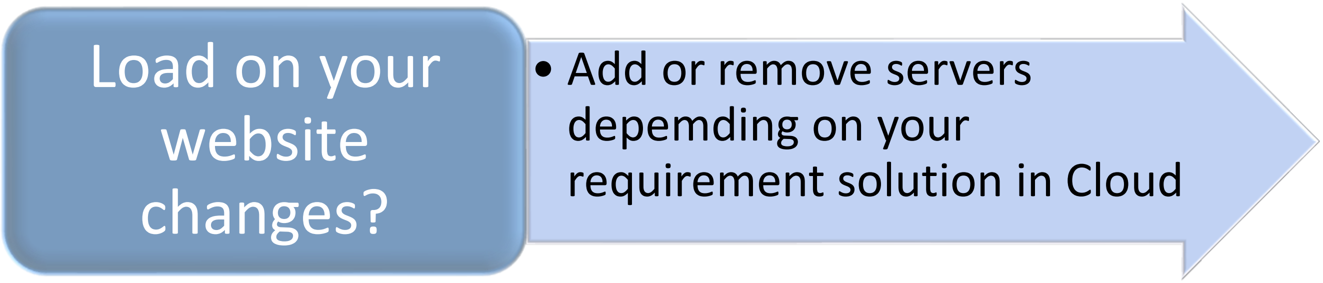
# CLOUD CONCEPTS

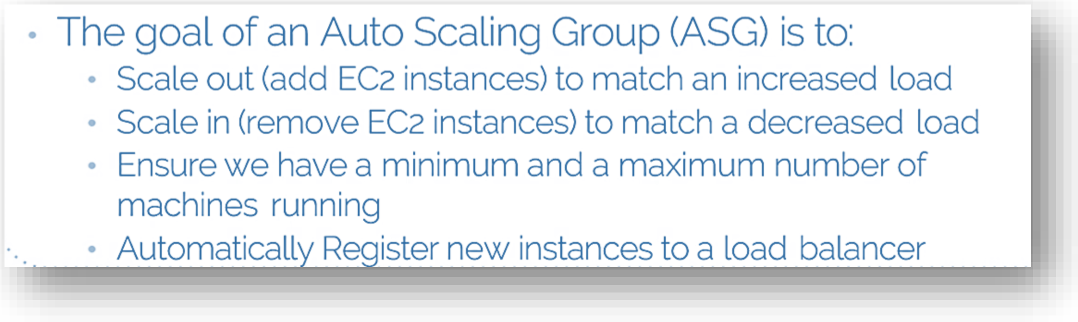
## (Auto Scaling Group- ELB)

The load on your website varies in real time, how your website will react to such a problem? Rapid action is required in this context cloud provide the scaling and elasticity in which servers can be added or removed from the auto scaling group. Min and Max machines must be mentioned earlier. To ensure high availability ASG is synced with load balancer deployed in max no. of AZs, and servers are registered and synced automatically with the Load Balancer.

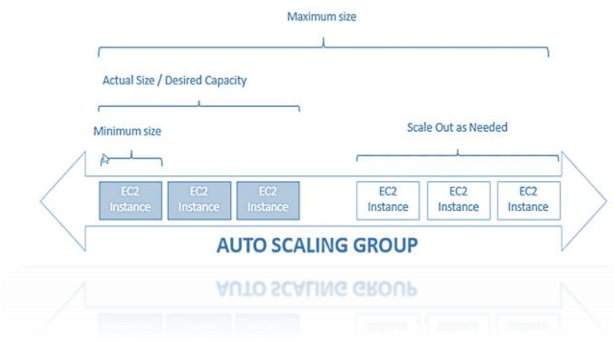


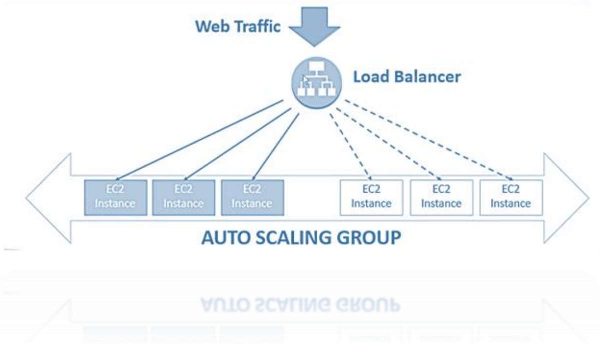
Load on your website changes?

- Add or remove servers depemding on your requirement solution in Cloud



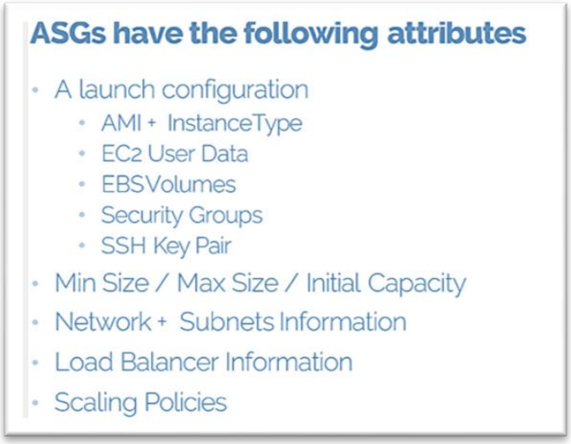
- The goal of an Auto Scaling Group (ASG) is to:
  - Scale out (add EC2 instances) to match an increased load
  - Scale in (remove EC2 instances) to match a decreased load
  - Ensure we have a minimum and a maximum number of machines running
  - Automatically Register new instances to a load balancer





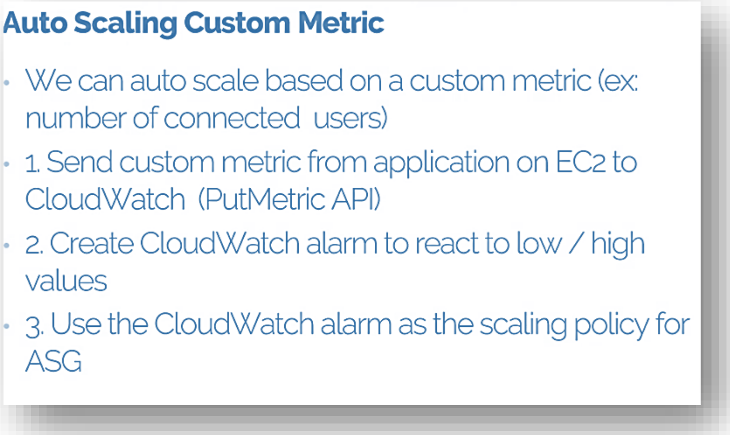
E.g., the main server must be 1, your desired capacity is 3 and maximum size is 6. Servers under unhealthy conditions will automatically launch the server from ASG. The traffic must be routed through load balancer.

**Cloud watch** is used to monitor services like LAOD Monitoring, checking health of the server and generating triggers or alarms. It is important to monitor peridically as some times there is some load which may generate the spike but its is not directly related to the traffic load. Basic monitoring by Amzaon is free. Triggers are generated after every 5 minutes, if trigger is to be generated after 1 minute it will be charged.



