LAB: EC2 ECS Cluster with Capacity Providers

You need:

An AWS Account

Duration of the Lab: 30 Minutes.

Difficulty: hard

Create an EC2 ECS Cluster

Create a new EC2 Launch Type Cluster:

Select cluster template

The following cluster templates are available to simplify cluster creation. Additional configuration and integrations can be added later.

Networking only

Resources to be created:

Cluster

VPC (optional)

Subnets (optional)

Powered by AWS Fargate

EC2 Windows + Networking

Resources to be created:

Cluster

VPC

Subnets

Auto Scaling group with Windows AMI

EC2 Linux + Networking

Resources to be created:

Cluster

VPC

Subnets

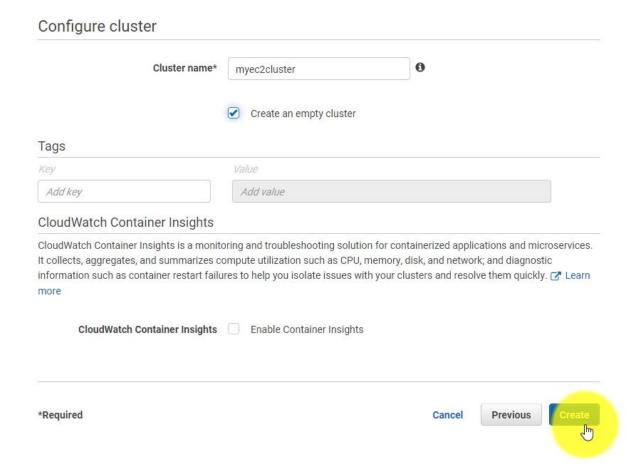
Auto Scaling group with Linux AMI

*Required

Cancel



Give the cluster a name, e.g. "myec2cluster", but this time choose an empty cluster:



Create an Auto Scaling Group

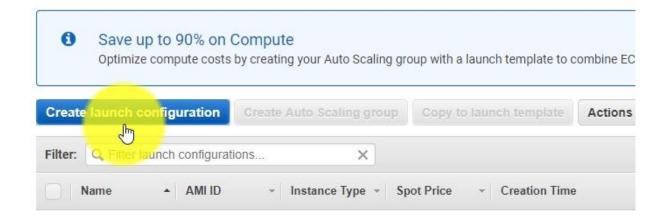
Before we can create a capacity provider, we need an ASG.

Create a new Launch Configuration

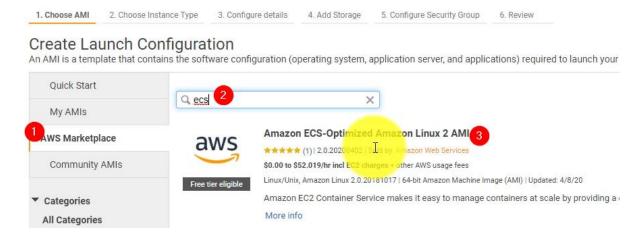
Open the EC2 Dashboard and find Launch Configurations in the left panel:



Create a new Launch Configuration:



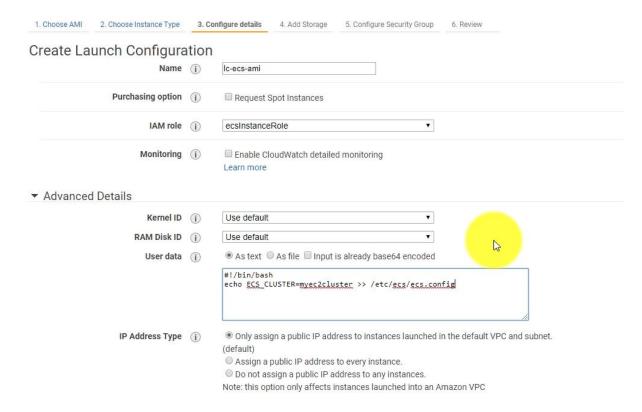
Find an Amazon ECS optimized AMI from the AWS Marketplace:



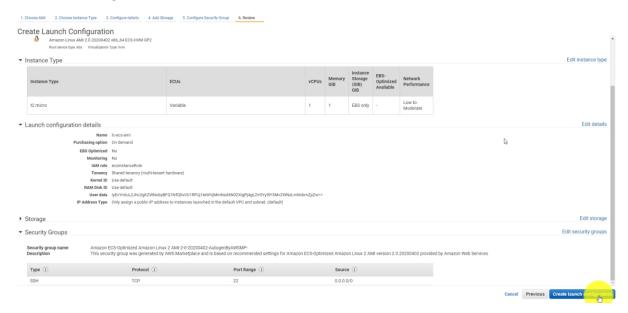
Select a t2.micro instance.

