**Placement Empowerment Program**

***Cloud Computing and DevOps Centre***

**Implement Auto-scaling in the Cloud**

Set up an auto-scaling group for your cloud VMs to handle variable workloads.

Name: Madhu Smitha Department: CSE



**Introduction and Overview**

Auto-scaling in AWS is a powerful feature that enables cloud environments to automatically adjust computing resources based on dynamic workloads. This ensures applications maintain performance while optimizing costs. By automatically increasing or decreasing the number of EC2 instances, AWS Auto Scaling helps achieve high availability and efficient resource utilization.

**Objectives**

* To set up an Auto Scaling Group (ASG) in AWS.
* To automate the scaling of EC2 instances based on workload demands.
* To improve application availability and reliability.
* To optimize cloud resource utilization and reduce costs.

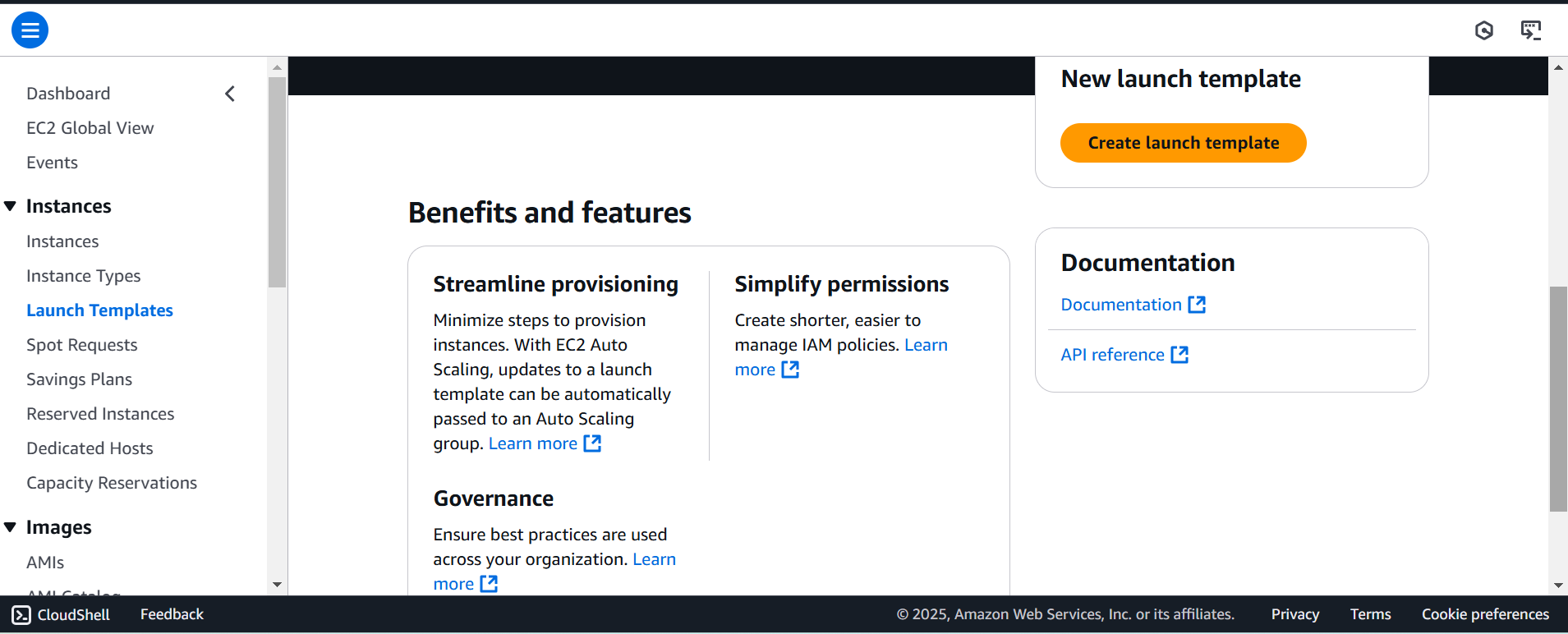
**Importance**

* **Cost Efficiency:** Reduces costs by scaling down unused resources.
* **High Availability:** Maintains application uptime even during traffic spikes.
* **Performance Optimization:** Ensures consistent performance under varying loads.
* **Automation:** Eliminates manual intervention for scaling resources.

