Introduction

An Auto Scaling group is a representation of multiple Amazon EC2 instances that share similar characteristics and that are treated as a logical grouping for the purposes of instance scaling and management. For example, if a single application operates across multiple instances, you might want to increase or decrease the number of instances in that group to improve the performance of the application. You can use the Auto Scaling group to automatically scale the number of instances or maintain a fixed number of instances. You create Auto Scaling groups by defining the minimum, maximum, and the desired number of running EC2 instances the group must have at any given point of time.

An Auto Scaling group starts by launching the minimum number (or the desired number, if specified) of EC2 instances and then increases or decreases the number of running EC2 instances automatically according to the conditions that you define. Auto Scaling also maintains the current instance levels by conducting periodic health checks on all the instances within the Auto Scaling group. If an EC2 instance within the Auto Scaling group becomes unhealthy, Auto Scaling terminates the unhealthy instance and launches a new one to replace the unhealthy instance. This automatic scaling and maintenance of the instance in an Auto Scaling group is the core value of the Auto Scaling service. It's what puts the "elastic" in EC2.

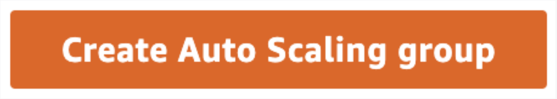
In this lab step, you will create an auto-scaling group using the launch template you created previously.

Instructions

1. Navigate to the EC2 service in the AWS Management Console.

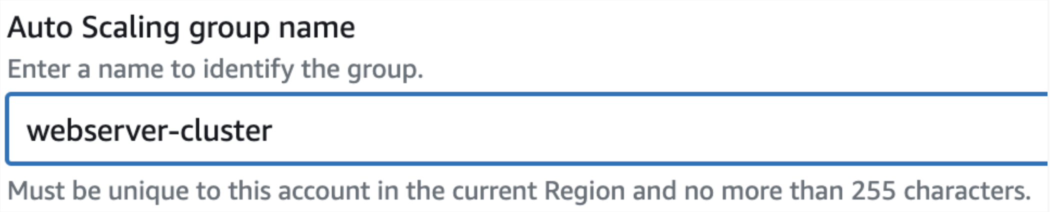
2. In the left-hand menu, click **Auto Scaling Groups**:

3. To start creating an auto-scaling group, click **Create Auto Scaling group**:



A multi-step wizard will start.

4. In the **Auto Scaling group name** text-box, enter *webserver-cluster*:



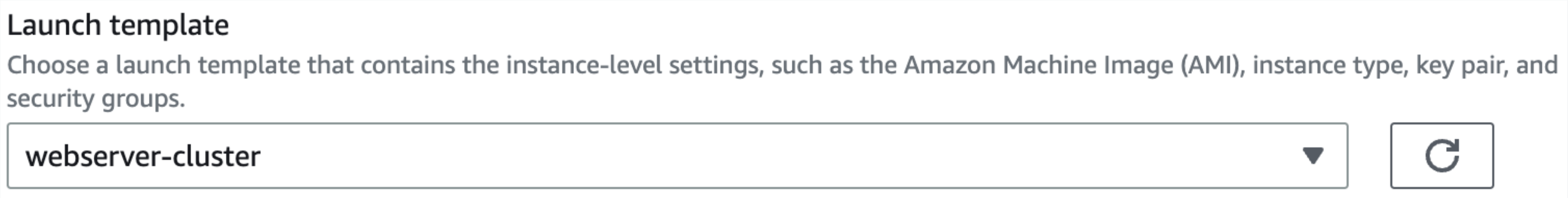
Auto Scaling group name

Enter a name to identify the group.

webserver-cluster

Must be unique to this account in the current Region and no more than 255 characters.

5. In the **Launch template** field, select **webserver-cluster**:



Launch template

Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.

