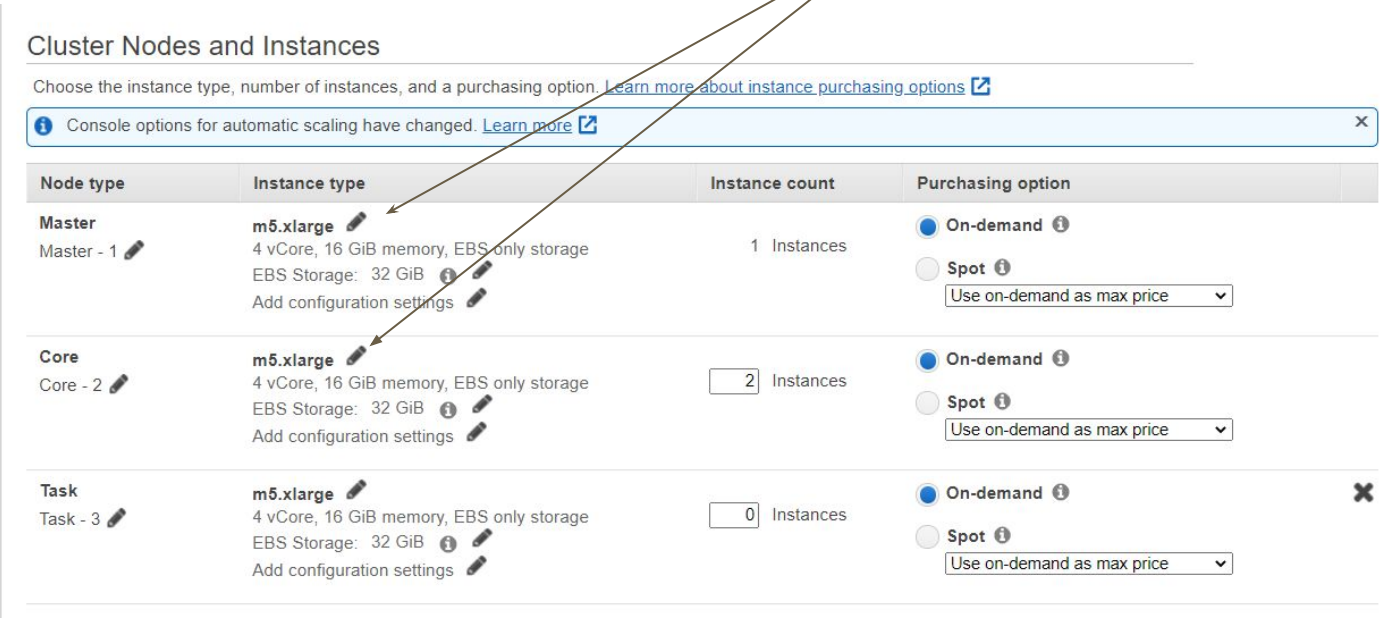
Elastic Map Reduce - EMR

- In step 1: software and steps
  - Leave release on default.
  - Select Hadoop and spark. (ganglia and zeppelin might be useful to you but not required)
  - No need to change anything else
- In Step 2: Hardware

Elastic Map Reduce - EMR

- In Step 2: Hardware

Change the specs of each machine.



Elastic Map Reduce - EMR

- In Step 2: Hardware

Change Space. SSDs are costly and not really needed use magnetics 30 / 40 GB should be ok