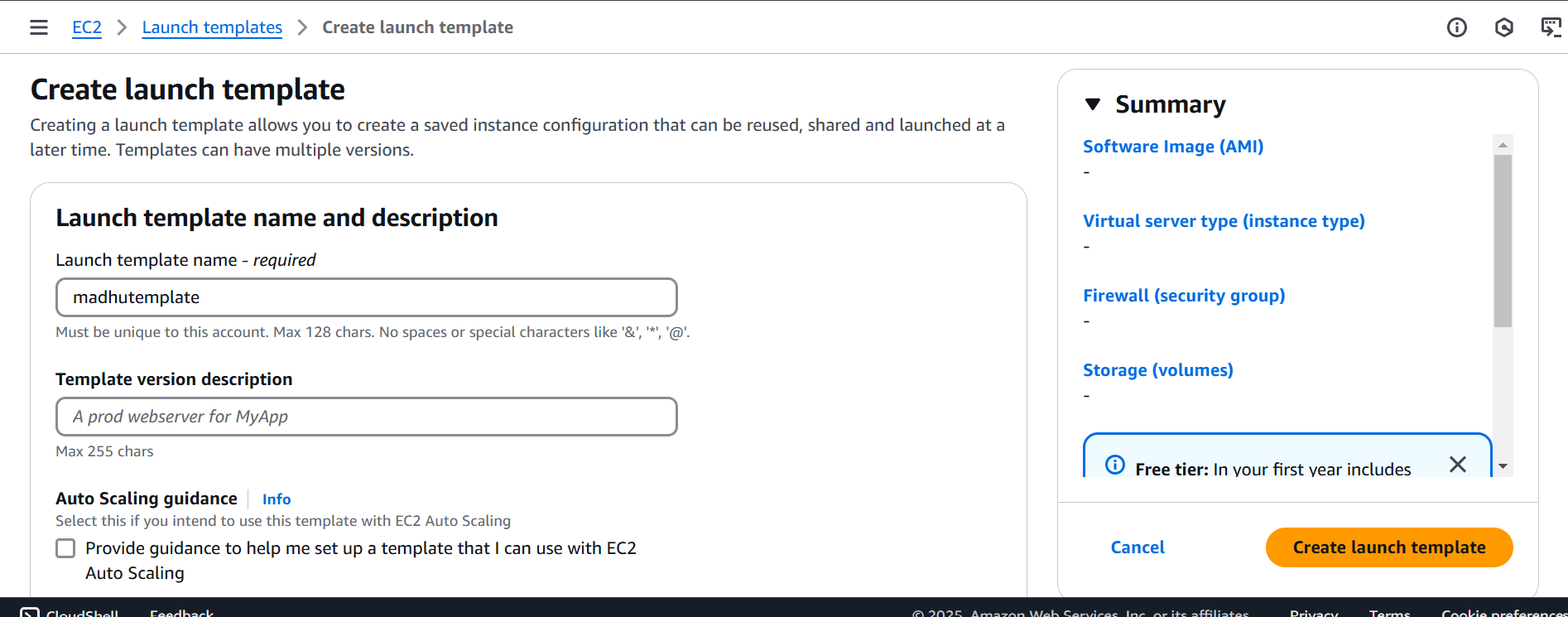
**STEPS:**

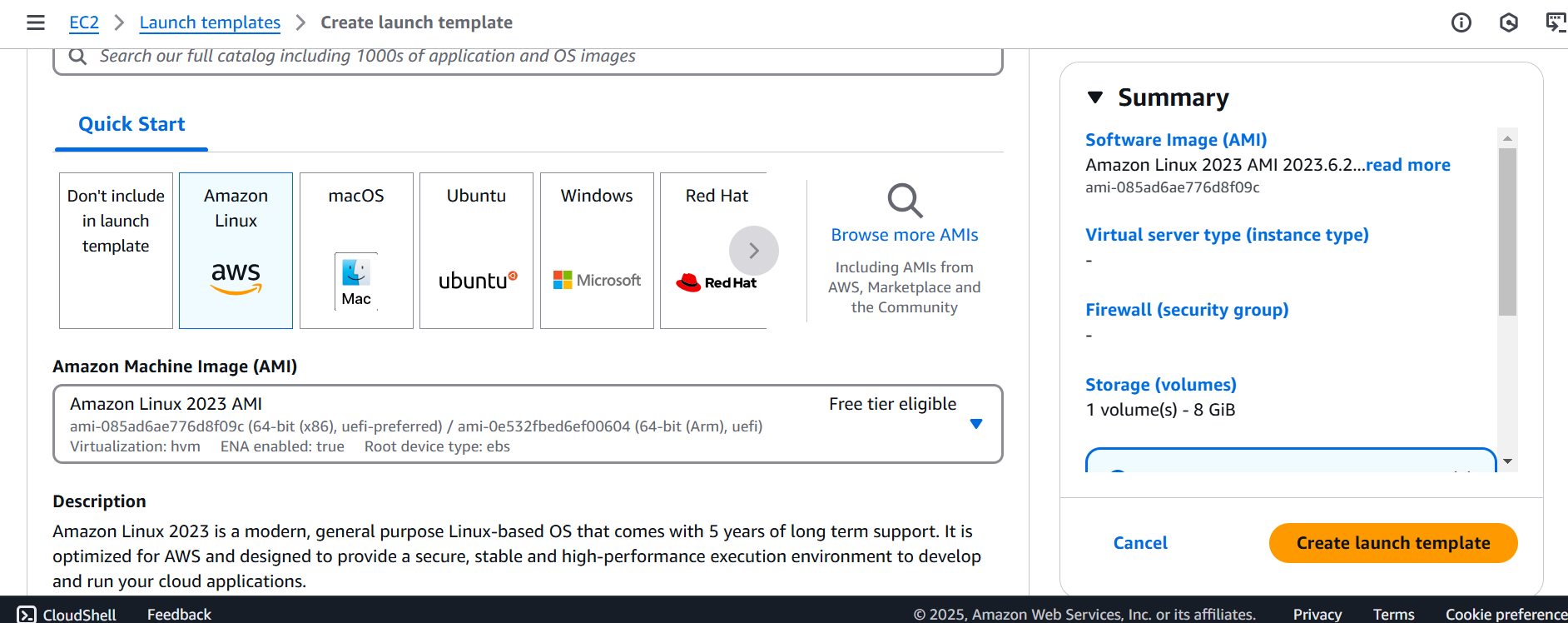
**STEP 1: Create a Launch Template**

* Go to the AWS Management Console → EC2 → Launch Templates → Create Launch Template.

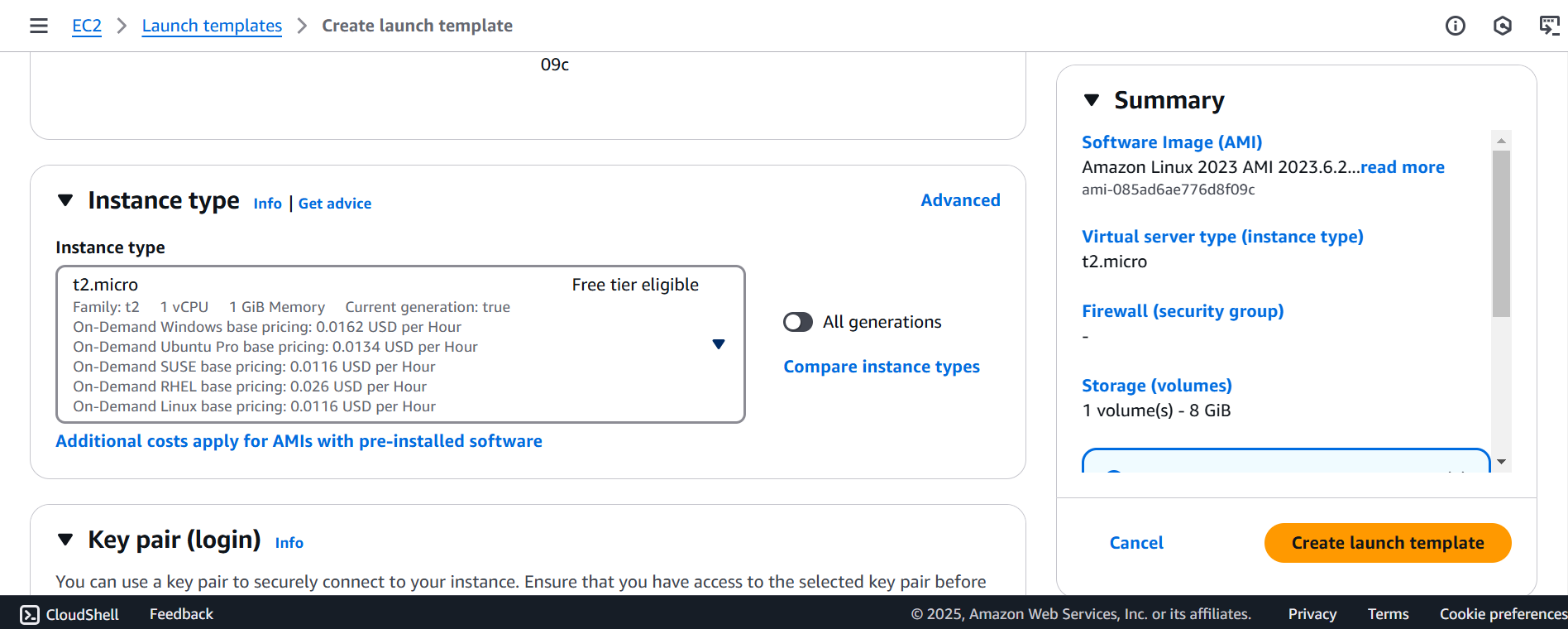


* Provide a name for your template.
* Choose the Amazon Linux 2 AMI.