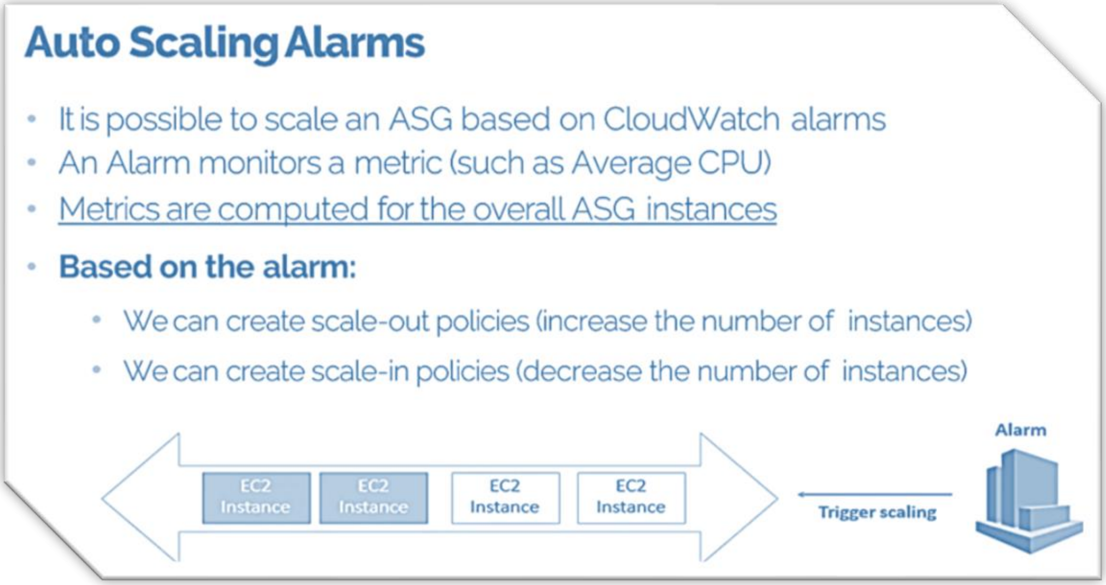
## ASGs have the following attributes

- A launch configuration
  - AMI + InstanceType
  - EC2 User Data
  - EBSVolumes
  - Security Groups
  - SSH Key Pair
- Min Size / Max Size / Initial Capacity
- Network + Subnets Information
- Load Balancer Information
- Scaling Policies



## Auto Scaling Custom Metric

- We can auto scale based on a custom metric (ex: number of connected users)
- 1. Send custom metric from application on EC2 to CloudWatch (PutMetric API)
- 2. Create CloudWatch alarm to react to low / high values
- 3. Use the CloudWatch alarm as the scaling policy for ASG



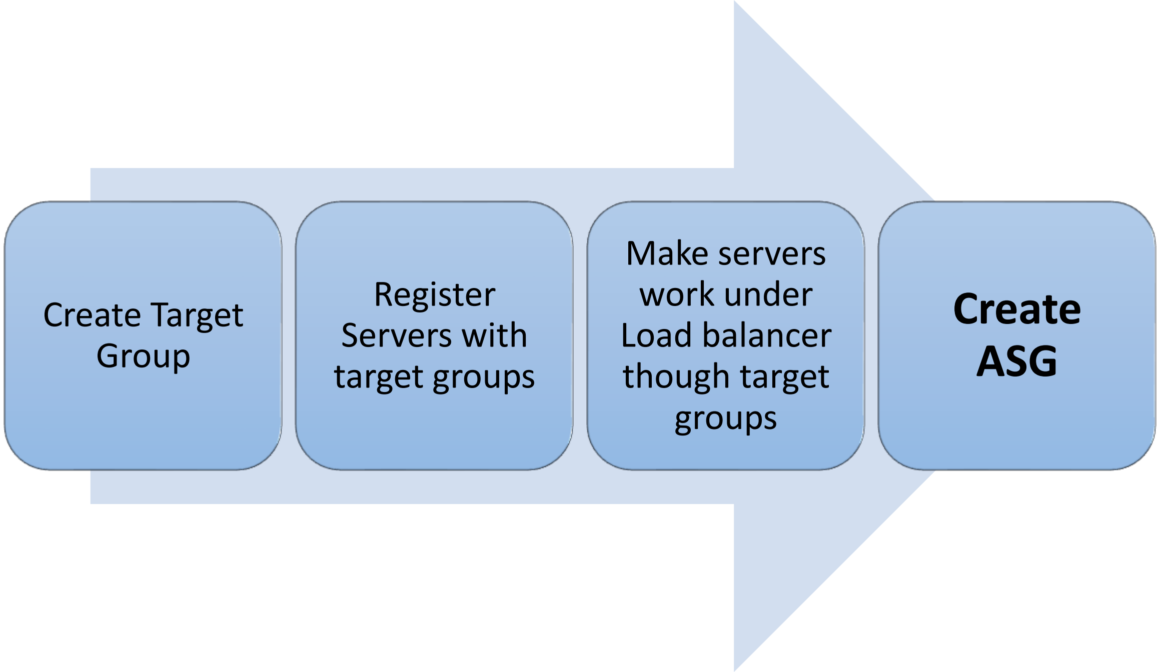
## Auto Scaling Alarms

- It is possible to scale an ASG based on CloudWatch alarms
- An Alarm monitors a metric (such as Average CPU)
- Metrics are computed for the overall ASG instances
- **Based on the alarm:**
  - We can create scale-out policies (increase the number of instances)
  - We can create scale-in policies (decrease the number of instances)

EC2 Instance
EC2 Instance
EC2 Instance
EC2 Instance
Trigger scaling
Alarm

- Scalin can be on CPU, Network or any custom metric based on schdeule.
- ASG use launch configurations which you update by providing new launch configurations.
- IAM roles attached to ASG will get assigned to EC2 instancees.
- ASG is free. You are charged for underlying resources.
- Instances under ASG will restart if they are terminated.
- ASG terminate the instance marked as unhealthy by a load balancer.

## LAB (Automatic Scaling Group using Load Balancer)



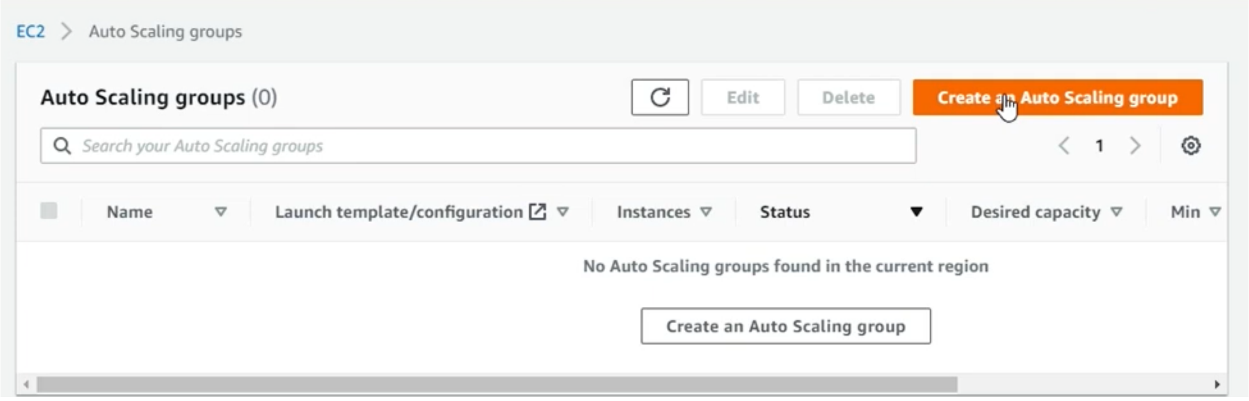
Create Target Group

Register Servers with target groups

Make servers work under Load balancer though target groups

**Create ASG**

- Scroll on left side for Auto Scaling Group and create an ASG.