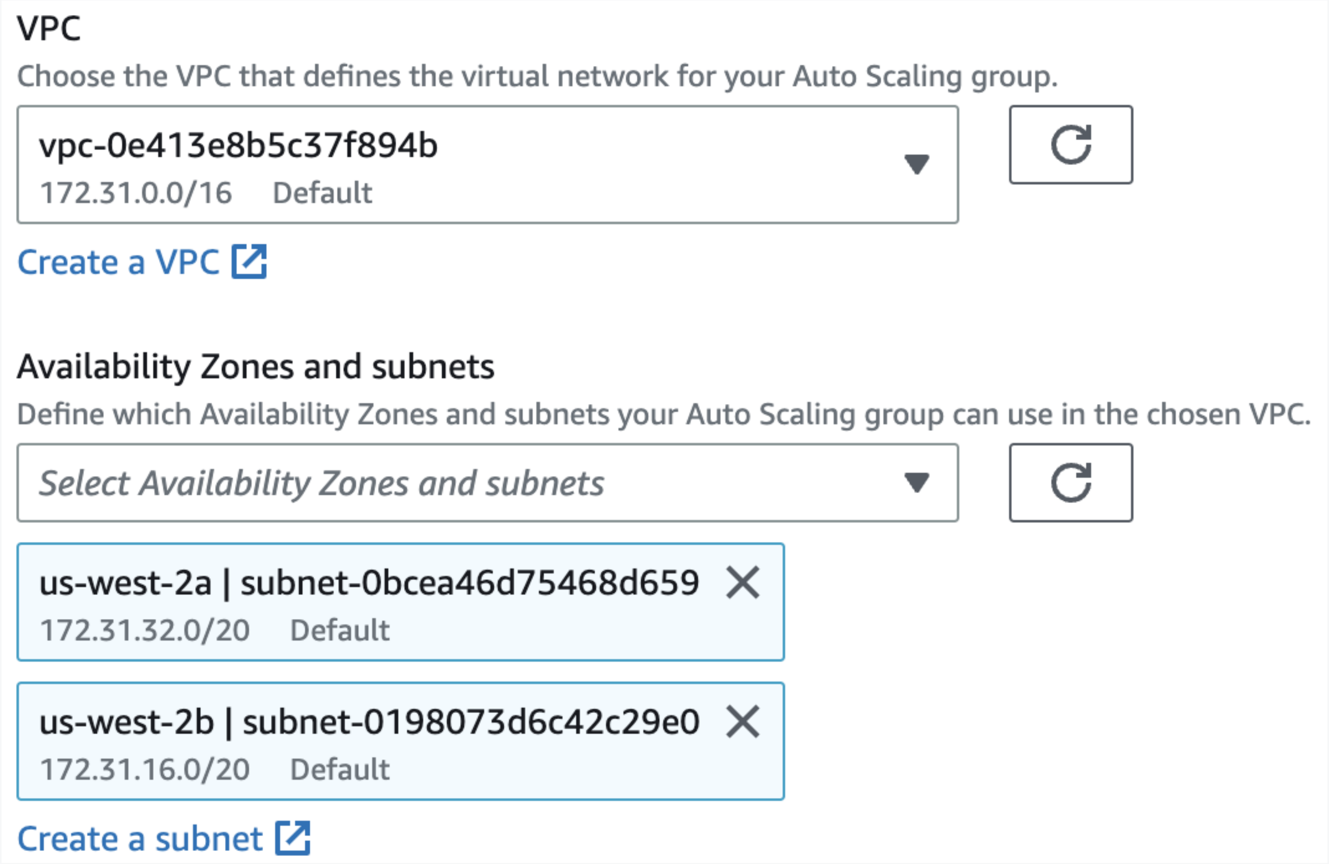
webserver-cluster

This is the launch template you created previously.

6. To advance to the next page of the wizard, click **Next**:

7. To configure your auto-scaling group, enter the following values:

- **VPC**: Select the only option available
- **Subnets**: us-west-2a and us-west-2b



VPC

Choose the VPC that defines the virtual network for your Auto Scaling group.

vpc-0e413e8b5c37f894b
172.31.0.0/16 Default

Create a VPC

Availability Zones and subnets

Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.