* Select the instance type t2.micro.
* Configure key pair and security group (allow HTTP, HTTPS, and SSH).



* Do not add a subnet during this process.
* Under Advanced Details, add the following User Data script:

#!/bin/bash

yum update -y

yum install -y httpd

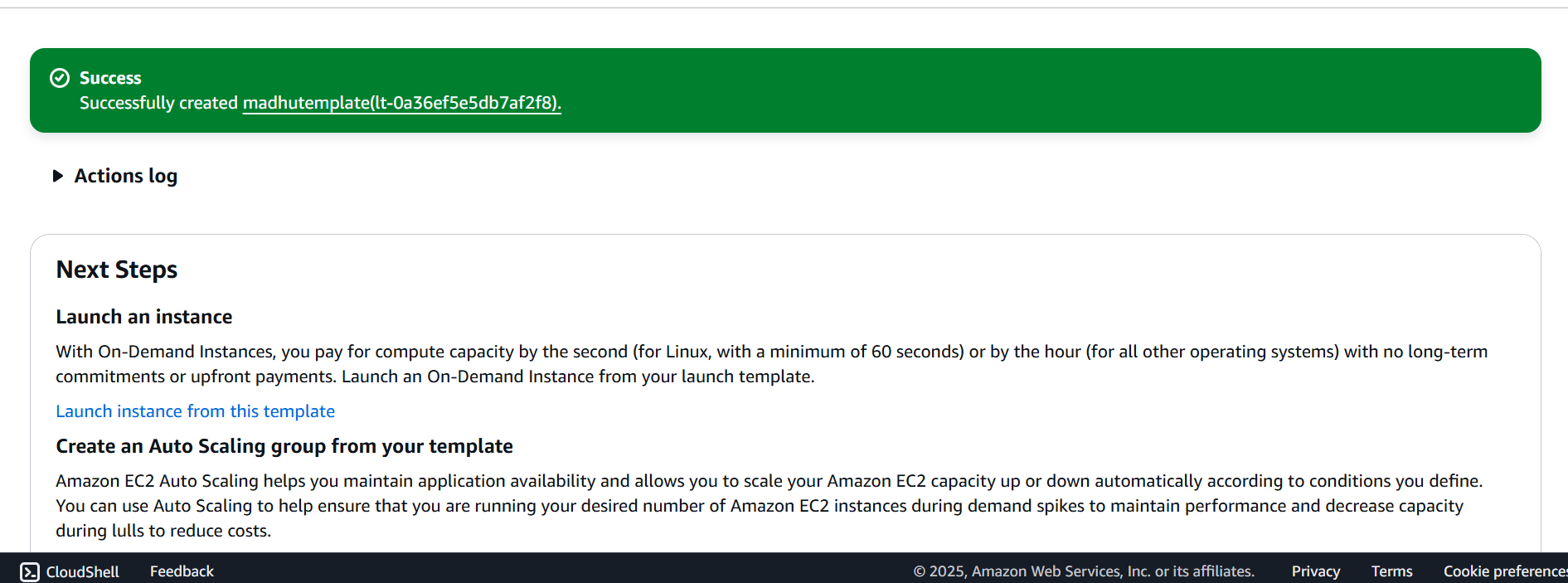
systemctl start httpd

systemctl enable httpd

echo "<h1>Hello World from $(hostname -f)</h1>" > /var/www/html/index.html

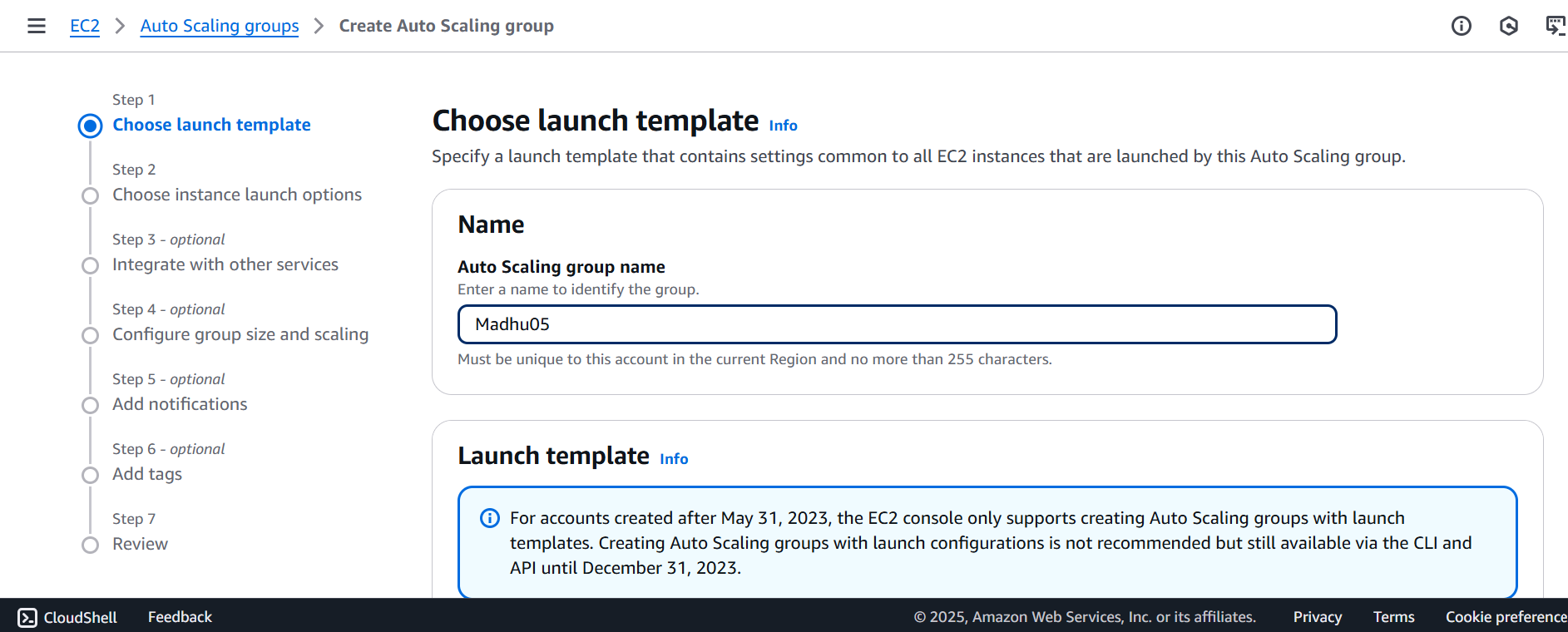


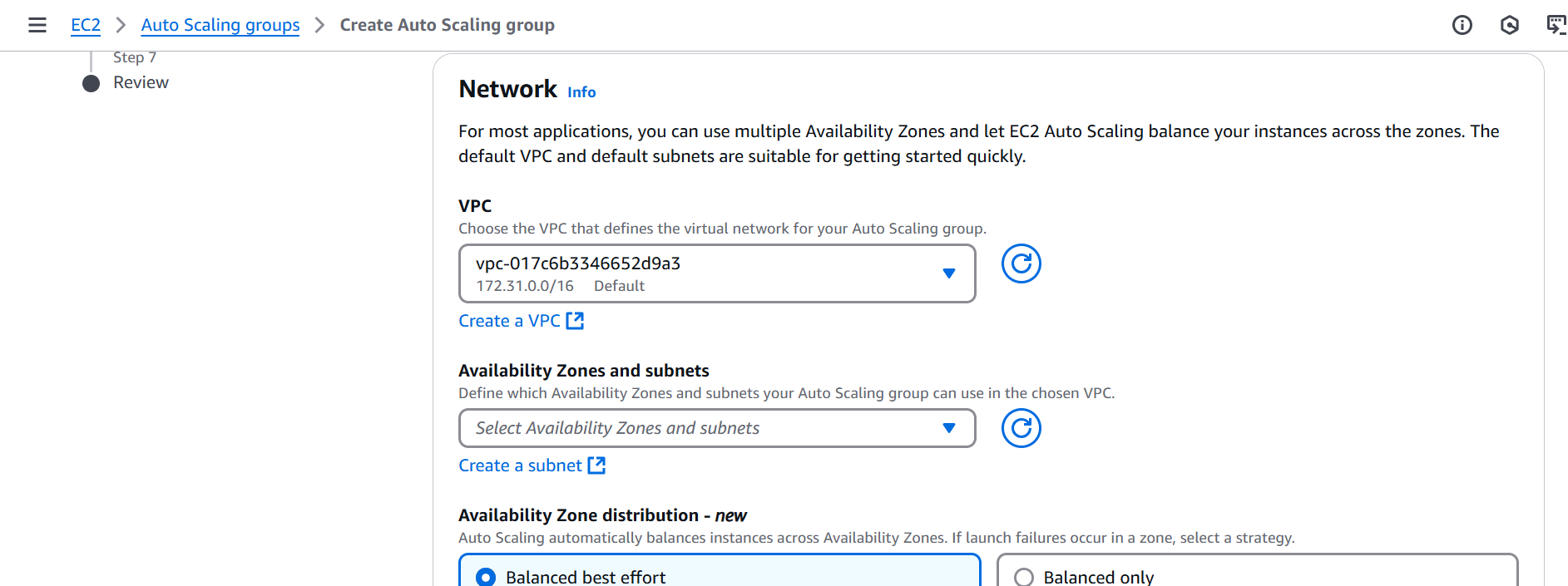
* Review and create the launch template.

