Getting started with Amazon EC2 Auto Scaling

The content of this tutorial is based on information found at <https://docs.aws.amazon.com/autoscaling/ec2/userguide/GettingStartedTutorial.html>

# Overview

When you use Amazon EC2 Auto Scaling, you must use certain building blocks to get started. This tutorial walks you through the process for setting up building blocks to create a basic infrastructure for Amazon EC2 Auto Scaling.

Before you create an Auto Scaling group for use with your application, review your application thoroughly as it runs in the AWS Cloud. Consider the following:

* How many Availability Zones the Auto Scaling group should span.
* What existing resources can be used, such as security groups or Amazon Machine Images (AMIs).
* Whether you want to scale to increase or decrease capacity, or if you just want to ensure that a specific number of servers are always running. Keep in mind that Amazon EC2 Auto Scaling can do both simultaneously.
* What metrics have the most relevance to your application's performance.
* How long it takes to launch and configure a server.

The better you understand your application, the more effective you can make your Auto Scaling architecture.

# Step 1: Create a launch template

In this step, you sign in to the Amazon EC2 console with your AWS account credentials and create a launch template that specifies the type of EC2 instance that Amazon EC2 Auto Scaling creates for you. Include information such as the ID of the Amazon Machine Image (AMI) to use, the instance type, the key pair, and security groups.

|  |  |
| --- | --- |
| Open the EC2 console | <https://awsacademy.instructure.com/courses> |
| Normally at this point we would choose the region our systems would be in. The student account is locked to a region (in our case, North Virginia), so we will skip this step.  The selection box is in the top right of the screen. |  |
| In the left navigation pane, choose **Launch Templates**, and then choose **Create launch template.** |  |
| For **Launch template name**, enter  “my-template-for-auto-scaling” |  |
| Under **Auto Scaling guidance**, select the check box. |  |
| For **Amazon machine image (AMI)**, choose a version of Amazon Linux 2 (HVM) from the **Quick Start** list. The AMI serves as a basic configuration template for your instances. |  |
| Under **Instance Type** select t2.micro (free tier eligible) |  |
| Under **Key pair**, select “Create new key pair” |  |
| Call the key “Session12Key”  Make the key a .ppk key to save using PuTTYGen later  Save your key somewhere safe |  |
| Within **Network settings leave the Subnet as Don’t include in launch template** |  |
| Under **security groups**, use the **default security group** |  |
| Choose **Create launch template**. |  |
| On the confirmation page, choose **Create Auto Scaling group**. |  |

# Step 2: Create a single-instance Auto Scaling group

Now use Amazon EC2 Auto Scaling to create an Auto Scaling group and add the launch template or launch configuration to the group. Also include information such as the VPC subnets for the instances.

|  |  |
| --- | --- |
| You will now be returned to the EC2 Dashboard |  |
| Scroll to the bottom and find **Auto Scaling** and select **Auto Scaling Groups** |  |
| Press Create Auto Scaling group |  |
| On the **Choose launch template or configuration** page, for **Auto Scaling group name**, enter  **my-first-asg**  Click “Next” |  |
| Under **Launch template**, press the drop-down arrow and select the template we created in Step 1  Press Next |  |
| Under “Network” make sure the “Default VPC” is selected. |  |
| For high availability, we will locate our machine in 2 datacenters in the North Virginia availability zone.  Under “Subnets” select:  Us-east-1a  Us-east-1b |  |
| We can keep defaults for the rest of the task.  Select “Skip to review” |  |
| For now, the default size of the group is “1 computer”  Select “Create Auto Scaling Group” |  |

# Step 3: Verify your Auto Scaling group

Now that you have created an Auto Scaling group, you are ready to verify that the group has launched an EC2 instance.

|  |  |
| --- | --- |
| You should now be on the “**Auto scaling groups**” page.  To return here, select **Auto Scaling > Auto Scaling Groups** |  |
| Under “Status” it will say “**Updating Capacity**”. It may stay here for a few minutes.  Hit the refresh button after a few minutes, and the message will disappear. |  |
| Press the checkbox to select our Auto Scaling group |  |
| Here we can see we should **have 1 machine always running**:  Desired capacity: The normal number of machines we will have running  Minimum capacity: Running machines should never fall below this number.  Maximum capacity: The maximum amount we will scale up to |  |
| Under the Activity tab, look at the “Activity History”.  So far all we have in our log is the machine being created.  Note the text: The machine was created in response to a difference between the desired capacity, 1, and the actual capacity, 0 (i.e. none were running at the time) |  |
| Under “Instance Management” you should see the status of the instance. It should be “Healthy”  The lifecycle should say “In-service” which means the machine is currently in use. |  |

# Step 4: Terminate an instance in your Auto Scaling group

Use these steps to learn more about how Amazon EC2 Auto Scaling works, specifically, how it launches new instances when necessary. The minimum size for the Auto Scaling group that you created in this tutorial is one instance. Therefore, if you terminate that running instance, Amazon EC2 Auto Scaling must launch a new instance to replace it.

|  |  |
| --- | --- |
| If not done already, return to the “**Auto Scaling Groups**” section of the console |  |
| Select the check box next to your Auto Scaling group. |  |
| On the **Instance management** tab, under **Instances**, select the ID of the instance.  This takes you to the **Instances** page in the Amazon EC2 console, where you can terminate the instance. |  |
| Choose **Actions**, **Instance State**, **Terminate**.  When prompted for confirmation, choose **Yes, Terminate**. |  |
| On the navigation pane, under **AUTO SCALING**, choose **Auto Scaling Groups**. Select your Auto Scaling group and choose the **Activity** tab.  The default cooldown for the Auto Scaling group is 300 seconds (5 minutes), so it takes about 5 minutes until you see the scaling activity. |  |
| In the activity history, when the scaling activity starts, you see an entry for the termination of the first instance and an entry for the launch of a new instance. |  |
| On the **Instance management** tab, the **Instances** section shows the new instance only. |  |
| In the navigation menu, select “Instances” |  |
| Here, you can see the terminated instance, and the new instance started up to replace it. |  |
|  |  |