**EC2 AUTO SCALLING**

· Amazon EC2 autoscaling is designed as a fully managed service that controls the number of running instances, in case your workload is higher, it will match by launching more instances instantaneously.

· Unlike load balancer, you don’t provision EC2 Instances in advance and register as a target but you define the whole configuration in the Launch template which will scale out or scale in based on the traffic.

· In auto-scaling, you also define what will be minimum capacity, desired capacity, and maximum capacity. Autoscaling makes sure that the number of instances specified in desired capacity is always running when traffic is normal, it will scale in when traffic is lowest and scale-out when running there is a spike.

· While creating auto-scaling, it requires a Launch template where you specify which AMI to choose, what will be instance type, security group associated, and all other options required to launch an EC2 Instance including key pair.

**Lab Steps:**

Task 1: Sign in to AWS Management Console

1. Click on the Open Console button, and you will get redirected to AWS Console in a new browser tab.

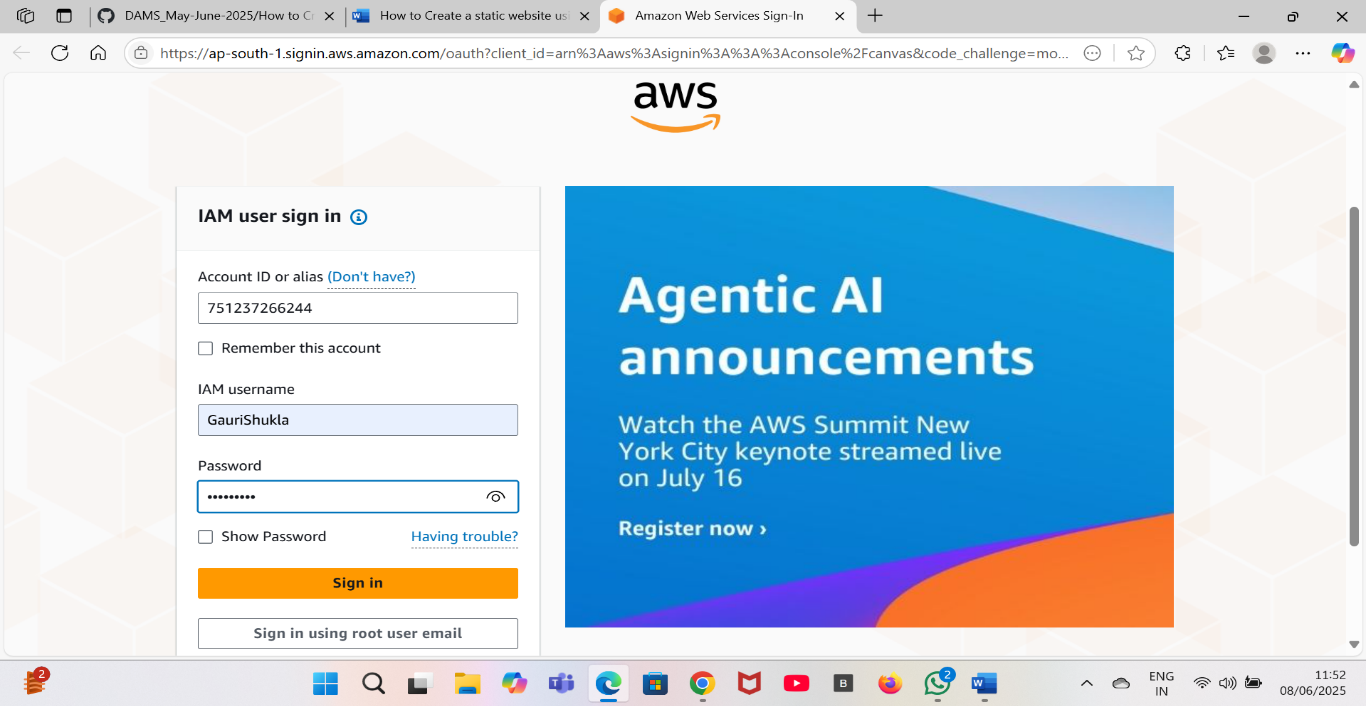
2. On the AWS sign-in page,

· Leave the Account ID as default

·Now enter your username and password.

3. click on sign-in.

4.After signing in select US East (N. Virginia) us-east- as AWS region.

1

Task 2: Creating Launch Template:

1.Choose region as US East (N. Virginia) us-east-1.

2.Click on search bar and enter EC2 then click on EC2.

3.In left side scroll down and click on launch template.

4.Click on create Launch template.

5.Enter the details:

a. Enter template name: Enter your template name (gauri2).

b. Template version description: Enter Launch template version 1.

c. Amazon machine image (AMI): Select Amazon Linux 2 AMI

d. Instance type: Select t2.micro.

e. Key pair name: Don’t include in launch template.

f. Network Settings:

Security groups: Select the Default security group of Default VPC.

