

BU.330.740 Large Scale Computing with Hadoop

Lab 2. Set Up a Hadoop Cluster on Amazon AWS

Learning Goal: practice using AWS to create a big data processing environment

1. Download scripts.zip from Blackboard, unzip and store them in a local folder. Login into your AWS Educate account, then go to **AWS Account -> AWS Management Console**. In **Find Services**, search for S3 and enter/return.

The screenshot shows the AWS Management Console homepage. At the top, there's a navigation bar with the AWS logo, services dropdown, resource groups dropdown, a bell icon, user information (vocstartsoft/user326267=xu.m...), region selection (N. Virginia), and support dropdown. Below the navigation is the title "AWS Management Console". On the left, there's a sidebar with "AWS services" and a "Find Services" search bar (which is highlighted with a red box). The search bar has placeholder text "Example: Relational Database Service, database, RDS". Below the search bar are sections for "Recently visited services" (Support) and "All services". To the right of the sidebar is a "Build a solution" section with "Launch a virtual machine" and "Build a web app" buttons. Further right is a "Access resources on the go" section with a mobile device icon and text about the AWS Console Mobile App. At the bottom of the sidebar is a "Explore AWS" section with "Free Digital Training" and "Amazon DynamoDB" links. The main content area is mostly empty at the moment.

2. At S3 console, create an S3 bucket. You need to name your S3 bucket, and leave all other things unchanged. Because of Hadoop requirements, bucket and folder names that you use with Amazon EMR must contain only letters, numbers, periods (.), and hyphens (-). After you specify the name, do not go to the next steps but click **Create**. Note that your region may be different from mine.

The screenshot shows the AWS S3 console. On the left, there's a sidebar with "Amazon S3" and options like "Buckets", "Batch operations", "Access analyzer for S3", "Block public access (account settings)", and "Feature spotlight". The main area has a message about S3 Replication being temporarily re-enabled. Below that is a "S3 buckets" section with a search bar, a "Discover the console" link, and a "+ Create bucket" button (which is highlighted with a red box). It also shows "0 Buckets" and "0 Regions". A message at the bottom says "You do not have any buckets. Here is how to get started with Amazon S3.". To the right, there's a "Name and region" step of a wizard. It has a "Bucket name" field containing "330740lab2" (highlighted with a red box), a "Region" dropdown set to "US East (N. Virginia)", and a "Create" button at the bottom (also highlighted with a red box).

After you create the bucket, choose it from the list and then choose **Create folder**. Create two folders, input and scripts, one by one. After specify the name of the folder, do not change anything else, but choose **Save**.

Then click input folder, choose **Upload** in the upper left corner, add bigata.txt to the folder. Do not change other settings, but click **Upload** directly. Go back to your bucket, click scripts folder, and upload mapper.py and reducer.py to the folder.

3. Create an EC2 Key Pair. Go back to **AWS Management Console**, search for EC2 to open the EC2 console. In the navigation pane, choose **Key Pairs**.

Choose **Create key pair**.

Key pairs (1)			
		Actions	Create key pair
<input type="text"/> Filter key pairs			
Name	Fingerprint	ID	
<input type="checkbox"/> minghongxu	06:32:01:f3:98:27:70:30:cb:3d:a5:c2:d...	key-0a5af9d1d43f273e	

For **Name**, enter a descriptive name for the key pair. For **File format**, choose **pem**. Choose **Create key pair**.

Key pair
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name	<input type="text" value="minghonghadoop"/>
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.	
File format	<input checked="" type="radio"/> pem For use with OpenSSH
	<input type="radio"/> ppk For use with PuTTY
Tags (Optional)	No tags associated with the resource.
<input type="button" value="Add tag"/> You can add 50 more tags	
<input type="button" value="Cancel"/> <input style="background-color: #e57373; color: white; border-radius: 5px; padding: 5px; font-weight: bold; border: none;" type="button" value="Create key pair"/>	

The private key file is automatically downloaded by your browser. Save the private key file in a safe place.

4. Launch the EMR Cluster. Go back to **AWS Management Console**, search for EMR to open the EMR console. Choose **Create cluster**.