Select Availability Zones and subnets

us-west-2a | subnet-0bcea46d75468d659
172.31.32.0/20 Default

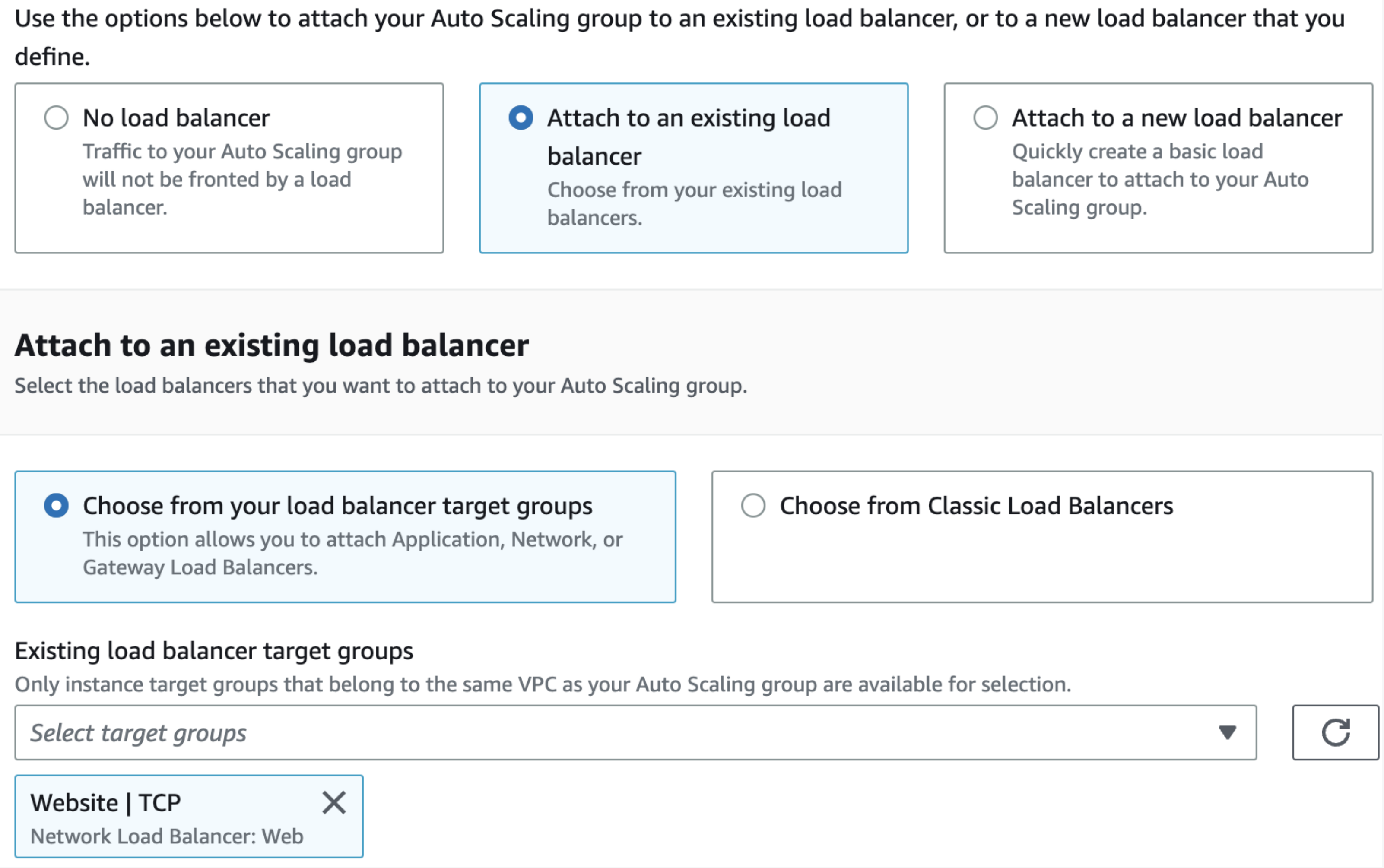
us-west-2b | subnet-0198073d6c42c29e0
172.31.16.0/20 Default

Create a subnet

8. To advance to the next page of the wizard, click **Next**:

9. Under the **Load balancing** section, enter the following leaving unspecified fields at their defaults:

- **Load balancing**: Select **Attach to an existing load balancer**
- **Choose a target group for your load balancer**: Select **Website**



Use the options below to attach your Auto Scaling group to an existing load balancer, or to a new load balancer that you define.

