

AUTO-SCALING PRAC

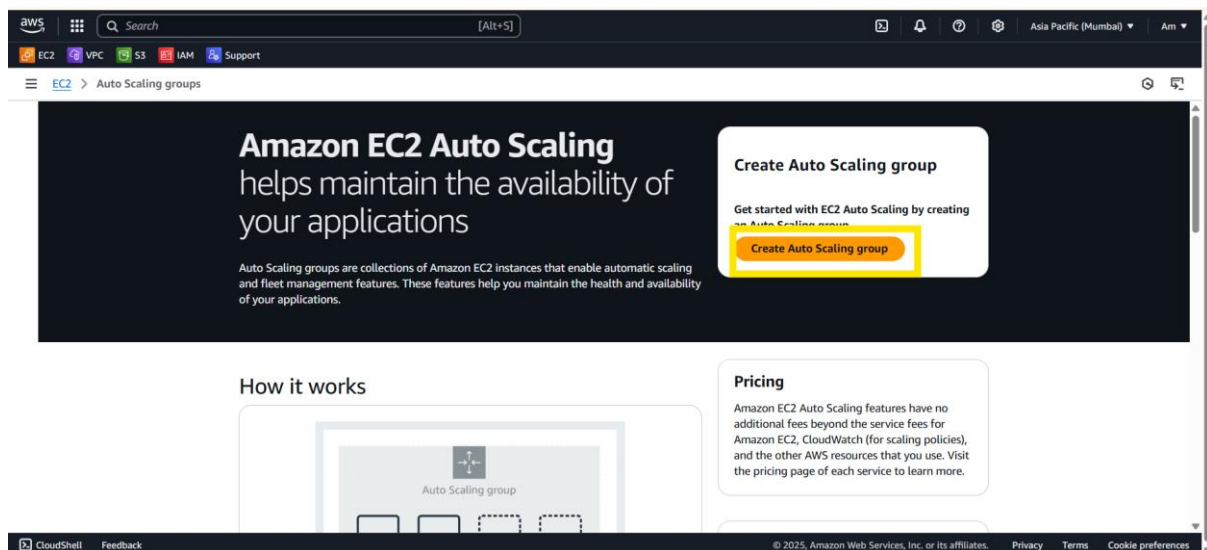
Define Auto-scaling:

system that automatically adjusts your cloud resources (like servers or databases) to match your application's needs.

Why use Auto-scaling:

automatically adjusts your resources to match the demands of your application, ensuring optimal performance and cost efficiency.

Click on create autoscaling group:-



STEP 1: name it and create launch template to launch instance through launch template.

aws Search [Alt+S] Asia Pacific (Mumbai) Am

EC2 VPC S3 IAM Support

EC2 > Auto Scaling groups > Create Auto Scaling group

Step 2 Choose instance launch options
Step 3 - optional Integrate with other services
Step 4 - optional Configure group size and scaling
Step 5 - optional Add notifications
Step 6 - optional Add tags
Step 7 Review

Name

Auto Scaling group name
Enter a name to identify the group.

MyASC

Must be unique to this account in the current Region and no more than 255 characters.

Launch template [Info](#)

For accounts created after May 31, 2023, the EC2 console only supports creating Auto Scaling groups with launch templates. Creating Auto Scaling groups with launch configurations is not recommended but still available via the CLI and API until December 31, 2023.

Launch template
Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.

Select a launch template

Create a launch template

Cancel Next

CloudShell Feedback

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

STEP 2: all other things will be same as we create a launch template I have only added a new SG name all traffic, other details can be seen in SC.

aws [Alt+S]

EC2 VPC S3 IAM Support

EC2 > Launch templates > Create launch template

Security group name - *required*

1

This security group will be added to all network interfaces. The name can't be edited after the security group is created. Max length is 255 characters. Valid characters: a-z, A-Z, 0-9, spaces, and _-./()#,@[]+=&:~!\$*

Description - *required* [Info](#)

VPC [Info](#)

(default)

172.31.0.0/16

Inbound Security Group Rules

▼ Security group rule 1 (All, All, 0.0.0.0/0)

Type [Info](#) Protocol [Info](#) Port range [Info](#)

2 All All

Source type [Info](#) Source [Info](#) Description - *optional* [Info](#)

3

▼ Security group rule 2 (TCP, 22, 0.0.0.0/0)

4 Type [Info](#) Protocol [Info](#) Port range [Info](#)

TCP 22

Source type [Info](#) Source [Info](#) Description - *optional* [Info](#)

5

STEP 3: Adding script to install nginx and apache.

The screenshot shows the 'Launch an instance' page in the AWS Management Console. The 'User data' section is highlighted with a black border. It contains a text area with the following script:

```
#!/bin/bash

sudo apt update -y
sudo install nginx -y
systemctl start nginx
systemctl enable nginx

sudo install apache2 -y
systemctl start apache2
systemctl enable apache2
```

Below the text area, there is a checkbox labeled 'User data has already been base64 encoded' which is currently unchecked.

USING THIS SCRIPT SO THAT, WHEN THE INSTANCE IS LAUNCHED WE CAN DIRECTLY ACCESS BOTH THESE PROXY SERVERS BY USING PUBLIC IP OF THE INSTANCE.