Give the Launch Configuration a name and select the ecsInstanceRole which was hopefully already created earlier by the ecs cluster lab. Then also add a user-data to register the EC2 instances in the right cluster:

```
#!/bin/bash
echo ECS_CLUSTER=myec2cluster >> /etc/ecs/ecs.config
```



Skip to review and Create launch configuration, you can have, but you don't need to have SSH access for the remaining lab:



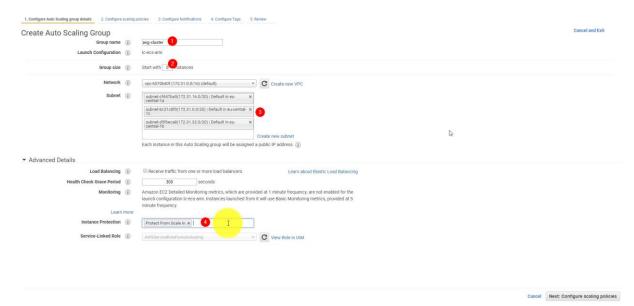
Then directly proceed to create an Auto Scaling grou:



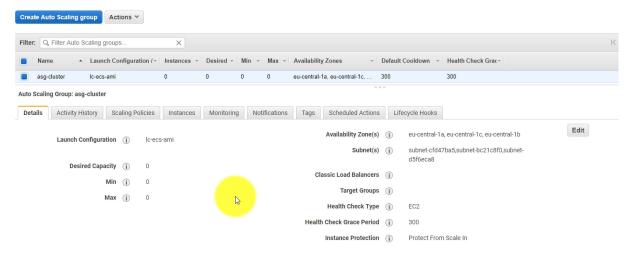
Create an Auto Scaling Group

Create the ASG:

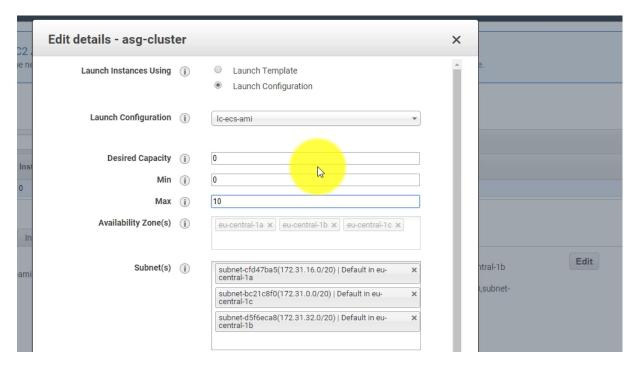
- 1. Give it a name
- 2. Start with 0 Instances
- 3. Select all available Subnets from your default VPC
- 4. Enable Instance Protection



Then proceed until the ASG is created.



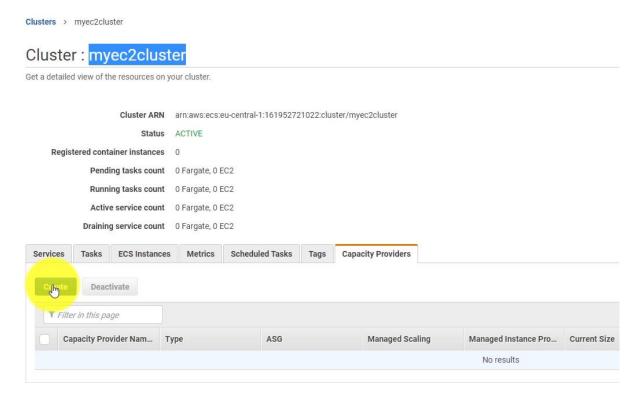
Now edit the ASG to set the maximum capacity to a higher number than 0:



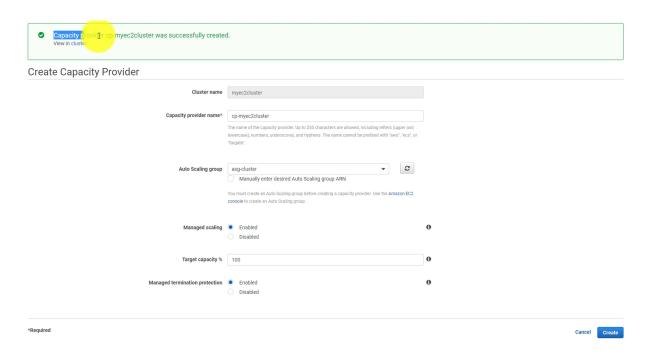
Now we need to attach a new Capacity Provider to the Auto Scaling Group.

