Hands-on with AWS Auto Scaling

What is Scaling?

Scaling is the process of **increasing or decreasing the number of servers** based on requirements.

Types of Scaling

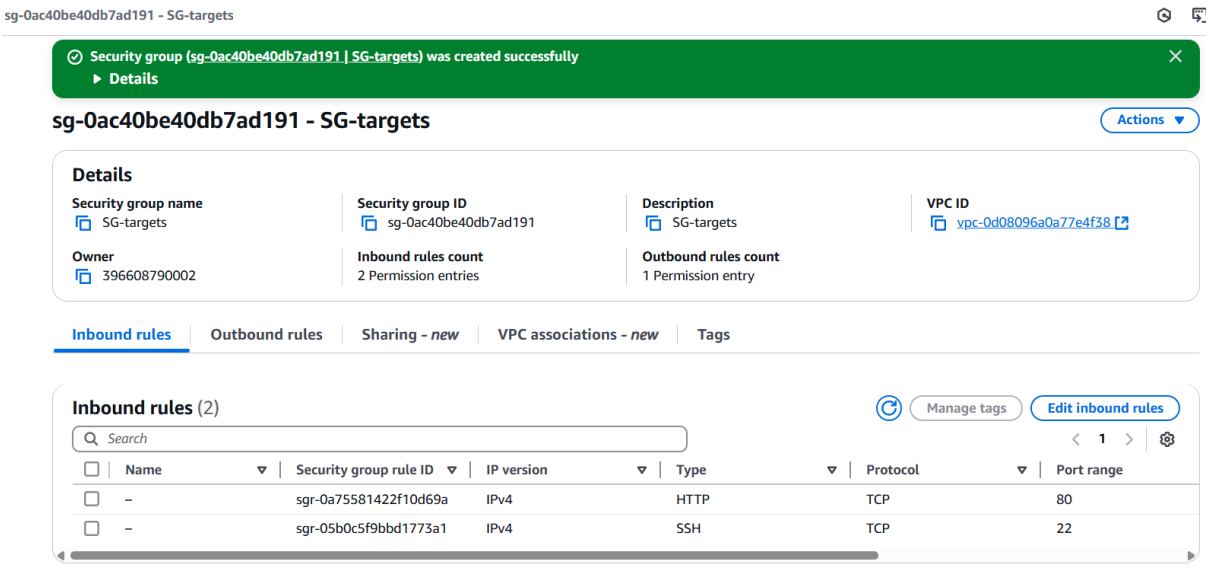
1. **Vertical Scaling** – Add or reduce resources like CPU/RAM in a **single machine**
    - *Scale Up* – Increase resources
    - *Scale Down* – Decrease resources
2. **Horizontal Scaling** – Add or remove **multiple machines**
    - ➕ *Scale Out* – Add more machines
    - ➖ *Scale In* – Remove machines

Auto Scaling means adjusting the number of servers **automatically** based on system load.

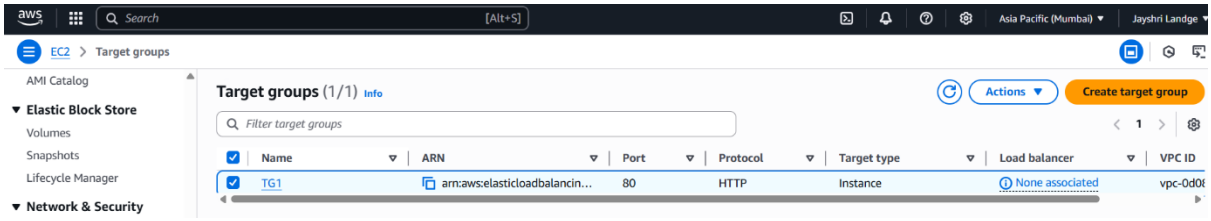
Auto Scaling Practical Steps (AWS)

Step 1: Initial Setup

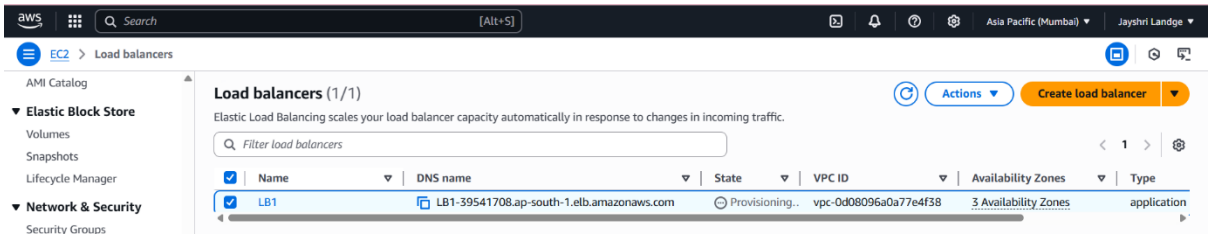
- Create a **Security Group** with port **22 (SSH)** and **80 (HTTP)**
- Create an **empty Target Group (TG)**
- Launch a **Load Balancer (LB)** and attach the TG