STEP 4: Choosing the instance requirements of how many instances to be launched at the initial start of the instance.

The screenshot shows the 'Create Auto Scaling group' page in the AWS Management Console. The 'Choose instance launch options' step is highlighted. The 'Instance type requirements' section is expanded, showing the 'Specify instance attributes' option selected. The 'Required instance attributes' section is visible, with the following settings:

- vCPUs:** Minimum 0, Maximum 1 (No maximum checkbox is unchecked).
- Memory (GiB):** Minimum 0, Maximum 1 (No maximum checkbox is unchecked).

The 'Additional instance attributes - optional' section is also visible at the bottom.

STEP 5: NETWORK CONFIG—Selecting all the AZ's

quickly.

VPC
Choose the VPC that defines the virtual network for your Auto Scaling group.

vpc-0077c0b100707fd7e
172.31.0.0/16 Default

[Create a VPC](#)

Availability Zones and subnets
Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.

Select Availability Zones and subnets

ap-south-1a | subnet-0b4a2733ea77b0881
172.31.32.0/20 Default

ap-south-1b | subnet-057511227b8b41e9b
172.31.0.0/20 Default

ap-south-1c | subnet-022b390090e2a32de
172.31.16.0/20 Default

[Create a subnet](#)

Availability Zone distribution - new
Auto Scaling automatically balances instances across Availability Zones. If launch failures occur in a zone, select a strategy.

☒ **Balanced best effort**
If launches fail in one Availability Zone, Auto Scaling will attempt to launch in another healthy Availability Zone.

☐ **Balanced only**
If launches fail in one Availability Zone, Auto Scaling will continue to attempt to launch in the unhealthy Availability Zone to preserve balanced distribution.

[Cancel](#) [Skip to review](#) [Previous](#) [Next](#)

STEP 6: In health checks we give 120 seconds.

Explain why 120 seconds?

- 120 seconds mean the servers we have taken nginx and apache will refresh after 120 seconds.
- If the load increases on one primary instance, autoscaling will re-direct the traffic to other instance.

Application Recovery Controller (ARC) zonal shift - new [Info](#)
During an Availability Zone impairment, target instance launches towards other healthy Availability Zones.

☐ **Enable zonal shift**
New instance launches will be retargeted towards healthy Availability Zones until the zonal shift is cancelled.

Health checks
Health checks increase availability by replacing unhealthy instances. When you use multiple health checks, all are evaluated, and if at least one fails, instance replacement occurs.

EC2 health checks
[Always enabled](#)

Additional health check types - optional [Info](#)

☐ **Turn on Elastic Load Balancing health checks**
Elastic Load Balancing monitors whether instances are available to handle requests. When it reports an unhealthy instance, EC2 Auto Scaling can replace it on its next periodic check.

☐ **Turn on VPC Lattice health checks**
VPC Lattice can monitor whether instances are available to handle requests. If it considers a target as failed a health check, EC2 Auto Scaling replaces it after its next periodic check.

☐ **Turn on Amazon EBS health checks**
EBS monitors whether an instance's root volume or attached volume stalls. When it reports an unhealthy volume, EC2 Auto Scaling can replace the instance on its next periodic health check.

Health check grace period [Info](#)
This time period delays the first health check until your instances finish initializing. It doesn't prevent an instance from terminating when placed into a non-running state.

120 seconds

[Cancel](#) [Skip to review](#) [Previous](#) [Next](#)

CloudShell Feedback

© 2025, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

STEP 7: Set desired capacity and max. desired capacity.

The screenshot shows the AWS Management Console interface for the 'Create Auto Scaling group' wizard. The breadcrumb navigation at the top indicates the path: EC2 > Auto Scaling groups > Create Auto Scaling group. The left-hand navigation pane lists seven steps: Step 1: Choose launch template, Step 2: Choose instance launch options, Step 3 - optional: Integrate with other services, Step 4 - optional: Configure group size and scaling (highlighted with a blue circle), Step 5 - optional: Add notifications, Step 6 - optional: Add tags, and Step 7: Review.

The main content area is titled 'Configure group size and scaling - optional' with an 'info' icon. Below the title is a descriptive sentence: 'Define your group's desired capacity and scaling limits. You can optionally add automatic scaling to adjust the size of your group.'

The 'Group size' section includes an 'info' icon and a description: 'Set the initial size of the Auto Scaling group. After creating the group, you can change its size to meet demand, either manually or by using automatic scaling.' It contains a 'Desired capacity type' section with a dropdown menu set to 'Units (number of instances)'. Below this is a 'Desired capacity' section with a text input field containing the value '1'.

The 'Scaling' section includes an 'info' icon and a description: 'You can resize your Auto Scaling group manually or automatically to meet changes in demand.' It contains a 'Scaling limits' section with a description: 'Set limits on how much your desired capacity can be increased or decreased.' This section has two input fields: 'Min desired capacity' with a value of '1' and 'Max desired capacity' with a value of '3'. Below these fields are two small labels: 'Equal or less than desired capacity' under the min field and 'Equal or greater than desired capacity' under the max field.

At the bottom of the main content area, there is a section for 'Automatic scaling - optional' which is currently collapsed. The footer of the console shows 'CloudShell', 'Feedback', and copyright information for Amazon Web Services, Inc.