Create a Capacity Provider

Go back to your ECS Cluster and open the Capacity Provider Tab:

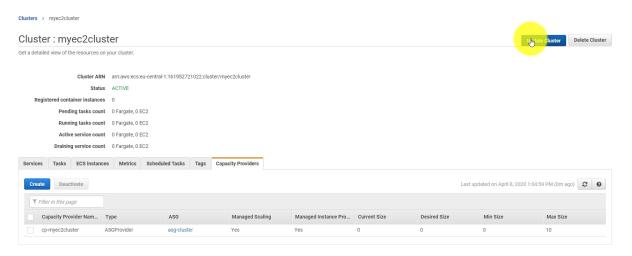


Give it a name, select the ASG you created in the previous step. Leave Managed Scaling to be enabled, set the target capacity to 100% and leave Managed termination protection also enabled:

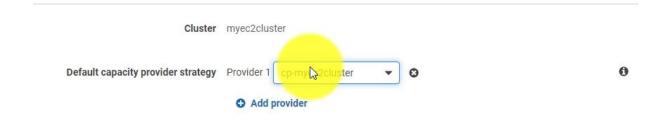


Attach the Capacity Provider to the Cluster

Now you need to update the Cluster and attach the Capacity Provider. Hit "Update Cluster"



Add the Capacity Provider we just created:



Save it and go back to the Cluster Dashboard to start a new service.

Start a Service

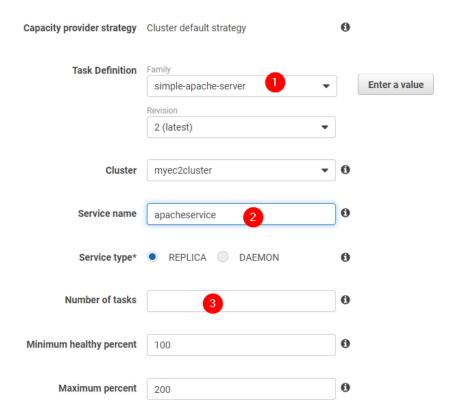
In this step we are starting a service to see how the Capacity Provider scales out and in.

Create a new Service:

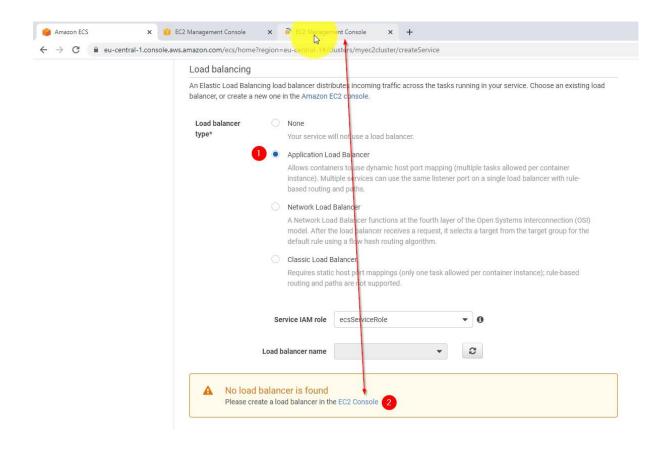
Select again the simple apache server task definition (1). Give it a service name (2). Launch 3 tasks initially (3). Leave the rest on default, then hit "Next step".

Configure service

A service lets you specify how many copies of your task definition to run and maintain in a cluster. You can optionally use an Elastic Load Balancing load balancer to distribute incoming traffic to containers in your service. Amazon ECS maintains that number of tasks and coordinates task scheduling with the load balancer. You can also optionally use Service Auto Scaling to adjust the number of tasks in your service.