○ No load balancer
Traffic to your Auto Scaling group will not be fronted by a load balancer.

● Attach to an existing load balancer
Choose from your existing load balancers.

○ Attach to a new load balancer
Quickly create a basic load balancer to attach to your Auto Scaling group.

Attach to an existing load balancer

Select the load balancers that you want to attach to your Auto Scaling group.

● Choose from your load balancer target groups
This option allows you to attach Application, Network, or Gateway Load Balancers.

○ Choose from Classic Load Balancers

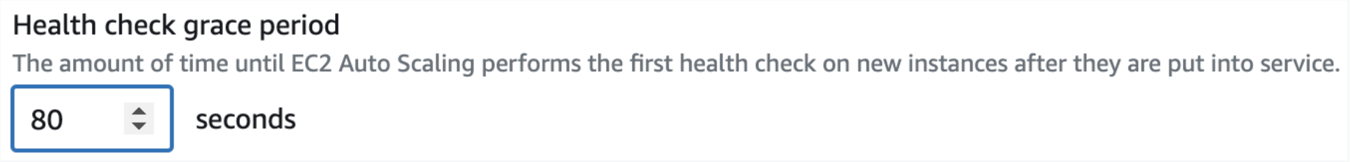
Existing load balancer target groups

Only instance target groups that belong to the same VPC as your Auto Scaling group are available for selection.

Select target groups

Website | TCP
Network Load Balancer: Web

10. Under **Health check grace period**, enter *80*:



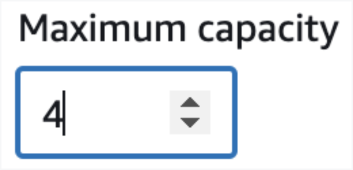
Health check grace period

The amount of time until EC2 Auto Scaling performs the first health check on new instances after they are put into service.

80 seconds

11. To advance to the next page of the wizard, click **Next**:

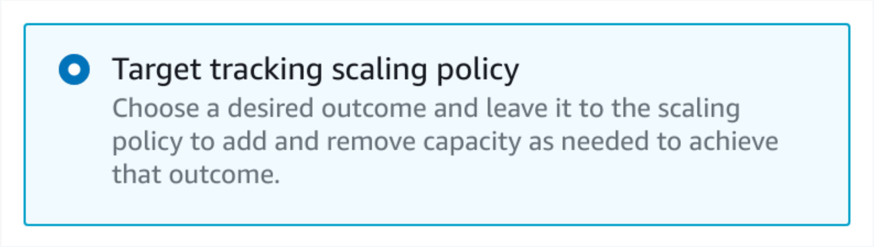
12. In the **Group size** section, in the **Maximum capacity** field, enter *4*:



Maximum capacity

4

13. Under **Scaling policies**, select **Target tracking scaling policy**:



● Target tracking scaling policy

Choose a desired outcome and leave it to the scaling policy to add and remove capacity as needed to achieve that outcome.

Take a look at the available options.

Auto-scaling groups allow you to scale out (add more instances) or scale in (remove instances) based on metrics such as:

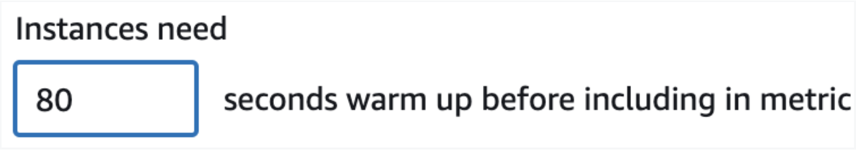
- Average CPU utilization

- Network traffic (ingress and egress)
- Application Load Balancer request counts

It's possible to create your own metrics and use these to decide when to scale, such metrics can be populated in CloudWatch directly from the application rather than inspecting the instance's resource usage. This is useful when the point that you need to scale at is determined by something unrelated to the instance, such as reaching a database connection limit or some other application-specific bottleneck.