EC2 > Auto Scaling groups

Auto Scaling groups (0)

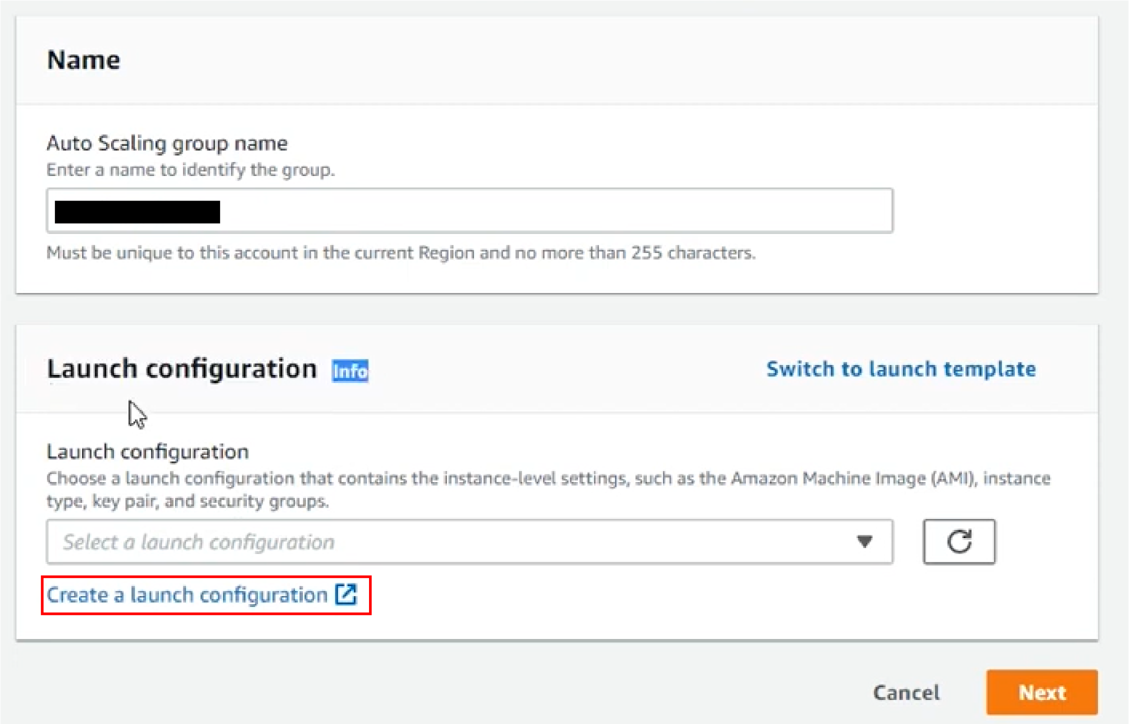
Edit
Delete
Create an Auto Scaling group
Search your Auto Scaling groups
1

| | Name | Launch template/configuration | Instances | Status | Desired capacity | Min |
|---|---|---|---|---|---|---|

No Auto Scaling groups found in the current region

Create an Auto Scaling group

- There are two options (Launch Template and launch configurations), switch to launch configurations. These configurations are for the server which is automatically launched when server scales out e.g., the Linux or windows AMI, type of resources, the boot strap script etc.
- Name the AGS and click on Create a launch configurations.



Name

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

[redacted]

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

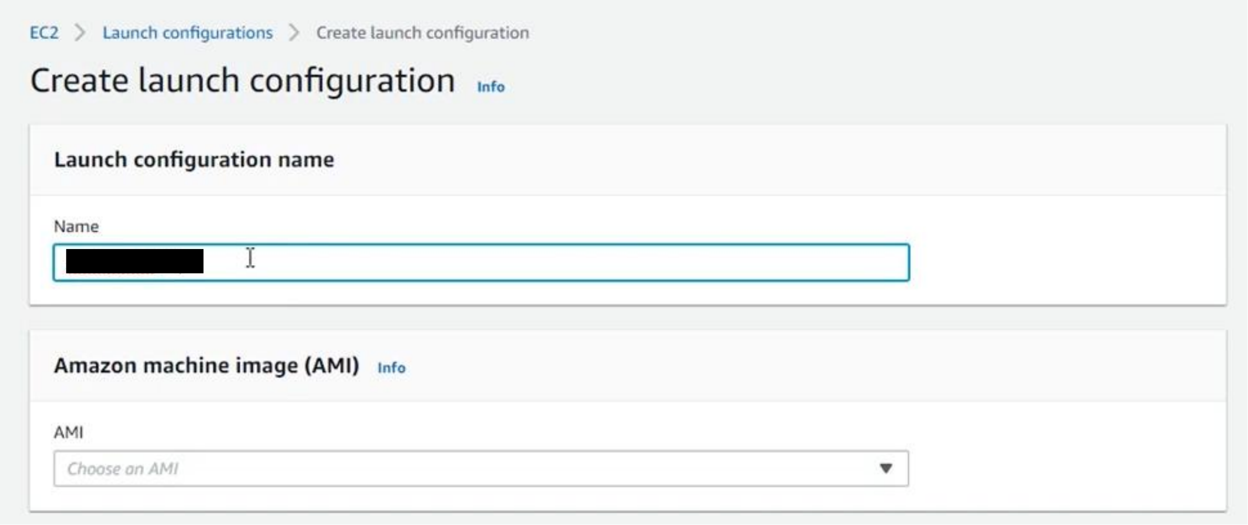
Launch configuration Info

Switch to launch template

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

Select a launch configuration
Create a launch configuration
Cancel
Next

- Now name your configuration and select an AMI using AMI id. You can find the AMI id in amazon AMI marketplace.
- Choose an instance type. (Use free tier t2 micro) and in optional field paste the bootstrap code in user data. For difference change the statement in index.html file.
- Create or use existing Security group and key pair and click Create Launch Configurations.