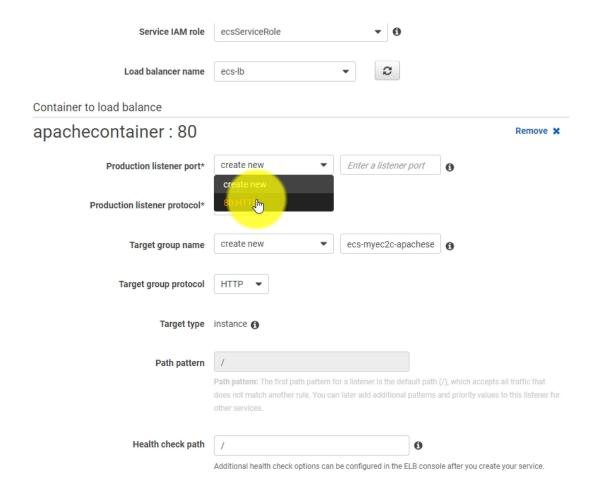


On the next page add an application load balancer. Again, go to the EC2 Dashboard in a new tab to create an application load balancer first:



Give the load balancer a name, and place it in all 3 AZ. Create a new target group type instance. Review -> create.

Then go back to your ECS Tab and reload the ALB Drop Down:

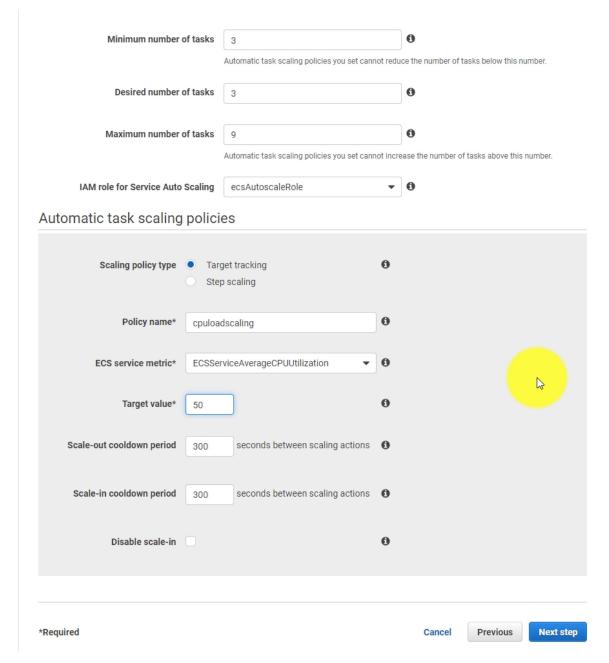


Remove the Service Discovery for this example and hit "Next Step".

On the next page add a Service Auto Scaling based on the CPU usage:

Minimum Tasks: 3, Desired: 3, Maximum: 9

Add a CPU Target Tracking with 50%:

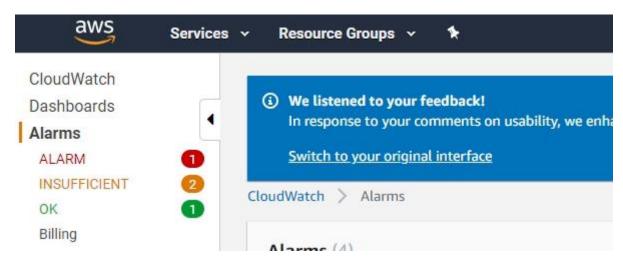


Then create the service.

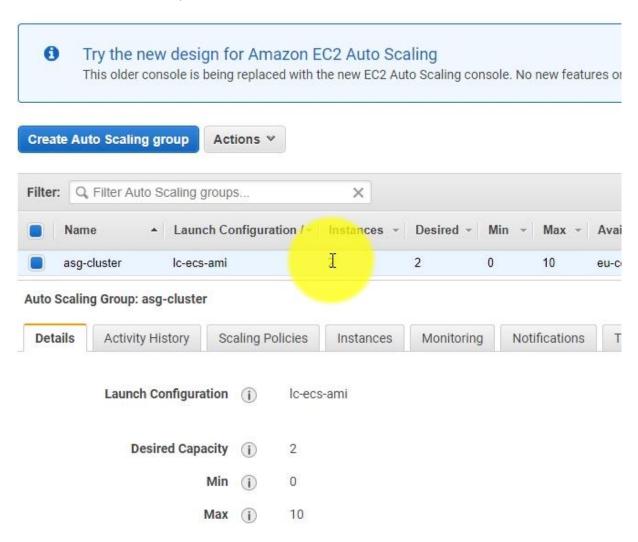
Watch the Scaling

The Capacity Provider needs a few minutes so the target capacity is tracked correctly. You can follow this also in the CloudWatch Alarms.