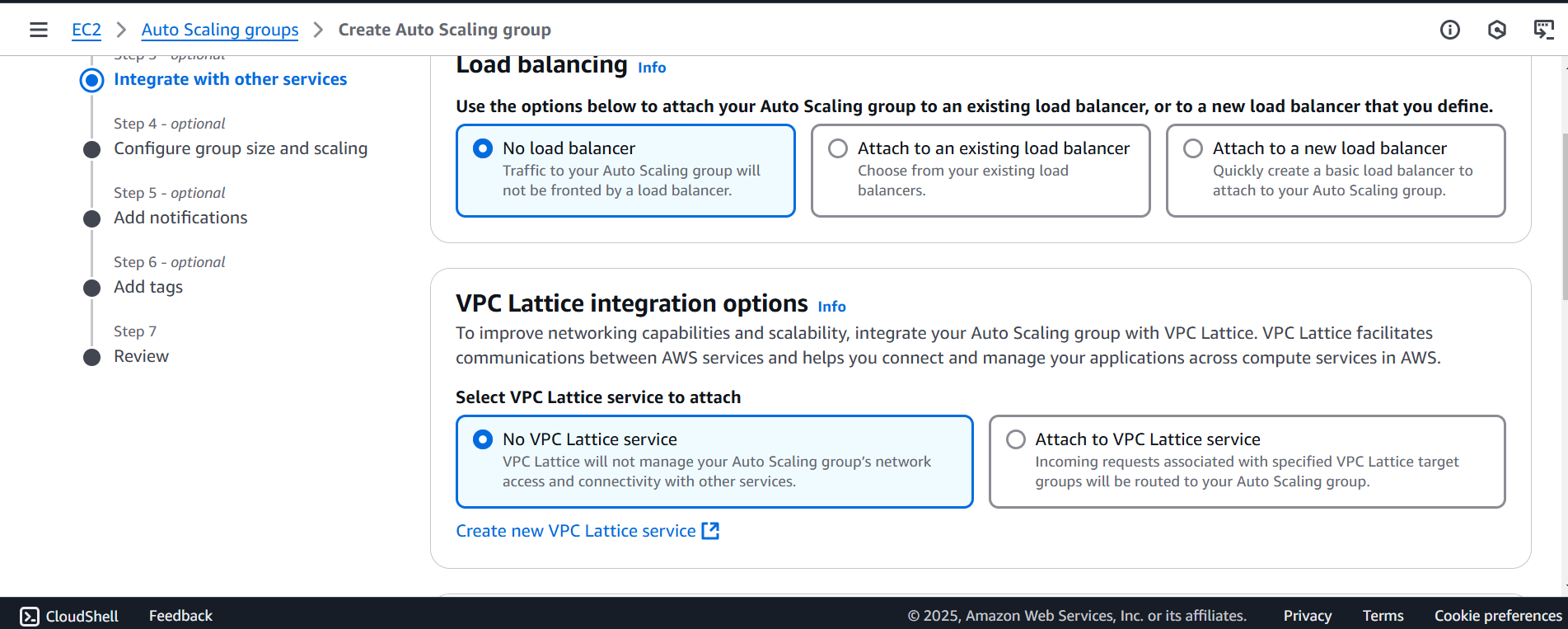


**STEP 2: Create an Auto Scaling Group (ASG)**