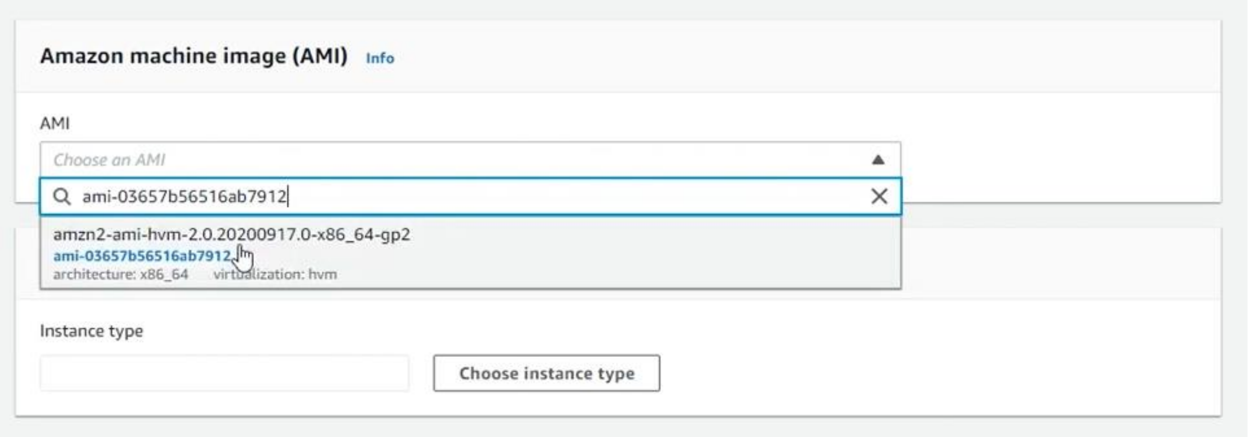
## Create launch configuration

**Launch configuration name**

Name: [redacted]

**Amazon machine image (AMI)**

AMI: *Choose an AMI*



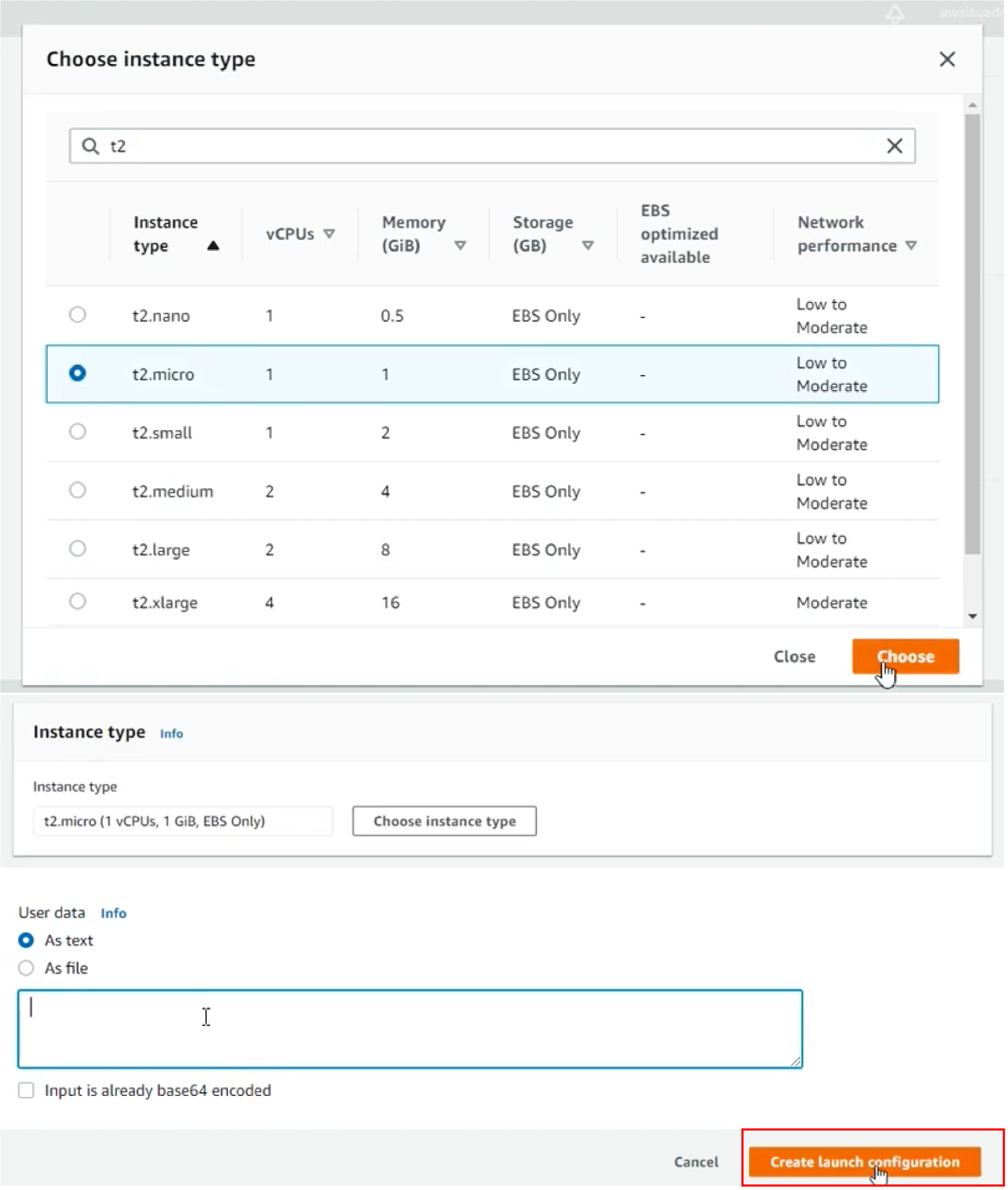
**Amazon machine image (AMI)**

AMI: *Choose an AMI*

ami-03657b56516ab7912

amzn2-ami-hvm-2.0.20200917.0-x86_64-gp2
ami-03657b56516ab7912
architecture: x86_64 virtualization: hvm

Instance type



Choose instance type

t2

| | Instance type ▲ | vCPUs ▽ | Memory (GiB) ▽ | Storage (GB) ▽ | EBS optimized available | Network performance ▽ |
|---|---|---|---|---|---|---|
| ○ | t2.nano | 1 | 0.5 | EBS Only | - | Low to Moderate |
| ● | t2.micro | 1 | 1 | EBS Only | - | Low to Moderate |
| ○ | t2.small | 1 | 2 | EBS Only | - | Low to Moderate |
| ○ | t2.medium | 2 | 4 | EBS Only | - | Low to Moderate |
| ○ | t2.large | 2 | 8 | EBS Only | - | Low to Moderate |
| ○ | t2.xlarge | 4 | 16 | EBS Only | - | Moderate |

Instance type Info

Instance type: t2.micro (1 vCPUs, 1 GiB, EBS Only)