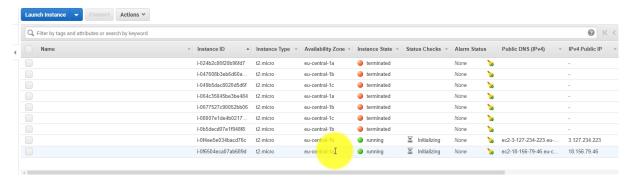
Open the CloudWatch Dashboard and have a look on the Alarms on the left side:



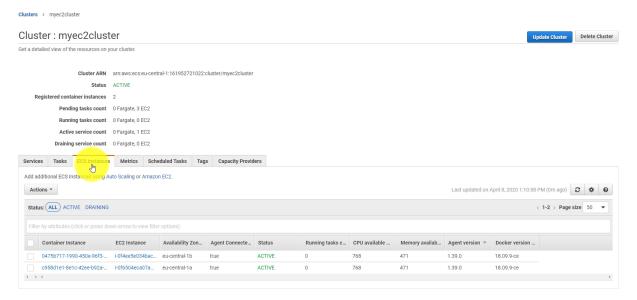
Have also a look at the ASG in the EC2 Dashboard. It will try to modify the desired capacity so that EC2 Instances are started by the ASG:



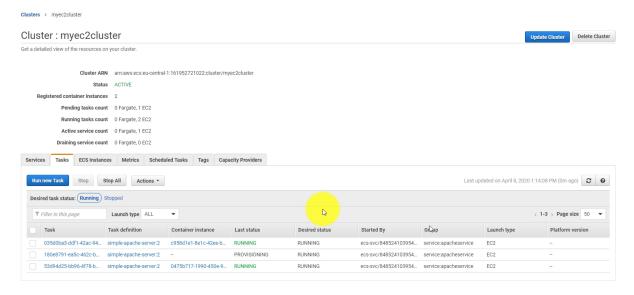
Have a look at the EC2 Dashboard, it automatically started the EC2 Instances through the ASG:



If you go back to the ECS Dashboard, to your cluster, and choose the ECS Instances, you will see the two EC2 Instances should have registered as Compute Resources for the Cluster:



If you check the service, then it should have started the Containers successfully:



Hint: If it does not start the containers successfully, have a look at the stopped containers failed message.

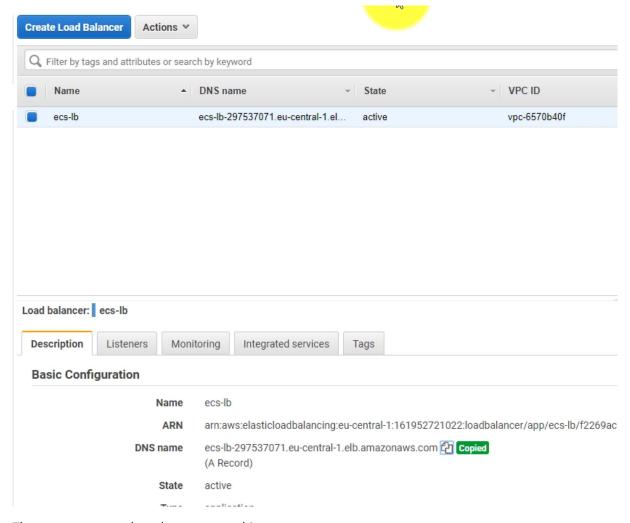
Check again the CloudWatch Alarms – they should be OK now. And if you drill down into the "AlarmHigh" you should see a spike where it went up to 200, then the ASG started the EC2 Instances and the Undercapacity was successfully mitigated by starting more compute resources, bringing the target capacity back to 100%:



Open the Website from Apache

Now, let's see the configuration for the Load Balancer. We can try and open up the DNS of the Load Balancer and see if we can reach out apache containers:

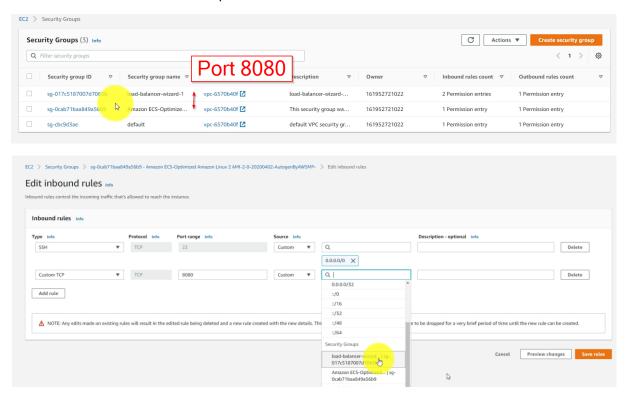
Open the EC2 Dashboard -> Load Balancers and copy the DNS of your ALB:



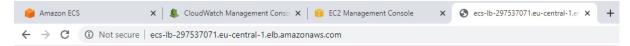
Then open a new tab and try to open this.

It times out, why? ... security groups!

We have to add an inbound rule to the Security group for our EC2 Instances, so that the Application Load Balancer is allowed to access port 8080 on the containers:

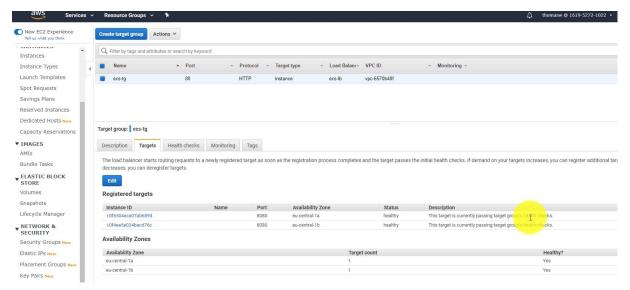


Try again to reload the browser tab, and it should work:

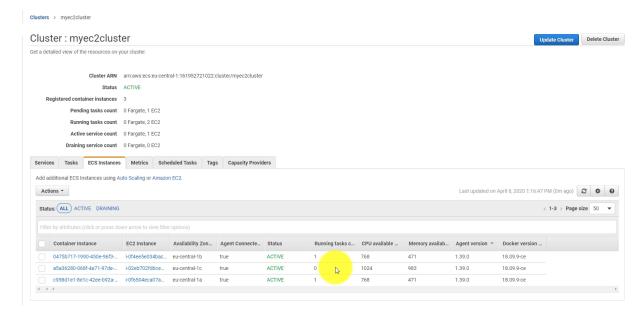


It works!

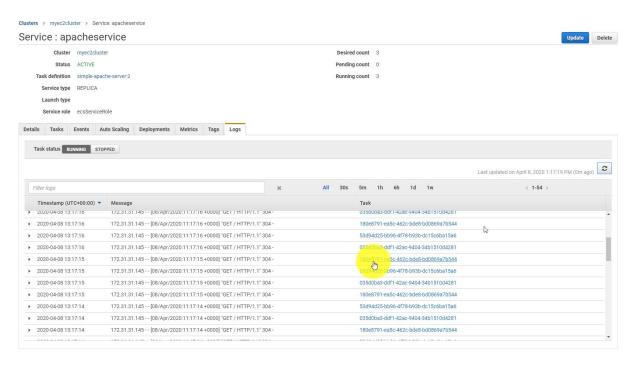
Also check out the Target Group of the Load Balancer and make sure the targets become healthy (might take a few minutes after updating the security groups):



It takes a while until the third instance is really running. The capacity provider is slowly ramping up the capacity until the desired capacity is met.



As soon as everything is running, make one last check at the logs if everything looks ok. Reload the apache website via the load balancer a few times and make sure the load is spread evenly across the instances:



Clean Up

To tear down everything:

- 1. Delete the Service
 - a. Potentially directly remove the tasks
- 2. The Capacity Provider *should* take care of scaling in, but it will take very long, so better directly delete the capacity provider
 - a. Update the Cluster -> Remove Capacity Provider
 - b. Deactivate the Capacity Provider
- 3. Open the Auto Scaling Group
 - a. Delete the ASG

- b. Delete the Launch Configuration
- 4. Load Balancer:
 - a. Delete the Application Load Balancer
 - b. Delete the Target Groups
- 5. Terminate any remaining EC2 Instances from the Lab manually
- 6. Remove the Security Groups
- 7. Also delete the cluster

Lab End	