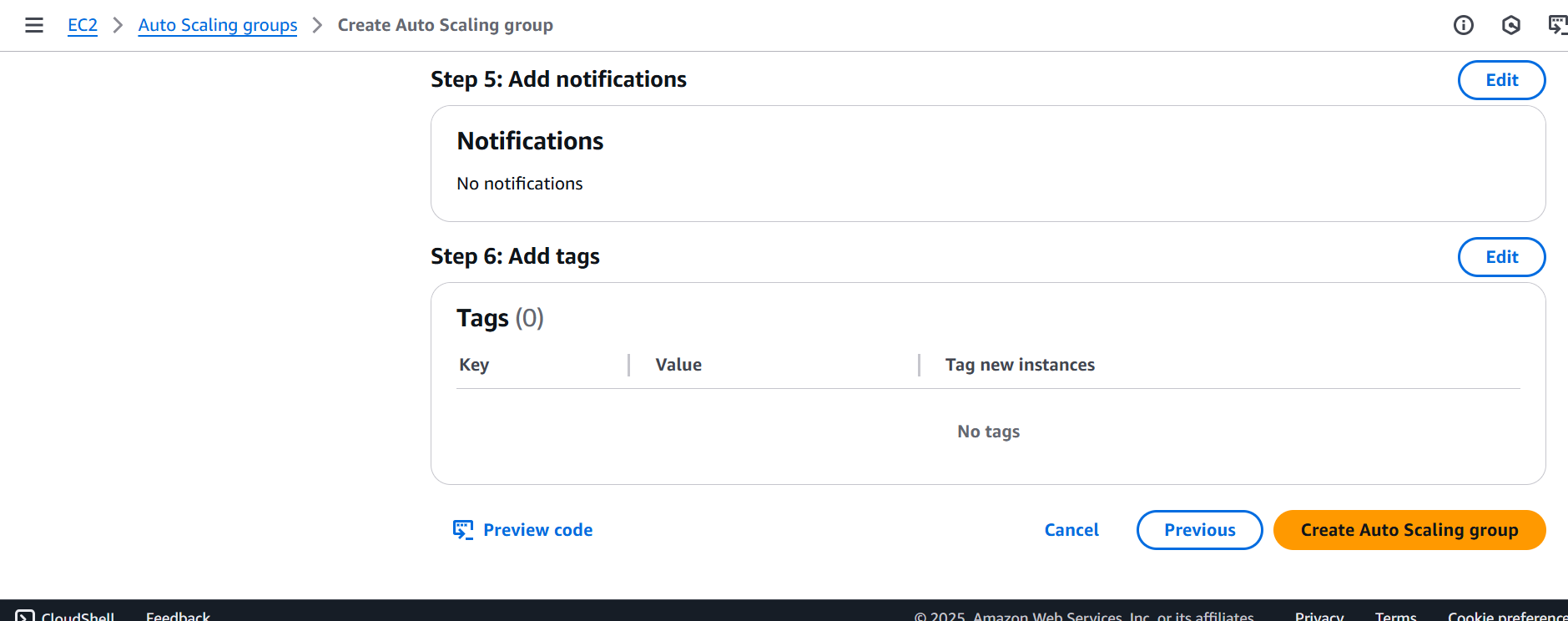
* Go to Auto Scaling Groups → Create Auto Scaling Group.



