Introduction to Amazon EC2 Auto Scaling

Overview

This lab shows you how to use Auto Scaling to automatically launch Amazon EC2 instances in response to conditions that you specify. You will then test Auto Scaling by terminating a running instance and watching while Auto Scaling automatically creates a replacement instance.

Topics covered

By the end of this lab you will be able to:

- Create a Launch Template
- Create an Auto Scaling group
- Test the Auto Scaling Infrastructure
- View the results of the Auto Scaling launch

Prerequisites

This hands-on lab assumes that you are familiar with launching Amazon EC2 instances and have already created and utilized key pairs and security groups.

Introducing the Technologies

Amazon EC2 Auto Scaling

Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle your application's workload. You create collections of EC2 instances, called *Auto Scaling groups*.

- You can specify the **minimum number** of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes *below* this size.
- You can specify the **maximum number** of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes *above* this size.

If you specify a **desired capacity**, Auto Scaling ensures that your group always has a fixed number of instances.

If you specify **scaling policies**, then Auto Scaling will launch new instances or terminate existing instances when the demand on your application increases or decreases.

Auto Scaling only launches new instances or terminates existing instances. It does not Stop or Start instances.

Auto Scaling Group

Your EC2 instances are organized into **Auto Scaling groups** and are treated as a logical unit for the purposes of scaling and management. When you create an Auto Scaling group, you can specify its minimum, maximum and desired number of EC2 instances.

Launch Template

Launch templates enable you to store launch parameters so that you do not have to specify them every time you launch an instance. For example, a launch template can contain the AMI ID, instance type, and network settings that you typically use to launch instances. When you launch an instance using the Amazon EC2 console, an AWS SDK, or a command line tool, you can specify the launch template to use.

Scaling plans

A **Scaling Plan** tells Auto Scaling when and how to scale. Types of plans are:

- **Maintain current instance levels at all times:** Auto Scaling performs a periodic health check on running instances within an Auto Scaling group. When Auto Scaling finds an unhealthy instance, it terminates that instance and launches a new one.
- **Manual scaling:** Manual scaling is the most basic way to scale your resources. You specify a change in the maximum, minimum, or desired capacity of your Auto Scaling group. Auto Scaling then manages the process of creating or terminating instances to maintain the updated capacity.

- **Scale based on a schedule:** Sometimes you know exactly when you will need to increase or decrease the number of instances in your group, simply because that need arises on a predictable schedule. Scaling by schedule means that scaling actions are performed automatically as a function of time and date.
- **Scale based on demand:** Define parameters that control the Auto Scaling process. For example, you can create a policy that calls for enlarging your fleet of EC2 instances whenever the average CPU utilization rate stays above ninety percent for fifteen minutes. This is useful when you can define how you want to scale in response to changing conditions, but you don't know when those conditions will change. You can set up Auto Scaling to respond for you.

Pricing for Auto Scaling

There are no additional fees with Auto Scaling. You simply pay for the Amazon EC2 instances that it launches.

Task 1: Create a Launch Template

Before you can create an Auto Scaling group using a launch template, you must create a launch template that includes the parameters required to launch an EC2 instance, such as the ID of the Amazon Machine Image (AMI) and an instance type.

In this task you will create a launch template.

3. In the **Services** menu, click **EC2**.

If you see **New EC2 Experience** at the top-left of your screen, ensure **New EC2 Experience** is selected. This lab is designed to use the new EC2 Console.

4. In the left navigation pane, below **Instances**, select **Launch Templates**.

5. Select **Create launch template** then configure:

6. In the **Launch template name and description** section configure:

- **Launch template name:** Lab-template
- **Template version description:** version 1

You will be asked to select an **Amazon Machine Image (AMI)**, which is a template for the root volume of the instance and can contain an operating system, an application server and applications. You use an AMI to launch an **instance**, which is a copy of the AMI running as a virtual server in the cloud.

AMIs are available for various versions of Windows and Linux. In this lab, you will launch an instance running *Amazon Linux*.

7. For **AMI**, select *Amazon Linux 2 AMI*.

This is directly below, **Quick Start**.