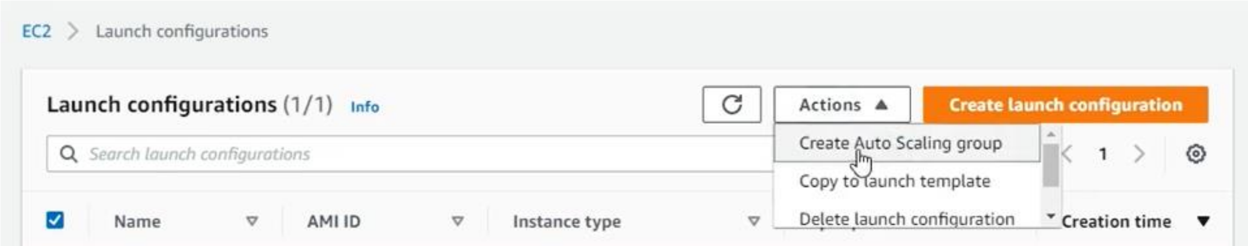
User data Info

● As text

○ As file

□ Input is already base64 encoded

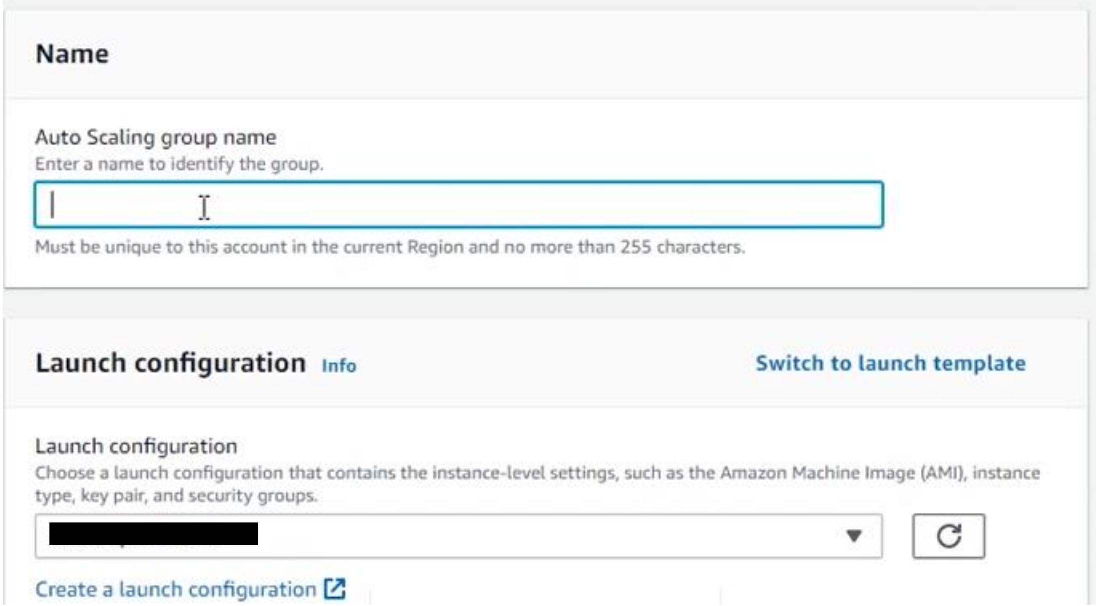
- After configuring “Launch Configurations” create an Auto Scaling Group. Give the name and select the created launch configurations.



EC2 > Launch configurations

Launch configurations (1/1) Info

Actions ▲
Create launch configuration
Search launch configurations
Create Auto Scaling group
Copy to launch template
Delete launch configuration
1
☑
Name
AMI ID
Instance type
Creation time



## Name

Auto Scaling group name

Enter a name to identify the group.

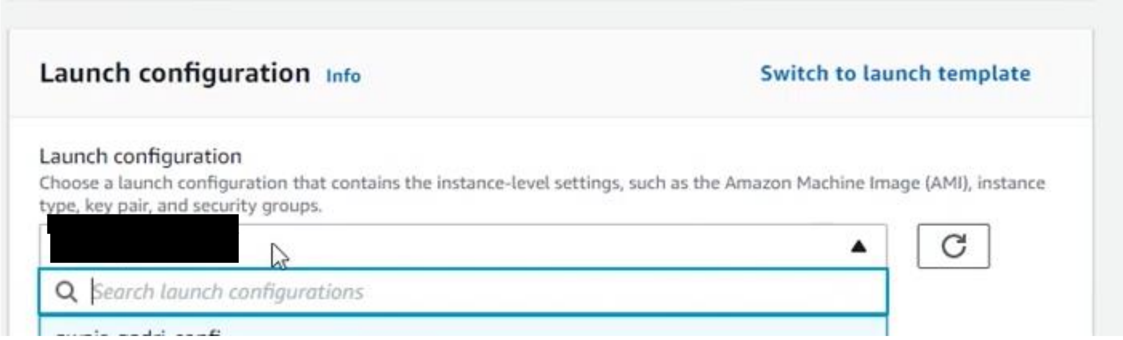
Must be unique to this account in the current Region and no more than 255 characters.

## Launch configuration Info

Launch configuration

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

[redacted]



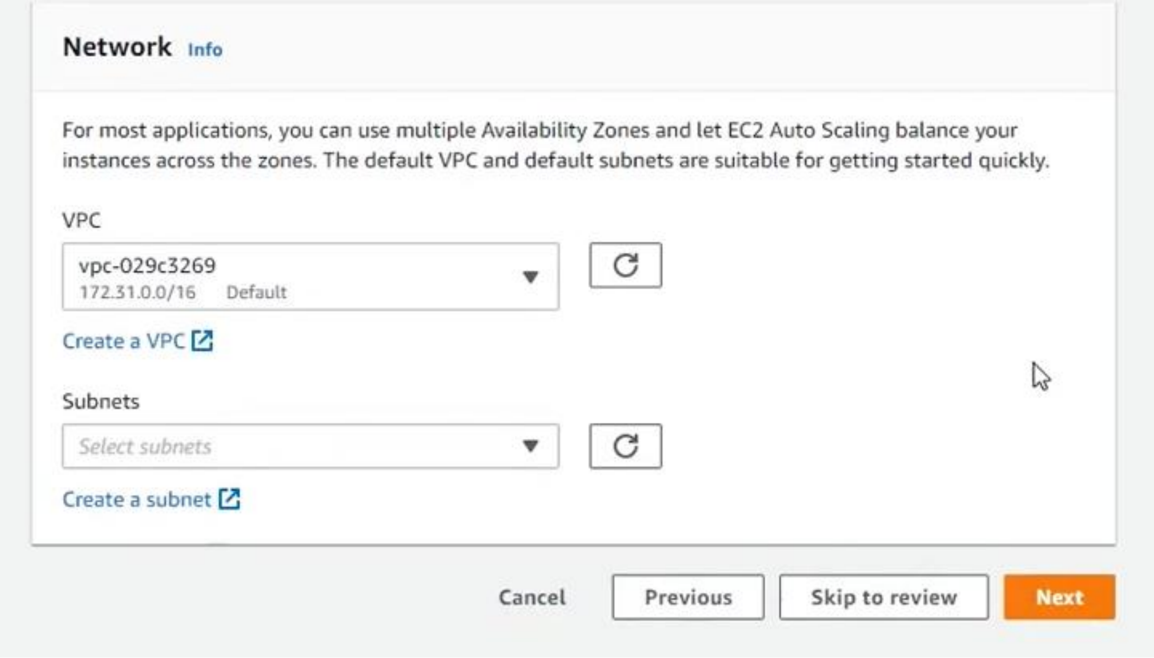
## Launch configuration Info

Launch configuration

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

[redacted]

- For High Availability select max (all) AZ = subnets in your region with default VPC.



Network Info

For most applications, you can use multiple Availability Zones and let EC2 Auto Scaling balance your instances across the zones. The default VPC and default subnets are suitable for getting started quickly.