Amazon EMR

- Clusters
- Notebooks
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events
- Help
- What's new

Create Cluster - Quick Options Go to advanced options

General Configuration
Cluster name <input type="text" value="330740lab2"/>
Logging <input type="checkbox"/>
S3 folder <input type="text" value="s3://aws-logs-684763980438-us-east-1/elastictmapreduce/"/>
Launch mode <input checked="" type="radio"/> Cluster <input type="radio"/> Step execution
Software configuration
Release emr-5.31.0
Applications
<input checked="" type="radio"/> Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
<input type="radio"/> HBase: HBase 1.4.13, Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and ZooKeeper 3.4.14
<input type="radio"/> Presto: Presto 0.238.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore
<input type="radio"/> Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2
<input type="checkbox"/> Use AWS Glue Data Catalog for table metadata

On the **Create Cluster - Quick Options** page, accept the default values except for the following fields:

- Enter a **Cluster name** that helps you identify the cluster.
- Under **Hardware configuration**, choose m4.large as the **Instance type**.
- Under **Security and access**, choose the **EC2 key pair** that you created.

Choose **Create cluster**.

Hardware configuration

Instance type <input type="text" value="m4.large"/>	<small>The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. Learn more</small>
Number of instances <input type="text" value="3"/> (1 master and 2 core nodes)	
Cluster scaling <input type="checkbox"/> scale cluster nodes based on workload	

Security and access

EC2 key pair minghongxu [Create EC2 key pair](#)

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 [Edit](#)

EC2 instance profile EMR_EC2_DefaultRole [Edit](#)

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

Now the cluster is starting, which means AWS will provision your cluster (find available nodes to fulfill your job request).

5. While you are waiting for the cluster to be ready, change the security group rules to allow SSH connections to the cluster from your own IP address. Under **Security and access** choose the **Security groups for Master** link.

Network and hardware

Availability zone: us-east-1f
Subnet ID: [subnet-c46c93ca](#)
Master: Bootstrapping 1 m4.large
Core: Provisioning 2 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: minghongxu
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All [Change](#)
Security groups for Master: [sg-0ee2453f9e50186ee](#) (ElasticMapReduce-master)
Security groups for Core & Task: slave: [sg-0d635656424b70c72](#) (ElasticMapReduce-task)

Choose **ElasticMapReduce-master** first from the list.

Security Groups (2) Info				
Actions Create security group				
<input type="text"/> Filter security groups				
Name	Security group ID	Security group name	VPC ID	
<input type="checkbox"/>	sg-0d635656424b70c72	ElasticMapReduce-slave	vpc-8d2273f7	Edit
<input type="checkbox"/>	sg-0ee2453f9e50186ee	ElasticMapReduce-mas...	vpc-8d2273f7	Edit

Scroll down to choose **Edit inbound rules**.

Inbound rules [Outbound rules](#) [Tags](#)

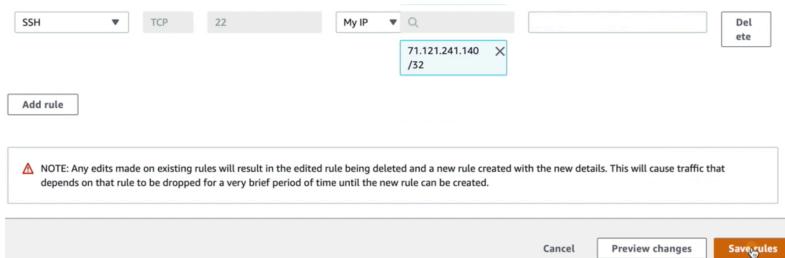
[Edit inbound rules](#)

Type	Protocol	Port range	Source	Description - optional
All TCP	TCP	0 - 65535	sg-0ee2453f9e50186ee (ElasticMapReduce-master)	-
All TCP	TCP	0 - 65535	sg-0d635656424b70c72 (ElasticMapReduce-slave)	-
SSH	TCP	22	71.121.241.140/32	-

Check for an inbound rule that allows public access with the following settings. If it exists, choose **Delete** to remove it.

- **Type:** SSH
- **Port:** 22
- **Source:** Custom 0.0.0.0/0

Scroll to the bottom of the list of rules and choose **Add Rule**. For **Type**, select **SSH**. This automatically enters **TCP** for **Protocol** and **22** for **Port Range**. For source, select **My IP**. Choose **Save**.



Go back to your EMR cluster and click **Security groups for Core & Task** link.