8. For **Instance type**, select *t3.micro*.

When you launch an instance, the **instance type** determines the hardware allocated to your instance. Each instance type offers different compute, memory, and storage capabilities and are grouped in **instance families** based on these capabilities.

9. For **Security groups** select *MySecurityGroup*.

10. Scroll to the bottom of the screen, then click **Create launch template**

11. Click **View launch templates**

Task 2: Create an Auto Scaling Group

An Auto Scaling group contains a collection of EC2 instances that share similar characteristics and 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 the number of instances in that group to improve the performance of the application, or decrease the number of instances to reduce costs when demand is low. You can use the Auto Scaling group to scale the number of instances automatically based on criteria that you specify, or maintain a fixed number of instances even if an instance becomes unhealthy. This automatic scaling and maintaining the number of instances in an Auto Scaling group is the core functionality of the Amazon EC2 Auto Scaling service.

In this task, you will configure an Auto Scaling group with one instance.

12. In the left navigation pane, below **Auto Scaling**, click **Auto Scaling Groups**.

13. Click **Create an Auto Scaling group** then configure:

- **Auto Scaling group name:** Lab-Group
- **Launch template:** select the launch template that you created.
- Click **Next**

14. In the **Network** section, configure:

- **VPC:** *Lab VPC*
- **Subnets:** select both subnets

15. Click **Next**

16. On the **Configure advanced options** page, configure:

- **Health check grace period:** 60
- **Monitoring:** *Enable group metrics collection within CloudWatch*
- Click **Next**

By default, the health check grace period is set to 300. Since this is a lab environment, you have set it to 60 to avoid having to wait very long to see auto scaling perform the first health check.

17. On the **Configure group size and scaling policies** page, configure:

- **Minimum capacity:** 1
- **Maximum capacity:** 2
- Click **Next**

By default, Auto Scaling will **Keep this group at its initial size**. This means that it will keep one instance always operating. If the instance fails, it will be automatically replaced.

18. Click **Next** till you get to the **Review** page.

19. Click **Create Auto Scaling group**

Task 3: Verify your Auto Scaling group

Now that you have created your Auto Scaling group, you can verify that the group has launched your EC2 instance.

20. Click on your Auto Scaling Group.

Examine the **Group Details** to view information about the Auto Scaling group.

21. Click the **Activity** tab.

The Status column contains the current status of your instance. When your instance is launching, the status column shows *PreInService*. The status changes to *Successful* once the instance is launched. You can click the refresh button to see the current status of your instance.

22. Click the **Instance management** tab.

You can see that your Auto Scaling group has launched your EC2 instance and it is in the *InService* lifecycle state. The Health Status column shows the result of the EC2 instance health check on your instance.

23. Click the **Monitoring** tab. Here you can see monitoring related info for your Autoscaling group.

Task 3: Test Auto Scaling

Try the following experiment to learn more about Auto Scaling. The minimum size for your Auto Scaling group is 1 instance. Therefore, if you terminate the running instance, Auto Scaling must launch a new instance to replace it.

24. In the **Instance management** tab, click the **Instance ID**. It will look similar to: **i-1234abcd1234**.

You will be taken to the Amazon EC2 console the Instances page.

If you receive error stating "AWS Compute Optimizer:
This user is not authorized to call AWS Compute Optimizer", then kindly ignore the error.

25. Select your instance.

26. On the **Instance state** menu, select **Terminate instance**

27. In the **Terminate instance** dialogue box, click **Terminate**

The instance will change to *shutting-down*.

Wait until the Instance State changes to *terminated*. Click refresh occasionally to update the state.

28. In the left navigation pane, click **Auto Scaling Groups**.

29. Click on your auto scaling group.

30. Click the **Instance management** tab.

You will see the initial instance status as *Terminating*. Soon thereafter you will see a new instance appear with a status of *Pending* or *InService*.

31. Click the **Activity** tab.

All scaling activities are logged here. After the scaling activity starts. Click the to view entries for the launch and termination of the first instance and then an entry for the launch of the new instance.

Conclusion

Congratulations! You now know how to:

- Create a Launch Template
- Create an Auto Scaling group
- Test the Auto Scaling Infrastructure
- View the results of the Auto Scaling launch