VPC

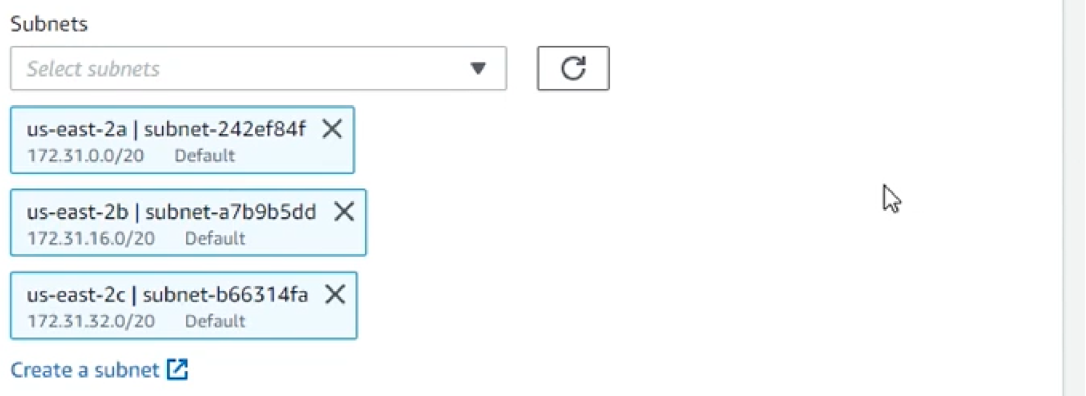
vpc-029c3269
172.31.0.0/16 Default

Create a VPC

Subnets

Select subnets

Create a subnet



Subnets

Select subnets

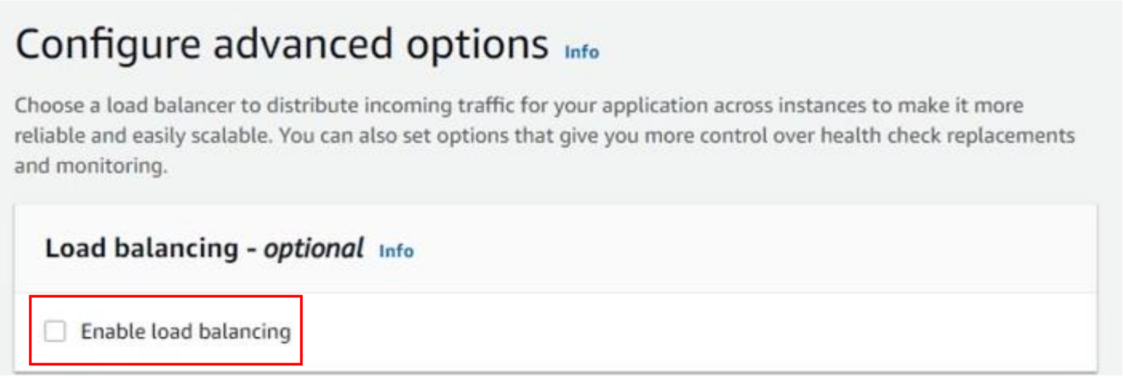
us-east-2a | subnet-242ef84f
172.31.0.0/20 Default

us-east-2b | subnet-a7b9b5dd
172.31.16.0/20 Default

us-east-2c | subnet-b66314fa
172.31.32.0/20 Default

Create a subnet

- Now enable the load balancer and select ALB. Select a target group from the created target groups and set the health checking period. When any server is unhealthy, a new server is automatically registered with the Load Balancer. Set the capacity values according to the requirements. Select scaling policies as None. Skip add notification as CLOUD watch is not checked and tagging. Click “Create Auto Scaling Group”.

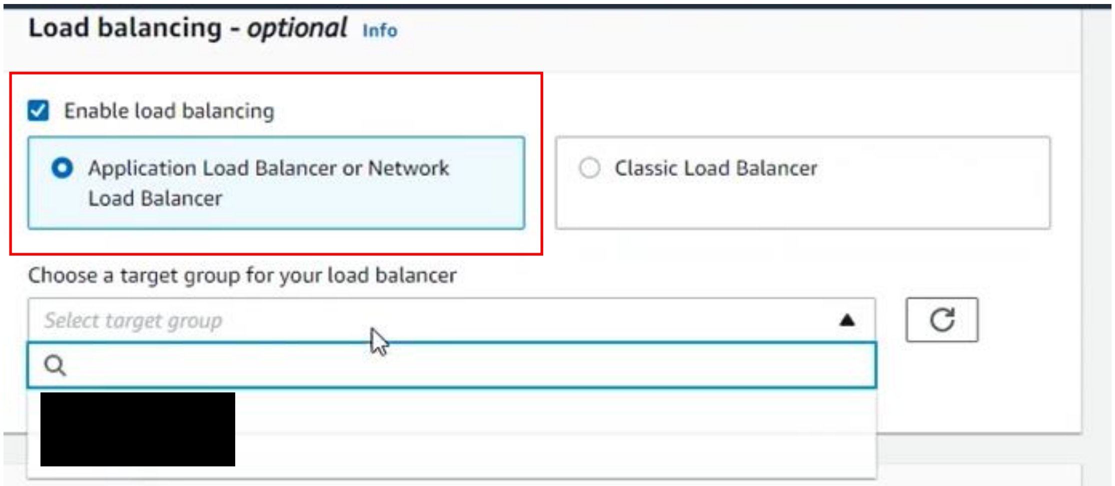


## Configure advanced options Info

Choose a load balancer to distribute incoming traffic for your application across instances to make it more reliable and easily scalable. You can also set options that give you more control over health check replacements and monitoring.

### Load balancing - *optional* Info

☐ Enable load balancing



### Load balancing - *optional* Info

☑ Enable load balancing

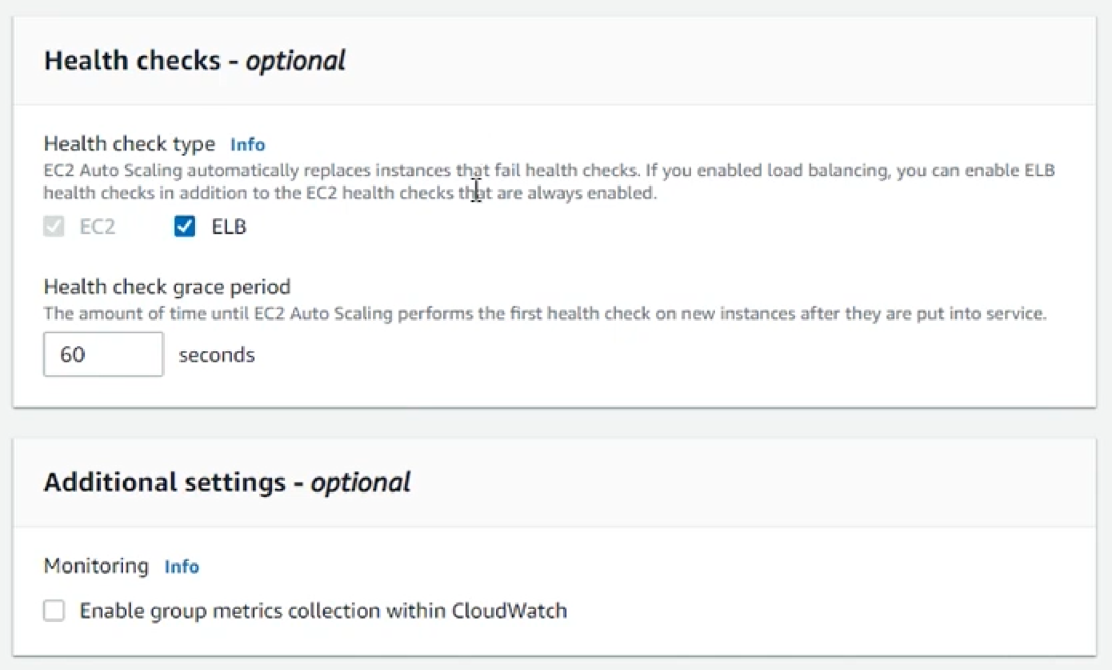
● Application Load Balancer or Network Load Balancer

○ Classic Load Balancer

Choose a target group for your load balancer

*Select target group*

[redacted]



### Health checks - *optional*

Health check type Info

EC2 Auto Scaling automatically replaces instances that fail health checks. If you enabled load balancing, you can enable ELB health checks in addition to the EC2 health checks that are always enabled.