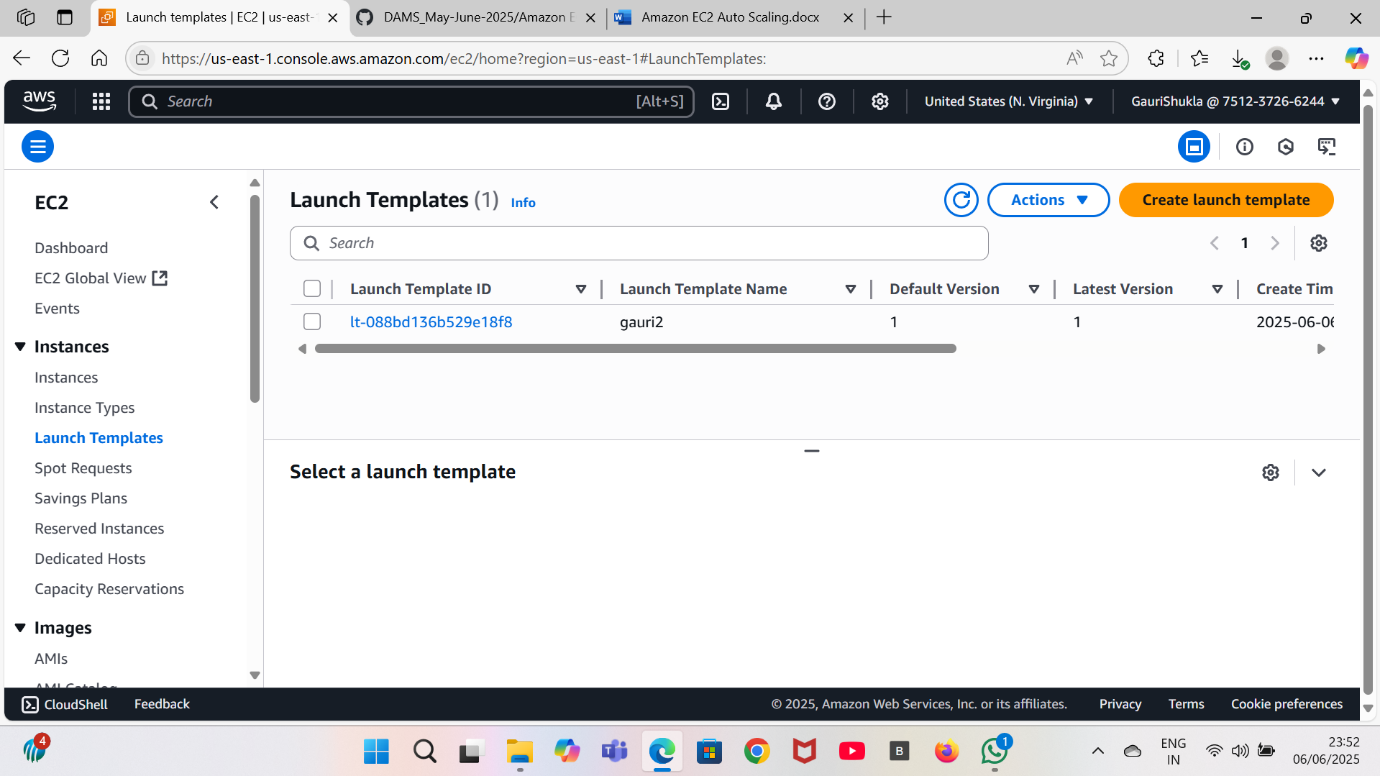
6. Leave all the settings as default.

7. Click on create launch template.

8. Your template is now created.

9. Click on view launch template.

10. You can see the template which you’ve launched .



Task 3: Create an Auto Scaling Group

1. Go to left menu under EC2 and scroll down to Auto Scaling .
2. Now click on Auto scaling group.
3. Click on Create Auto Scaling group.
4. Step 1: Choose launch template or configuration

Auto Scaling group name: Enter gauri21

Launch template: Select the launch template which you’ve created (gauri2)

Click on the Next button.

1. Step 2: Configure settings

VPC: Select the Default VPC .

Subnet: Select all the subnets.

Click on the Next button.

1. Step 3: Configure advanced options

No changes are needed on this page

Click on Next button.

1. Step 4: Configure group size and scaling policies.

Under Group size – optional.

Desired capacity: Enter 2.

Minimum capacity: Enter 2.

Maximum capacity: Enter 2.

Under Scaling policies – optional.

Select None.

Under Instance scale-in protection.

No changes are needed.

Click on Next button.

1. Step 5: Add notifications

No changes are needed on this page.

Click on Next button.

1. Step 6: Add tags.

Enter tags in key-value pairs to identify your auto-scaling group..