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

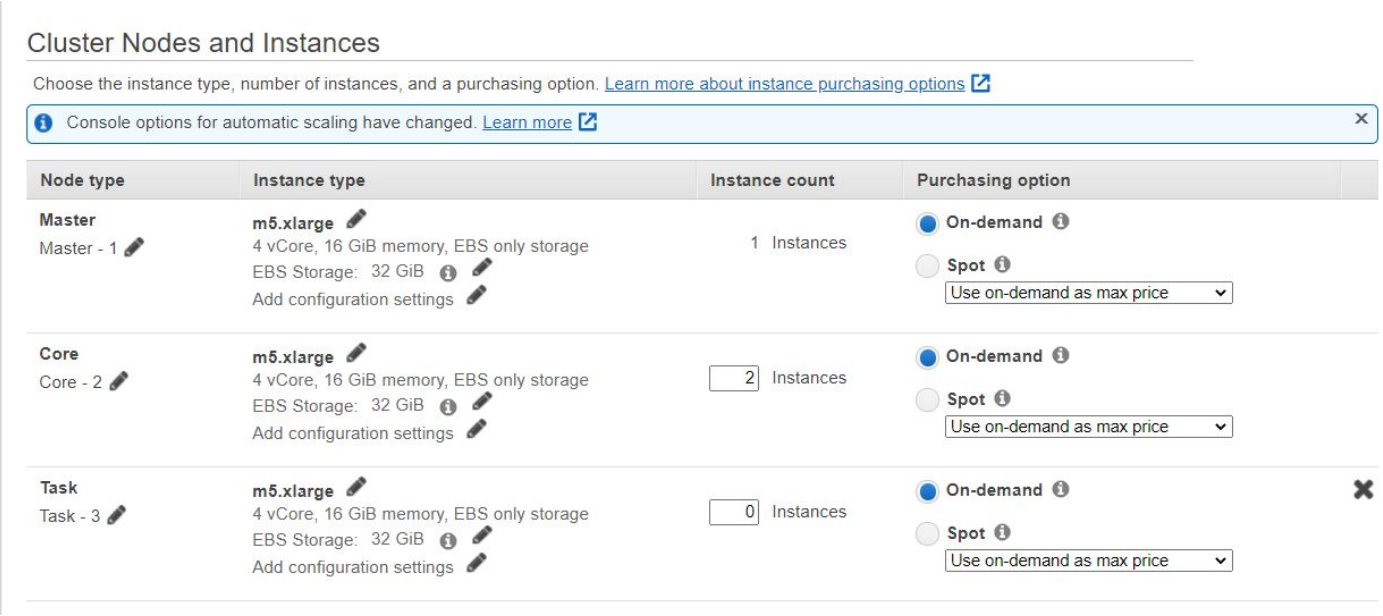
Console options for automatic scaling have changed. Learn more

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	2 Instances	On-demand Spot Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB Add configuration settings	0 Instances	On-demand Spot Use on-demand as max price

Elastic Map Reduce - EMR

- Consider using spot instead of on-demand
- Lower cost, same machines.
- If the cost goes up you are kicked out effective immediately
- Better for small tests than paying full price

- In Step 2: Hardware



Elastic Map Reduce - EMR

- In Step 2: Hardware
  - Change EBS root Volume to 30 GB for starters, keep in mind you might have to increase that.

Elastic Map Reduce - EMR

- In Step 3: General Cluster Settings
  - Provide a name for your cluster.
  - Turn of logging, not needed.
  - In S3 folder, check your S3 folder.
  - No need for any tags.

Elastic Map Reduce - EMR

- In Step 4: Security
  - From the drop-down menu, select your previously created key.
  - Press create Cluster.

Elastic Map Reduce - EMR

- You should see this screen and your cluster booting up

Filter: All clusters | Filter clusters ... | 2 clusters (all loaded)

Name	ID	Status	Creation time (UTC+2)	Elapsed time	Normalized instance hours
hy543-cluster	j-1SQLBIW9Y1JCG	Terminated User request	2021-03-03 18:17 (UTC+2)	2 hours, 11 minutes	48

- Click on your cluster.
- Click on the link next to “security groups for master”, it should open a new tab with:

Security Groups (2) Info

Actions | Create security group

Filter security groups

search: sg-07f8cbf249f14b087 | Clear filters

Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound r
–	sg-07f8cbf249f14b087	ElasticMapReduce-mas...	vpc-800ea4fd	Master group for Elasti..	639883095772	21 Permis
–	sg-0afcc92a68877c149	ElasticMapReduce-slave	vpc-800ea4fd	Slave group for Elastic ...	639883095772	6 Permissi