☑ EC2 ☑ ELB

Health check grace period

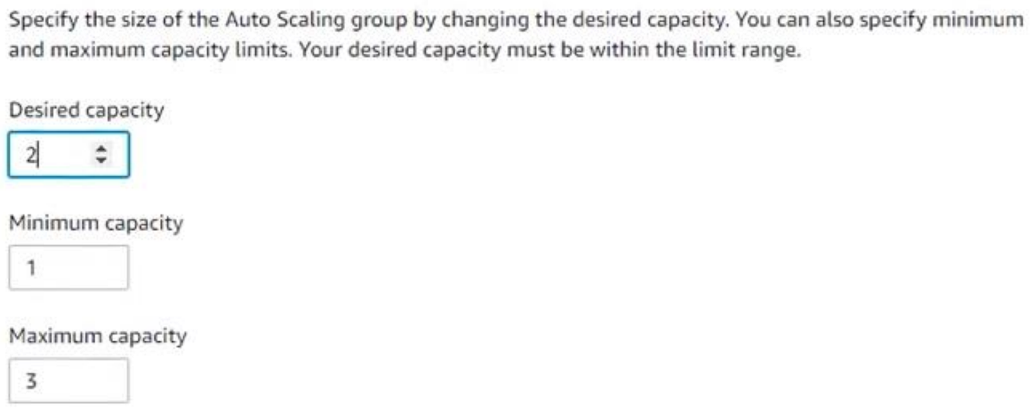
The amount of time until EC2 Auto Scaling performs the first health check on new instances after they are put into service.

60 seconds

### Additional settings - *optional*

Monitoring Info

☐ Enable group metrics collection within CloudWatch

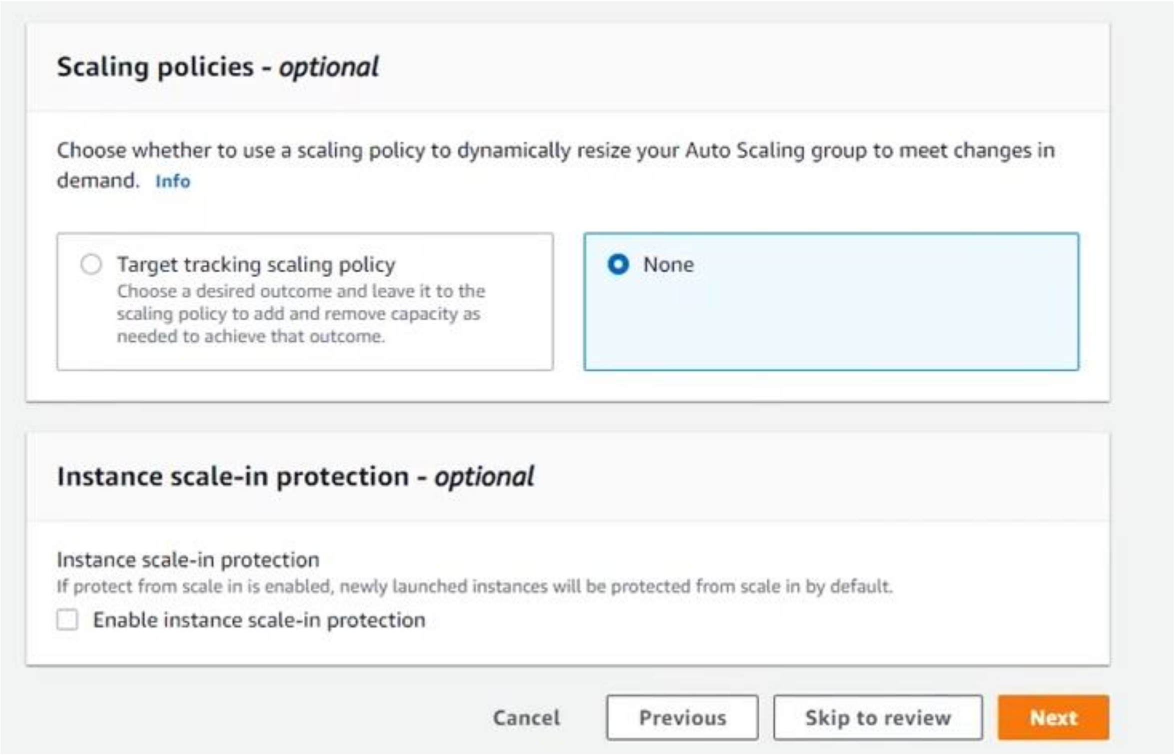


Specify the size of the Auto Scaling group by changing the desired capacity. You can also specify minimum and maximum capacity limits. Your desired capacity must be within the limit range.

Desired capacity: 2

Minimum capacity: 1

Maximum capacity: 3



### Scaling policies - *optional*

Choose whether to use a scaling policy to dynamically resize your Auto Scaling group to meet changes in demand. Info

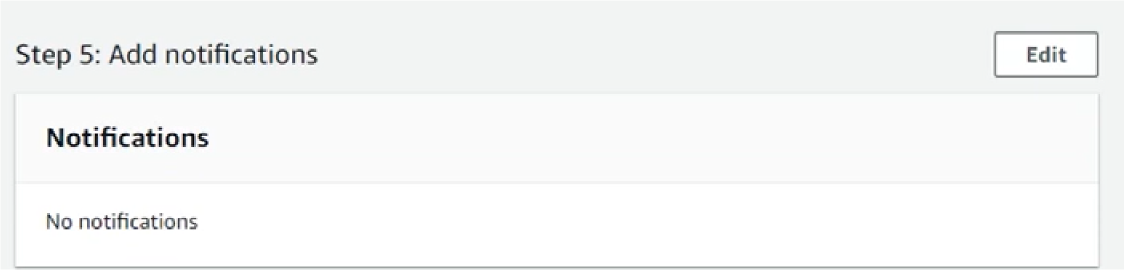
○ Target tracking scaling policy
Choose a desired outcome and leave it to the scaling policy to add and remove capacity as needed to achieve that outcome.

● None

### Instance scale-in protection - *optional*

Instance scale-in protection
If protect from scale in is enabled, newly launched instances will be protected from scale in by default.