Screenshot 1: Security Group



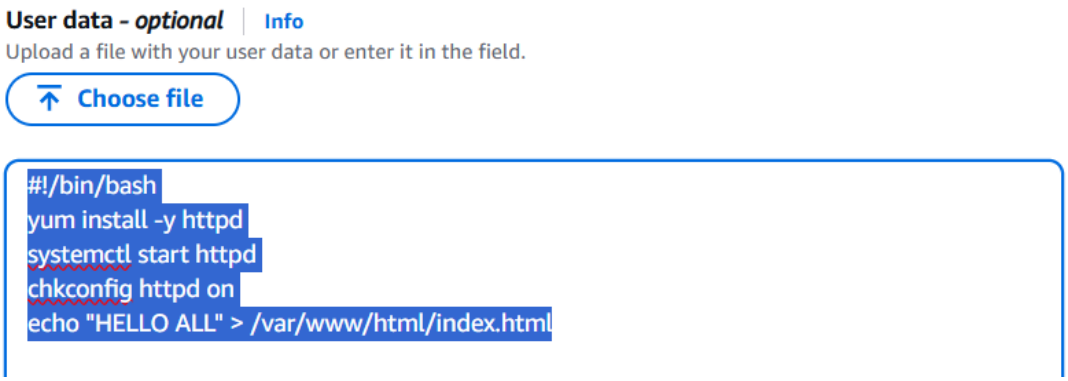
Screenshot 2: Empty Target Group

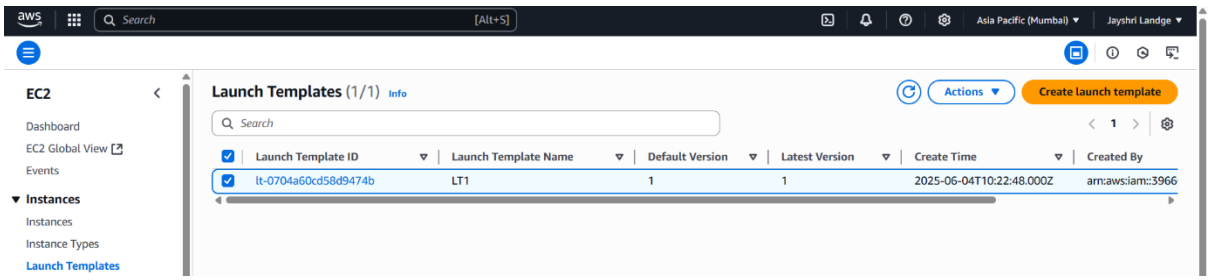


Screenshot 3: Load Balancer creation

Step 2: Launch Template

- Define OS, CPU, RAM, storage, key pair, SG, and **bootstrap script**
- This template helps to **automatically create EC2 instances** during high load

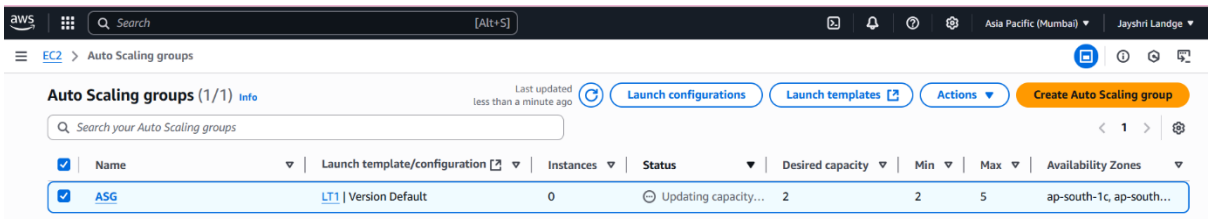




Screenshot 4: Launch Template with AMI, CPU, RAM, Key, SG, and script

Step 3: Auto Scaling Group (ASG)