Click on the security group ID of the master

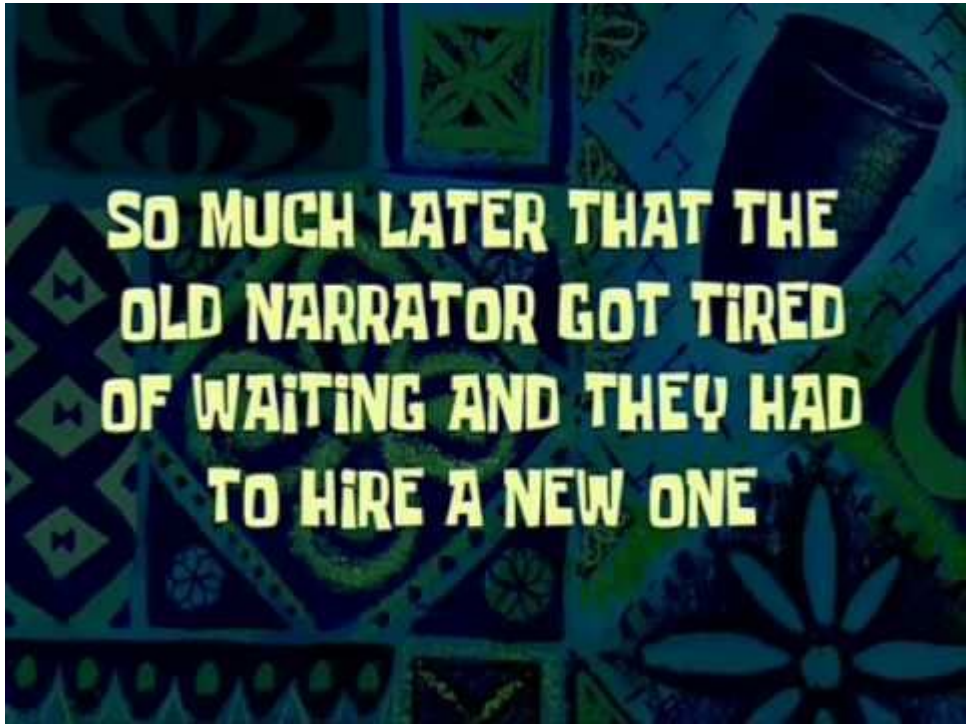
Elastic Map Reduce - EMR

- Scroll down a bit, you should see a bunch of Inbound rules. Press the “edit inbound rules”
- Press add rule
  - Select SSH
  - On the source drop-down menu select either your ip or anywhere. (to avoid any issues I would suggest anywhere.
- Press save rules.

Elastic Map Reduce - EMR

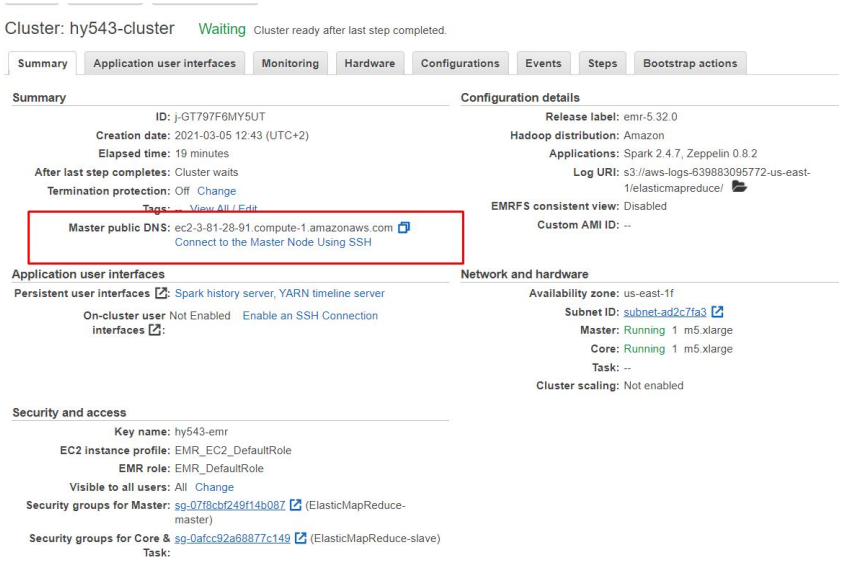
- In order to connect to your machine cluster, first wait for it to boot up.