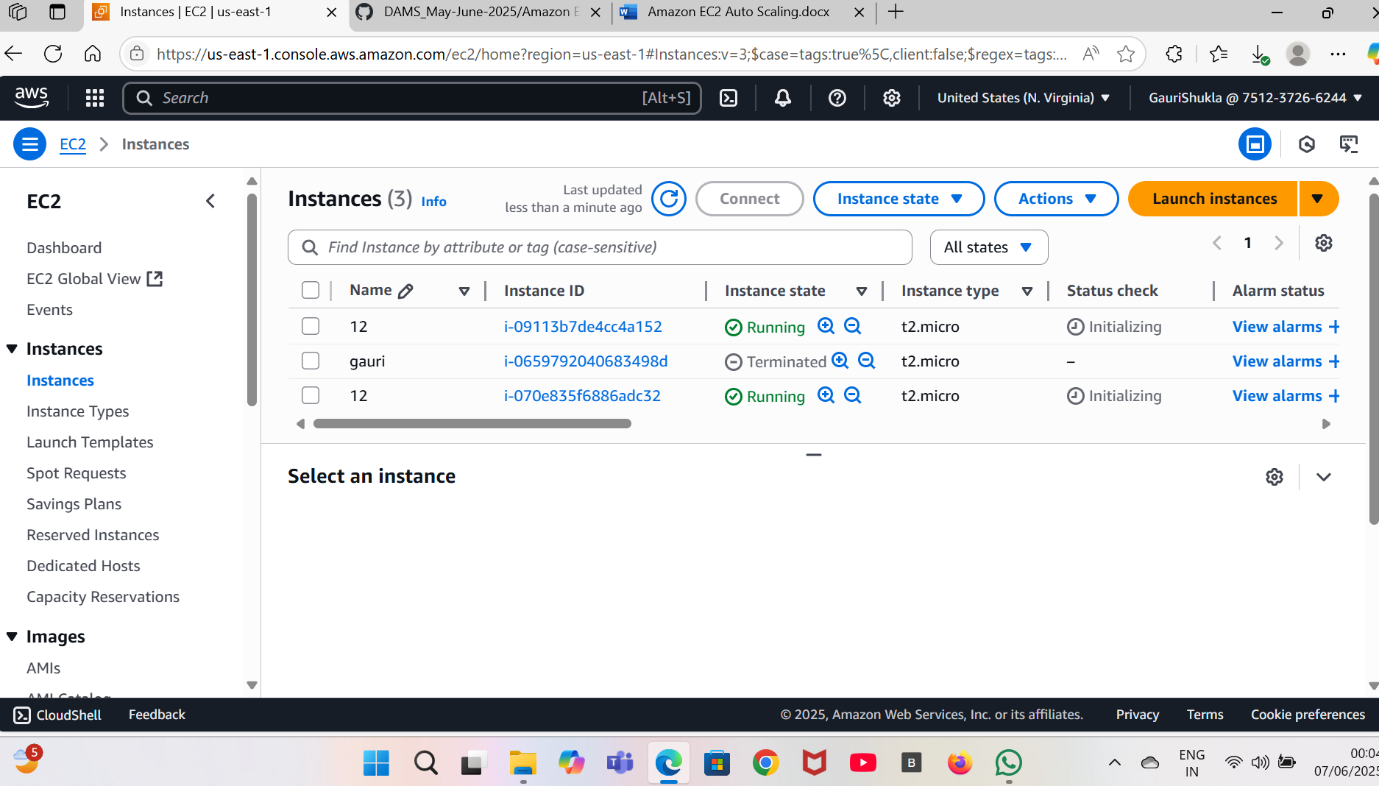
Click on the Add tag button.

Key: Name.

Value: ASG-EC.

Click on Next button.

1. Now scroll down and click on the Create Auto Scaling group button.
2. Auto scaling is created successfully.
3. You will be redirected to the autoscaling group page, you will be able to see that two instances are launched by the autoscaling group.
4. Now go to the EC2 instances list. You will see that there are two new running instances (which were created by your autoscaling group).
5. You have successfully created an autoscaling group with a policy of a minimum of 2 and a maximum of 2 instances.



**INTERVIEW QUESTIONS RELATED TO EC2 AUTO SCALING**

1. Explain Auto Scaling in EC2 and how it helps ensure application availability.

Answer: Auto Scaling is a service that allows you to automatically adjust the number of EC2 instances in response to changing application demands. It helps maintain application availability, improves fault tolerance, and optimizes resource usage by scaling out during periods of high demand and scaling in during lower demand.

2. . How does EC2 handle instance failures, and what strategies can you implement to ensure high availability for your applications?

Answer: EC2 instances can fail, but you can enhance availability by:  
— Using Auto Scaling to automatically replace failed instances.  
— Distributing your application across multiple Availability Zones.  
— Implementing load balancing with Elastic Load Balancing (ELB).  
— Utilizing Elastic IP addresses and DNS solutions for failover.

3.  How can you automate EC2 instance provisioning and scaling using AWS services like AWS Auto Scaling ?

Answer: AWS Auto Scaling allows you to automatically adjust the number of EC2 instances based on specified conditions, such as CPU utilization or application load.