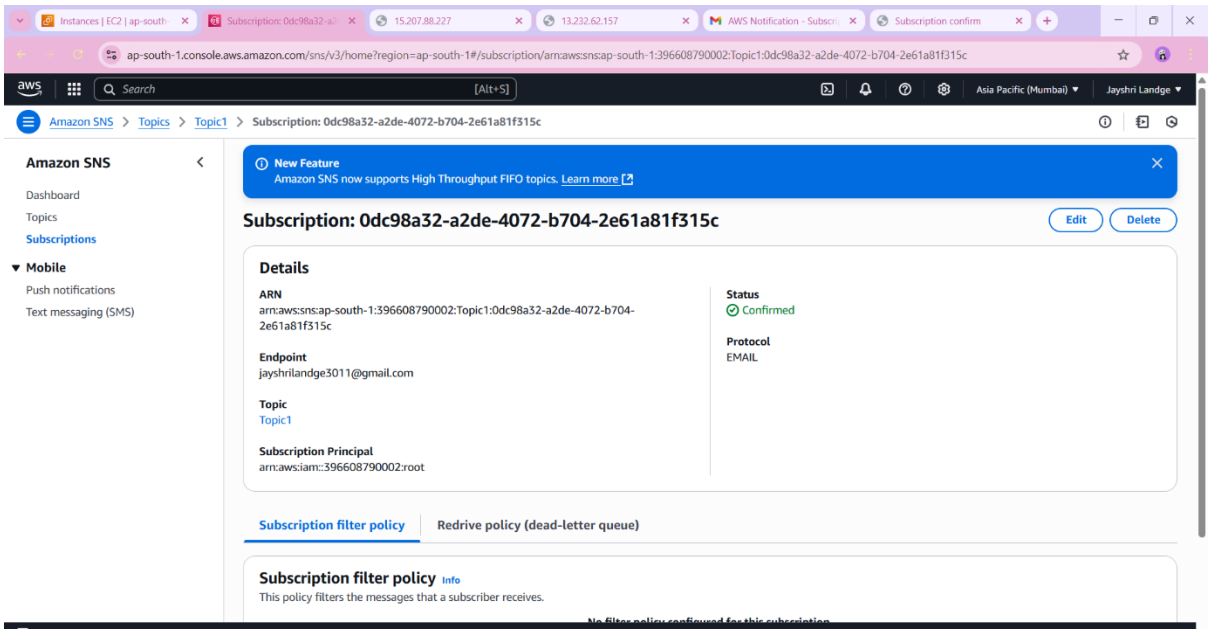
- Configure **minimum**, **maximum**, and **desired capacity**
    - ◆ *Minimum* → Instances during low load
    - ◆ *Maximum* → Instances during high load
    - ◆ *Desired* → Normal state between min & max



Screenshot 5: Auto Scaling Group setup with min, max, and desired values

Step 4: SNS Setup (Simple Notification Service)

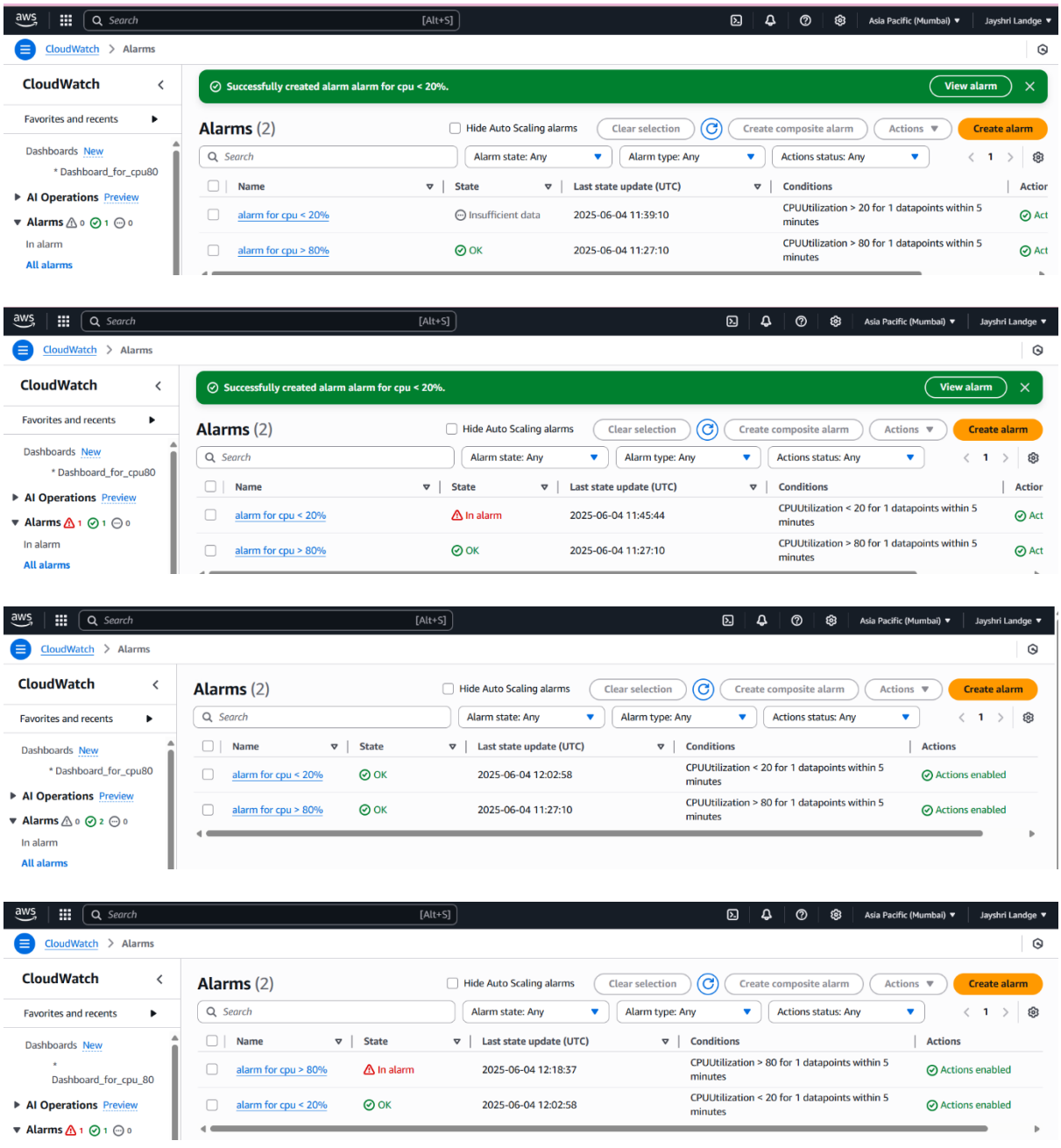
- SNS is a region-specific service used to send notifications
- Create an **SNS topic**, add a **subscription** (email of the recipient)