Elastic Map Reduce - EMR



Elastic Map Reduce - EMR

- In your clusters info you should see the following:



Elastic Map Reduce - EMR

- Pressing that hyperlink you can see how you can ssh to your machine.

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. Learn more.

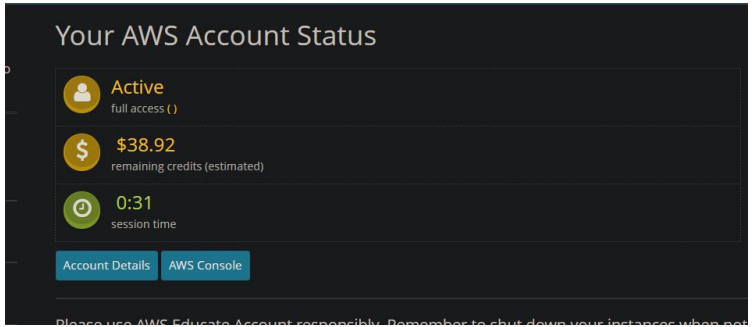
Windows | **Mac / Linux**

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/hy543-emr.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/hy543-emr.pem hadoop@ec2-3-81-28-91.compute-1.amazonaws.com
```

3. Type yes to dismiss the security warning.

Elastic Map Reduce - EMR



- Remember this ? ---------->
- Click on account details>AWS CLI>show
- Copy that in your ~/.aws/credentials file when you ssh to the cluster master server.

EMR - S3 connectivity

Once you have set up your credentials in .aws folder you can then access your S3 bucket contents. For example you can run:

```
aws s3 ls
```

Which will show you a list from your buckets or

```
aws s3 cp s3://bucket-name/your-file ~/ \
```

In order to copy files from your S3 bucket to your EC2 machine

Amazon Services Pricing

Amazon services are not free of charge. With your educate account you get 40$ to use, ration them well. Before starting VMs remember to choose a low cost location (e.d North Virginia). You can look up each service cost online:

EMR -> https://aws.amazon.com/emr/pricing/

S3 -> https://aws.amazon.com/s3/pricing/

Amazon Services Pricing

Explore various regions and find the minimum cost one!

Region: US East (Ohio)

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
a1.medium	1	N/A	2 GiB	EBS Only	$0.0255 per Hour
a1.large	2	N/A	4 GiB	EBS Only	$0.051 per Hour
a1.xlarge	4	N/A	8 GiB	EBS Only	$0.102 per Hour
a1.2xlarge	8	N/A	16 GiB	EBS Only	$0.204 per Hour
a1.4xlarge	16	N/A	32 GiB	EBS Only	$0.408 per Hour
a1.metal	16	N/A	32 GiB	EBS Only	$0.408 per Hour
t3.nano	2	Variable	0.5 GiB	EBS Only	$0.0052 per Hour
t3.micro	2	Variable	1 GiB	EBS Only	$0.0104 per Hour

Amazon Services Pricing

S3 charges you not only based on the storage occupy but also on the data transfers.

Amazon S3 pricing

Pay only for what you use. There is no minimum fee. There are four cost components to consider when deciding on which S3 storage class best fits your data profile – storage pricing, request and data retrieval pricing, data transfer and transfer acceleration pricing, and data management features pricing.

Storage | Requests and data retrievals | Data transfer | Management and replication

You pay for storing objects in your S3 buckets. The rate you're charged depends on your objects' size, how long you stored the objects during the month, and the storage class—S3 Standard, S3 Intelligent-Tiering, S3 Standard - Infrequent Access, S3 One Zone - Infrequent Access, S3 Glacier, and S3 Glacier Deep Archive, and Reduced Redundancy Storage (RRS). You pay a monthly monitoring and automation fee per object stored in the S3 Intelligent-Tiering storage class to monitor access patterns and move objects between access tiers in S3 Intelligent-Tiering.

There are per-request ingest fees when using PUT, COPY, or lifecycle rules to move data into any S3 storage class. Consider the ingest or transition cost before moving objects into any storage class. Estimate your costs using the AWS Simple Monthly Calculator.

Region: US East (Ohio)

	Storage pricing
S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data	
First 50 TB / Month	$0.023 per GB
Next 450 TB / Month	$0.022 per GB
Over 500 TB / Month	$0.021 per GB

Amazon Services Pricing

Keep an eye out of small details such as:

Amazon S3 pricing

Pay only for what you use. There is no minimum fee. There are four cost components to consider when deciding on which S3 storage class best fits your data profile – storage pricing, request and data retrieval pricing, data transfer and transfer acceleration pricing, and data management features pricing.

Storage | Requests and data retrievals | Data transfer | Management and replication

You pay for all bandwidth into and out of Amazon S3, except for the following:

- Data transferred in from the Internet.
- Data transferred out to an Amazon Elastic Compute Cloud (Amazon EC2) instance, when the instance is in the same AWS Region as the S3 bucket.
- Data transferred out to Amazon CloudFront (CloudFront).

The pricing below is based on data transferred "in" and "out" of Amazon S3 (over the public Internet)†††. Transfers between S3 buckets or from Amazon S3 to any service(s) within the same AWS Region are free. You also pay a fee for any data transferred using Amazon S3 Transfer Acceleration. Learn more about AWS Direct Connect pricing.

Region: US East (Ohio)

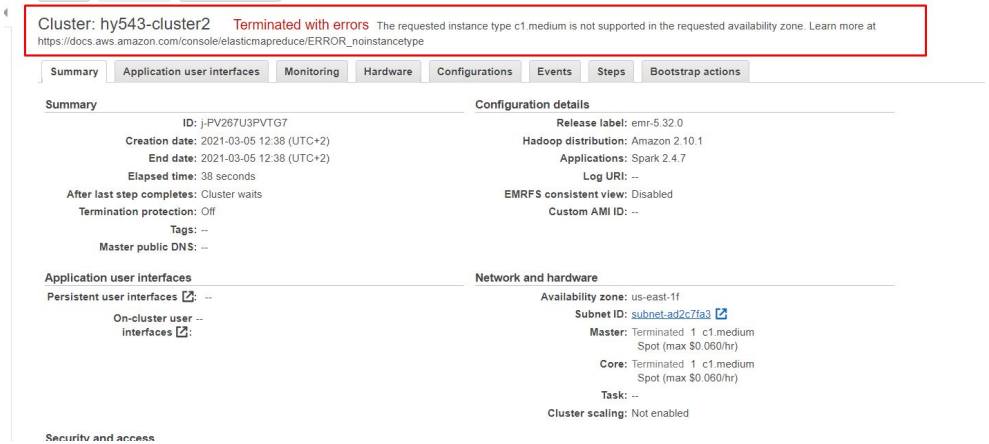
	Price
Data Transfer IN To Amazon S3 From Internet	
All data transfer in	$0.00 per GB
Data Transfer OUT From Amazon S3 To Internet	
Up to 1 GB / Month	$0.00 per GB

Amazon Services Pricing

As far as the pricing is concerned, remember to look up the up to date pricing of each service and don’t base your calculations only on what is shown in this tutorial !

Elastic Map Reduce - EMR

- Troubleshooting:
  - Your cluster may fail to start. This is most likely because the region you selected does not support the type of machine you selected.
    - For example, c1 medium is not available on us-east-1f
  - Follow the instructions given on the error and no further problems should occur.



Elastic Map Reduce - EMR

- Troubleshooting:
  - https://docs.aws.amazon.com/console/elasticmapreduce/ERROR_noinstancetype
- Amazon has very good troubleshooting guides and information. Don't forget to check carefully everything they throw at you.