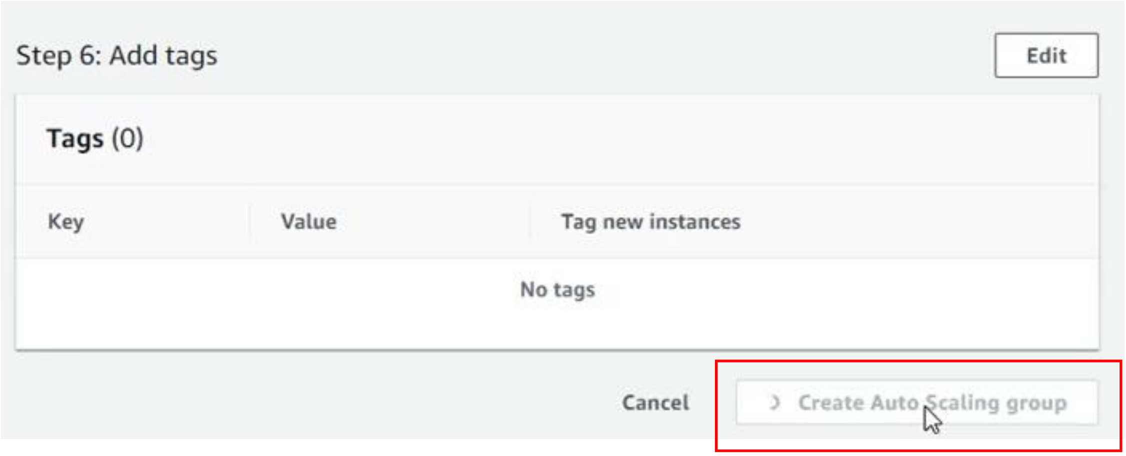
□ Enable instance scale-in protection



Step 5: Add notifications

### Notifications

No notifications



Step 6: Add tags

Edit

Tags (0)

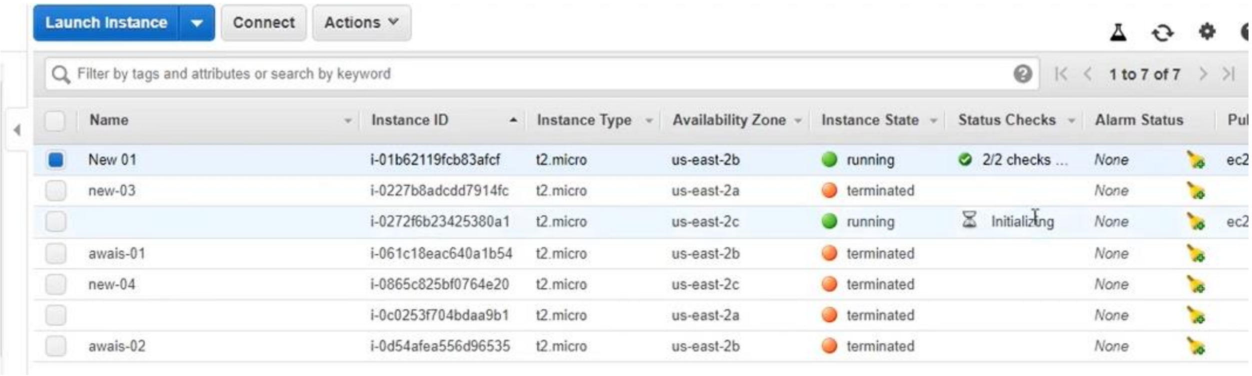
| Key | Value | Tag new instances |
|---|---|---|
| No tags | | |

Cancel

Create Auto Scaling group

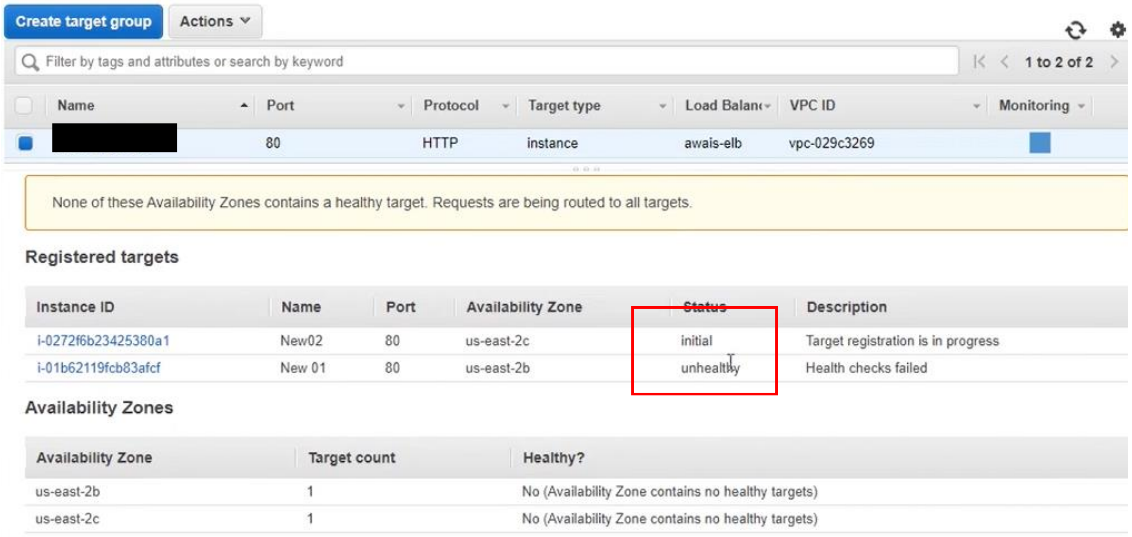
- Now visit EC2 instances to see if any instances are launched through AGS.
- Terminate any running instance to see how ASG automatically launches the servers. Delete/terminate all the servers to see new servers will start launching.

  DONOT FORGET to terminate all the servers now and release/delete all the AWS resources.



| | Name | Instance ID | Instance Type | Availability Zone | Instance State | Status Checks | Alarm Status | Pul |
|---|---|---|---|---|---|---|---|---|
| ☑ | New 01 | i-01b62119fcb83afcf | t2.micro | us-east-2b | ● running | ✓ 2/2 checks ... | None | ec2 |
| □ | new-03 | i-0227b8adcdd7914fc | t2.micro | us-east-2a | ● terminated | | None | |
| □ | | i-0272f6b23425380a1 | t2.micro | us-east-2c | ● running | ⌛ Initializing | None | ec2 |
| □ | awais-01 | i-061c18eac640a1b54 | t2.micro | us-east-2b | ● terminated | | None | |
| □ | new-04 | i-0865c825bf0764e20 | t2.micro | us-east-2c | ● terminated | | None | |
| □ | | i-0c0253f704bdaa9b1 | t2.micro | us-east-2a | ● terminated | | None | |
| □ | awais-02 | i-0d54afea556d96535 | t2.micro | us-east-2b | ● terminated | | None | |

- In target groups you can see the health status of the servers. This information takes time to sync.



| | Name | Port | Protocol | Target type | Load Balan | VPC ID | Monitoring |
|---|---|---|---|---|---|---|---|
| ☑ | [redacted] | 80 | HTTP | instance | awais-elb | vpc-029c3269 | ■ |

None of these Availability Zones contains a healthy target. Requests are being routed to all targets.

Registered targets

| Instance ID | Name | Port | Availability Zone | Status | Description |
|---|---|---|---|---|---|
| i-0272f6b23425380a1 | New02 | 80 | us-east-2c | initial | Target registration is in progress |
| i-01b62119fcb83afcf | New 01 | 80 | us-east-2b | unhealthy | Health checks failed |

Availability Zones

| Availability Zone | Target count | Healthy? |
|---|---|---|
| us-east-2b | 1 | No (Availability Zone contains no healthy targets) |
| us-east-2c | 1 | No (Availability Zone contains no healthy targets) |