Screenshot 6: SNS Topic and email subscription

Step 5: CloudWatch Alarms

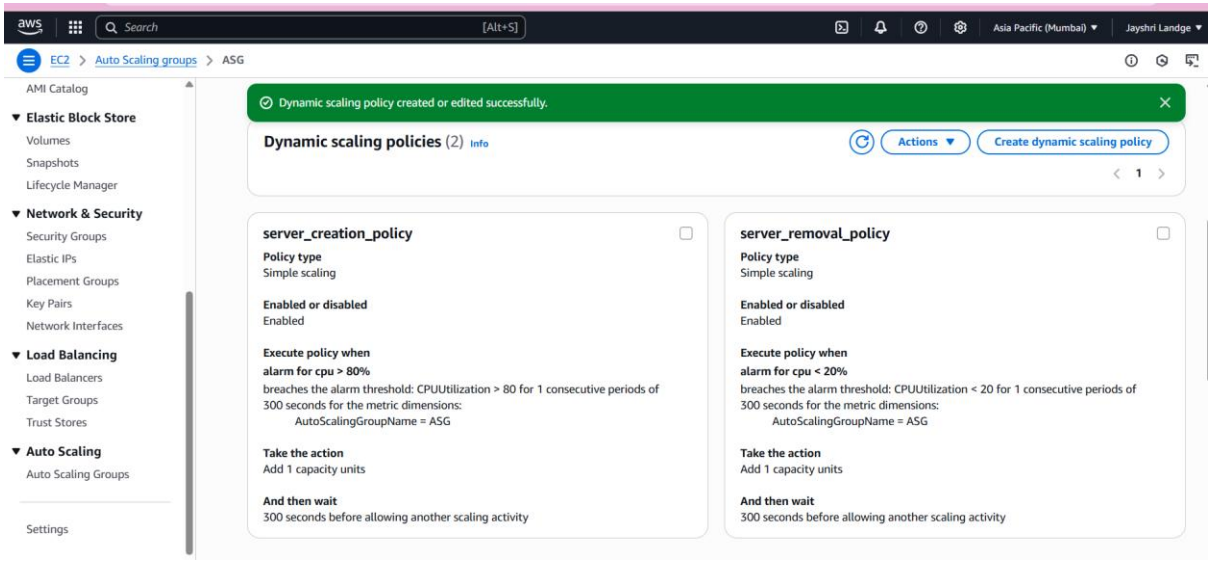
- Create 2 alarms:
  - 📈 CPU > 80% → Scale Out
  - 📉 CPU < 20% → Scale In



Screenshot 7: CloudWatch alarms - CPU > 80% and CPU < 20%

Step 6: Scaling Policies

- **Policy 1:** Create instances (attach CPU > 80% alarm)
- **Policy 2:** Remove instances (attach CPU < 20% alarm)



Screenshot 8: Scaling policies attached to alarms

Components of Auto Scaling (AWS)

1. EC2 Target
2. Target Group
3. Load Balancer (ALB)
4. Launch Template
5. Auto Scaling Group (ASG)
6. SNS Topic
7. CloudWatch Alarms
8. Scaling Policies

Conclusion

This hands-on exercise helped me understand how AWS Auto Scaling automatically handles instance provisioning and removal based on real-time metrics. It improves **availability**, **cost-efficiency**, and **performance** of applications hosted on EC2.