Network and hardware

Availability zone: us-east-1f
Subnet ID: [subnet-c46c93ca](#)
Master: Bootstrapping 1 m4.large
Core: Provisioning 2 m4.large
Task: --
Cluster scaling: Not enabled

Security and access

Key name: minghongxu
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All [Change](#)
Security groups for Master: [sg-0ee2453f9e50186ee](#) (ElasticMapReduce-master)
Security groups for Core & Task: [sg-0d635656424b70c72](#) (ElasticMapReduce-task)

Choose **ElasticMapReduce-slave** from the list and repeat the steps above to allow SSH client access to core and task nodes from your own IP address.

Security Groups (2) Info				
Actions Create security group				
<input type="text"/> Filter security groups				
<input type="button"/> Name <input type="button"/> Security group ID <input type="button"/> Security group name <input type="button"/> VPC ID				
<input type="checkbox"/>	-	sg-0d635656424b70c72	ElasticMapReduce-slave	vpc-8d2273f7
<input type="checkbox"/>	-	sg-0ee2453f9e50186ee	ElasticMapReduce-mas...	vpc-8d2273f7

- Under **Network and hardware**, find the **Master** and **Core** instance status. The status goes from **Provisioning** to **Bootstrapping** to **Waiting** during the cluster creation process.

Cluster: 330740lab2 Waiting Cluster ready after last step completed.

[Summary](#) [Application user interfaces](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

- After the cluster is ready, click **Steps**, and then **Add step**.

The screenshot shows the 'Steps' tab selected in the navigation bar. Below it, a table lists a single step:

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files
s-15DX3PS4EU3GW	Setup hadoop debugging	Completed	2020-10-24 11:39 (UTC-4)	10 seconds	View logs

- Choose **Streaming program** for **Step type**. Name your Streaming program; point **Mapper** to your mapper.py and **Reducer** to your reducer.py on your S3 instance; point **Input S3 location** to bigdata.txt; designate **Output S3 location** to a folder called output on your S3 instance. Note that this output folder should not be created before, and Hadoop will create it for you. You can use the folder icon (shown in red rectangle) to have access to folders on your own S3 instance. Click **Add** to execute the step.

The 'Add step' dialog box is shown with the following configuration:

- Step type:** Streaming program
- Name:** lab2wordcount
- Mapper:** s3://330740lab2/scripts/mapper.py
- Reducer:** s3://330740lab2/scripts/reducer.py
- Input S3 location:** s3://330740lab2/input/bigdata.txt
s3://<bucket-name>/<folder>/
- Output S3 location:** s3://330740lab2/output/ (highlighted with a red box and a red arrow pointing to the note)
- Arguments:** (empty)
- Action on failure:** Continue

Note: Type in output folder name directly, which does not exist in your S3 bucket

- Wait while the step is going from **Pending** to **Running** to **Complete**. You can always use **C** icon on the pages to refresh the status!

The screenshot shows the 'Steps' tab selected in the navigation bar. Below it, a table lists two steps:

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files
s-3S210BGUH3IYH	330740lab2	Pending	2020-10-24 11:44 (UTC-4)	1 minute	View logs
s-15DX3PS4EU3GW	Setup hadoop debugging	Completed	2020-10-24 11:39 (UTC-4)	10 seconds	View logs

10. After it's completed, you can check and download results from your S3 bucket -> output folder. Download these files to your local folder. And you can also check log files in your EMR cluster. Being able to read log files is a skill to acquire, especially if your task has a "failure" or "error" condition. You will get more practice as we go.

Name	Last modified	Size	Storage class
_SUCCESS	Jan 19, 2020 7:02:07 PM GMT-0500	0 B	Standard
part-00000	Jan 19, 2020 7:02:07 PM GMT-0500	508.0 B	Standard
part-00001	Jan 19, 2020 7:02:07 PM GMT-0500	595.0 B	Standard
part-00002	Jan 19, 2020 7:02:05 PM GMT-0500	545.0 B	Standard

11. DO NOT FORGET TO CLEAN UP RESOURCSES!! Otherwise, you will incur unnecessary charges to your account.

- Terminate the cluster
- Delete your S3 bucket
- Delete the log files bucket (we are not coming back to this instance)

Now that you are done with lab 2. Take some time to navigate your AWS Management Console. Double check that your S3 is empty, your EMR cluster is terminated, and no running instance in your EC2. And log out your AWS account every time you are done.

Reference:

<https://aws.amazon.com/getting-started/projects/analyze-big-data/>