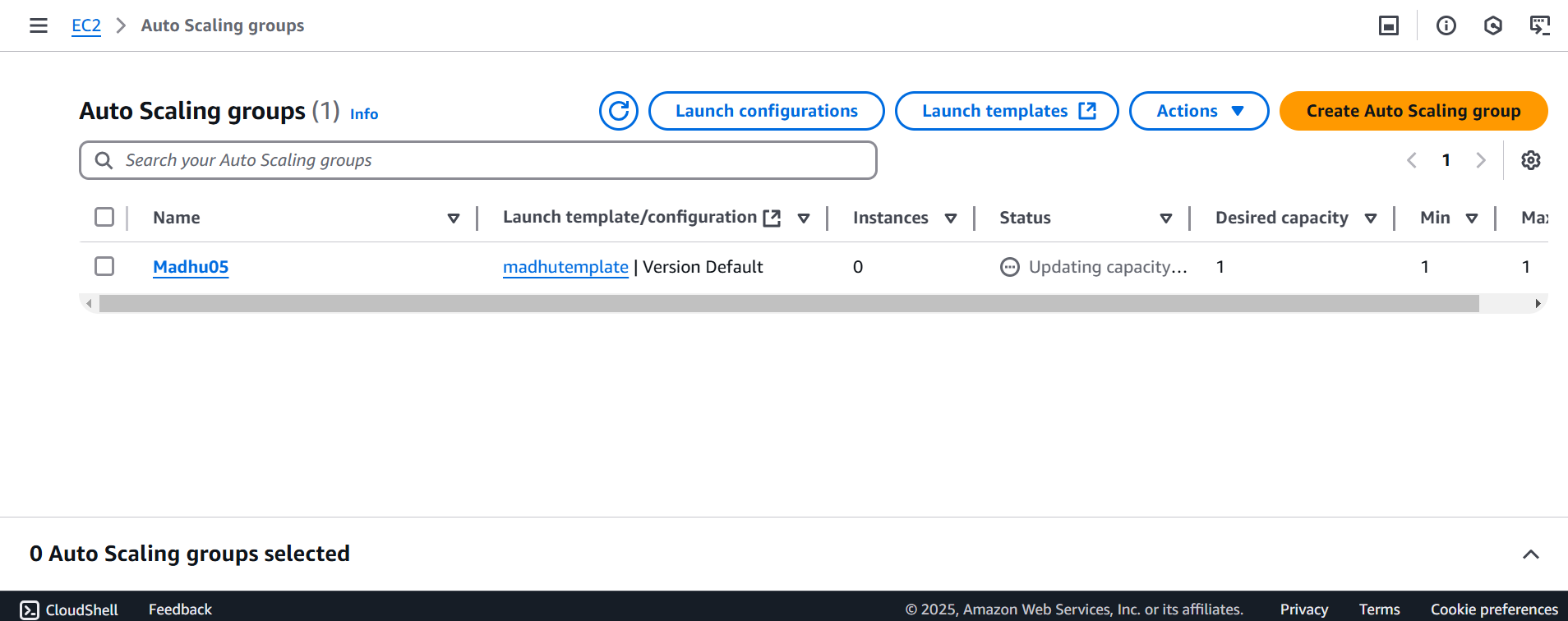
* Provide a name for the ASG and select the launch template created earlier.
* Choose the VPC and select any subnets. 
* Set desired, minimum, and maximum capacity (e.g., 1 min, 2 desired, 4 max).



* Add a Target Tracking Scaling Policy:
  1. Metric: Average CPU Utilization
  2. Target Value: 50%
  3. Warm-up Time: 300 seconds

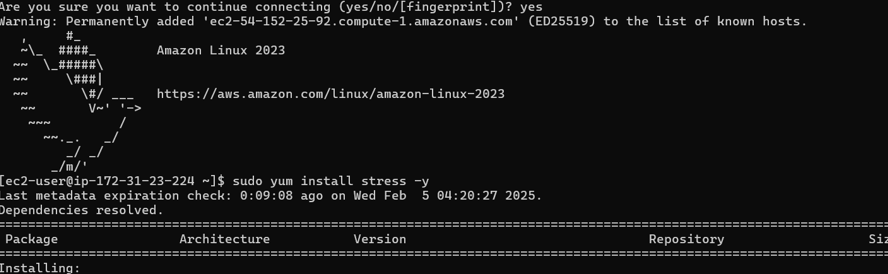


* Review and create the Auto Scaling Group**.**

****

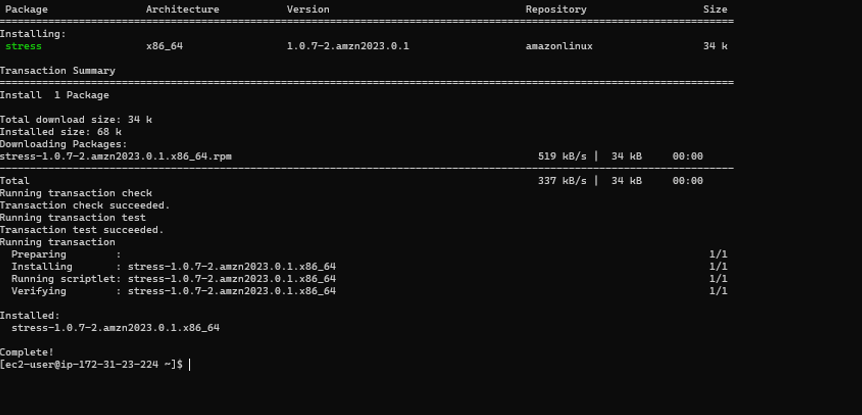
**STEP 3: Test the Auto-Scaling Setup**

1. After the ASG launches an instance, connect to it via SSH using the key pair.



1. Install the stress tool to simulate high CPU load:

**sudo yum install stress -y**

****

1. Generate CPU stress to trigger scaling:

**stress --cpu 2 --timeout 300**

1. Monitor scaling activities within the auto scaling group -> monitoring.
2. An instance will be launched due to one of the running instance has already reached it’s limit by the stress test.