By default, the **Metric type** is **Average CPU utilization**, and the **Target value** is 50 percent. Leave these values unchanged.

14. In the **Instances need** text-box, enter *80*:



Instances need

80 seconds warm up before including in metric

15. To advance to the next page of the wizard, click **Next**:

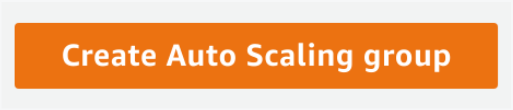
You will see a wizard step that allows you to configure notifications when scaling events occur. Notifications aren't used in this lab.

16. To advance to the **Add tags** page of the wizard, click **Next**.

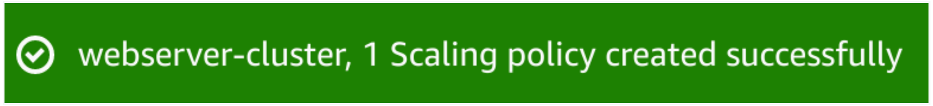
In a non-lab environment, it is best practice to tag resources when you create them so they can be easily filtered and discovered. Tags are not required in this lab.

17. To advance to the **Review** page of the wizard, click **Next**.

18. Check the details for accuracy and when ready, click **Create Auto Scaling group**:

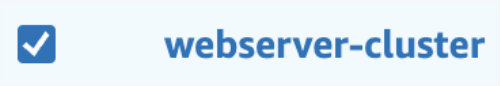


You will be taken to the auto-scaling group list page and you will see a notification that your group has been created:



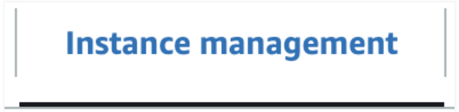
webserver-cluster, 1 Scaling policy created successfully

19. To see details about your auto-scaling group, check the box next to **webserver-cluster**:

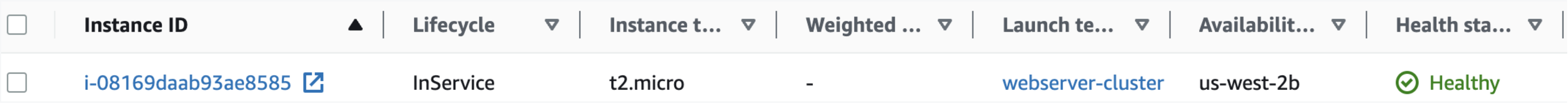


You will see some tabs appear under the list of groups.

20. To view instance information, click the **Instance management** tab:



You will see an instance:



□	Instance ID ▲	Lifecycle ▽	Instance t... ▽	Weighted ... ▽	Launch te... ▽	Availabilit... ▽	Health sta... ▽
□	i-08169daab93ae8585	InService	t2.micro	-	webserver-cluster	us-west-2b	✓ Healthy

If you don't see an instance that has a **LifeCycle** value of **InService**, wait a minute or two, and click the refresh button:

The auto-scaling group has started an instance, this is because the minimum capacity of the group is 1.

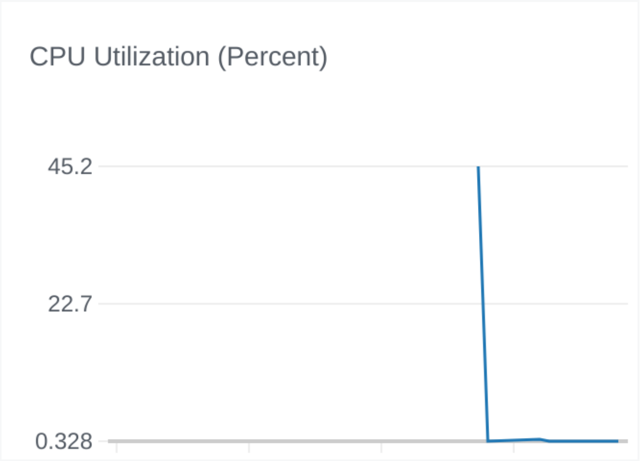
Because the webserver is not CPU-intensive and there is no load on the webserver, the high CPU alarm won't trigger. The number of instances will stay at 1 unless CPU utilization on the existing instance increases above 50 percent.

