

**Placement Empowerment Program**

***Cloud Computing and DevOps Centre***

**Implement Auto-scaling in the cloud**

Name: S Akshaya Department: ADS



# Introduction

Auto-scaling is a cloud computing feature that automatically adjusts the number of virtual machines (VMs) in response to workload demands. This ensures optimal performance while controlling costs by scaling resources up during high demand and down during low demand.

# Overview

Cloud providers like AWS, Azure, and Google Cloud offer auto-scaling services that help manage fluctuating workloads efficiently. By defining scaling policies, thresholds, and triggers, organizations can ensure that their applications remain highly available and responsive without manual intervention.

**Objectives**

* Understand the importance of auto-scaling in cloud environments.
* Set up an auto-scaling group for VMs on Windows.
* Configure scaling policies and triggers to manage workloads dynamically.

**Step-by-Step Overview**

**Step 1: Sign in to AWS Console**

1. Go to AWS management console
2. Navigate to the EC2 Dashboard.

Step 2: Create a Launch Template

1. In the EC2 Dashboard, select Launch Templates under "Instances".
2. Click Create Launch Template.
3. Provide a name and description.
4. Under Amazon Machine Image (AMI), choose an appropriate AMI (e.g., Amazon Linux 2 or Windows Server).
5. Select an Instance Type (e.g., t2.micro for Free Tier).
6. Configure key pair for SSH access (for Linux) or RDP (for Windows).
7. Under Network Settings, select an appropriate VPC and subnet.
8. Configure Security Group to allow necessary inbound traffic (e.g., HTTP, HTTPS, or SSH).
9. Click Create Launch Template.

Step 3: Create an Auto Scaling Group

1. Go to EC2 Dashboard > Auto Scaling Groups.
2. Click Create Auto Scaling Group.
3. Enter a name for the Auto Scaling group.
4. Choose the Launch Template created earlier.
5. Select a VPC and at least two subnets for high availability.
6. Set Desired Capacity, Minimum Capacity, and Maximum Capacity:
   * Desired: Number of instances initially running.
   * Minimum: Minimum number of instances.
   * Maximum: Maximum number of instances.

Step 4: Configure Scaling Policies

1. Select Target Tracking Scaling or Step Scaling.
2. If using Target Tracking, choose a metric like CPU Utilization:
   * Example: Scale out if CPU > 50% for 5 minutes.
   * Scale in when CPU < 20% for 5 minutes.
3. For Step Scaling, configure scaling adjustments based on predefined thresholds.

Step 5: Configure Load Balancer (Optional but Recommended)

1. Choose Attach to an existing load balancer or create a new one.
2. If using an Application Load Balancer (ALB):
   * Choose the ALB and Target Group.
   * Ensure health checks are configured.

Step 6: Enable Notifications (Optional)

1. Configure Amazon SNS to receive alerts for instance scaling events.
2. Subscribe via email or SMS.

Step 7: Review and Create

1. Review all settings.
2. Click Create Auto Scaling Group.
3. Monitor activity in Auto Scaling Group Dashboard.

# Outcome

Implementing Auto Scaling in AWS ensures that your application can automatically handle variable