21. To see monitoring information about your auto-scaling group, click the **Monitoring** tab:

22. Under **CloudWatch monitoring details**, click the **EC2** tab:

Take a look at the different metrics recorded.

Focus on the **CPU Utilization (Percent)** graph:



Please note it may take several minutes for data points to display. Click the refresh button periodically to update the chart.

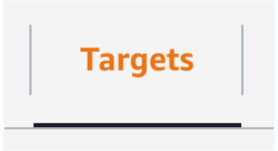
You will see the CPU utilization is currently near zero. You may see a higher value in the past, this spike in CPU utilization occurred when the instance was started and executed the commands you specified in the user data of your launch template.

You configured the auto-scaling group to have a minimum number of instances of one. This means that even though the CPU utilization is currently below 50 percent, the auto-scaling group is keeping this instance running instead of terminating it.

23. In the left-hand side menu, under **Load Balancing**, click **Target Groups**:

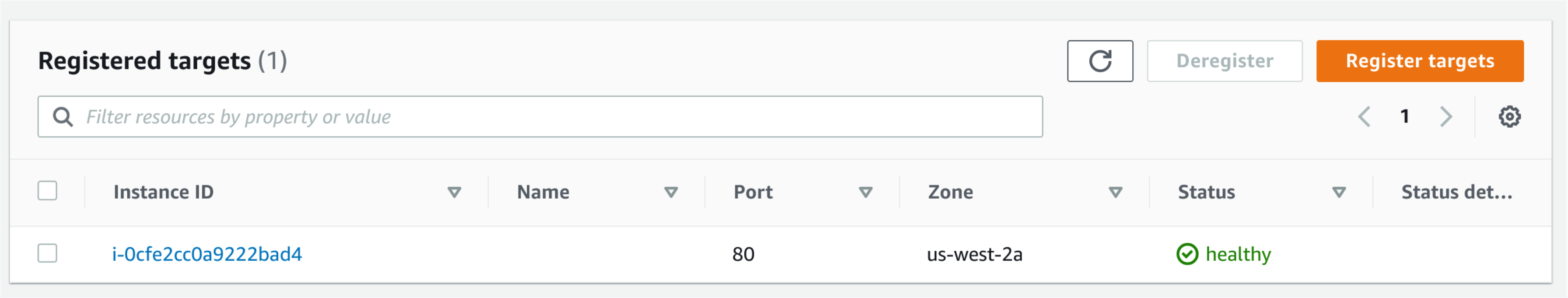
You will see the target group list page with one item, the **Website** target group you created earlier.

24. Select the **Website** target group, and then click the **Targets** tab:



Note you may need to click the refresh button to update the table if the instance is in the **initial** status while the load balancer waits for three successful health checks before assigning a healthy status.

Observe there is an instance added to the **Registered targets** and it is the same instance created by the Auto Scaling group. Also, notice the **Status** is **healthy** meaning the instance is reachable on TCP port 80 (HTTP). That means the launch template's user data script successfully completed to start the Apache webserver on the instance. Everything appears to be working. You will perform more thorough tests in the next lab step.



Registered targets (1)

□	Instance ID	Name	Port	Zone	Status	Status det...
□	i-0cfe2cc0a9222bad4		80	us-west-2a	✓ healthy	

If the instance doesn't become healthy after five minutes, a likely reason is the user data being absent or incorrect. Check the user data field on your Launch Template, if necessary, you can delete and re-create the Launch Template and Auto Scaling group.

Summary

In this lab step, you created an Auto Scaling group using a launch template. You defined a scaling policy to scale up or down based on the average CPU utilization of all the instances in the Auto Scaling group. The scaling policy makes use of CloudWatch metrics to trigger an alarm to cause a scale-in or scale-out event. You also configured the Auto Scaling group to automatically register its instances to a target group of a Network Load Balancer.