SCALING AND LOAD BALANCING AN ARCHITECTURE

AWS provides the feature of auto scaling and load balancing to make your architecture more robust.

ELB automatically distributes incoming application traffic across multiple Amazon Elastic Compute Cloud (Amazon EC2) instances. ELB provides the amount of load balancing capacity needed to route application traffic to help you achieve fault tolerance in your applications.

Auto scaling can automatically increase the number of EC2 instances during spikes in demand to maintain performance and can decrease capacity during lulls to reduce costs. Auto scaling is well suited to applications that have stable demand patterns or that experience hourly, daily, or weekly variability in usage.

With that said, let’s make it happen!!

REFERANCE DIAGRAMS:

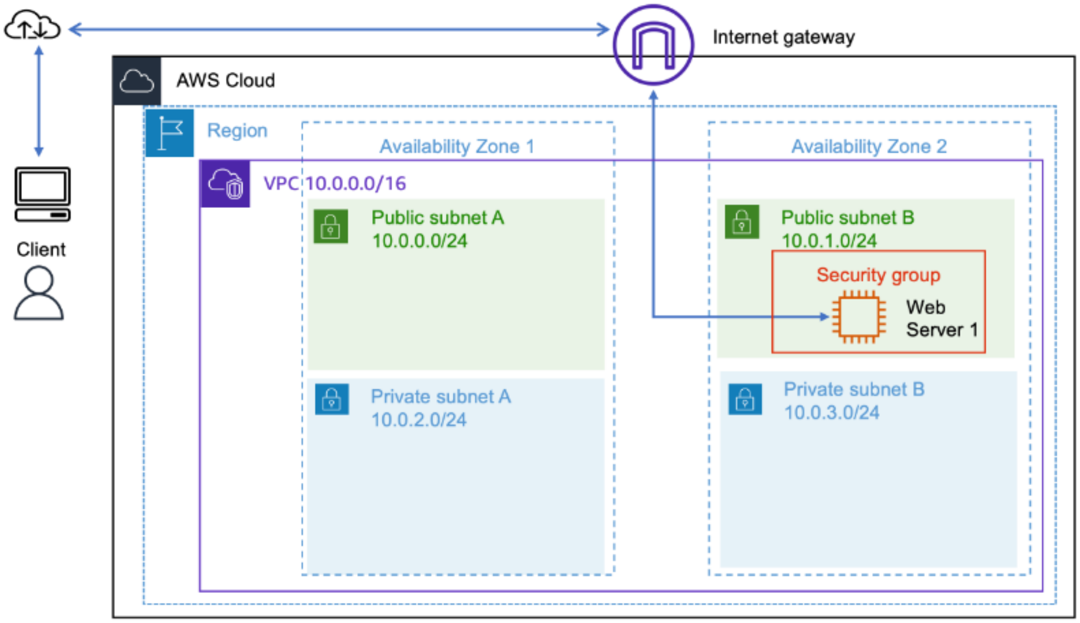


Fig 1. Initial Architecture

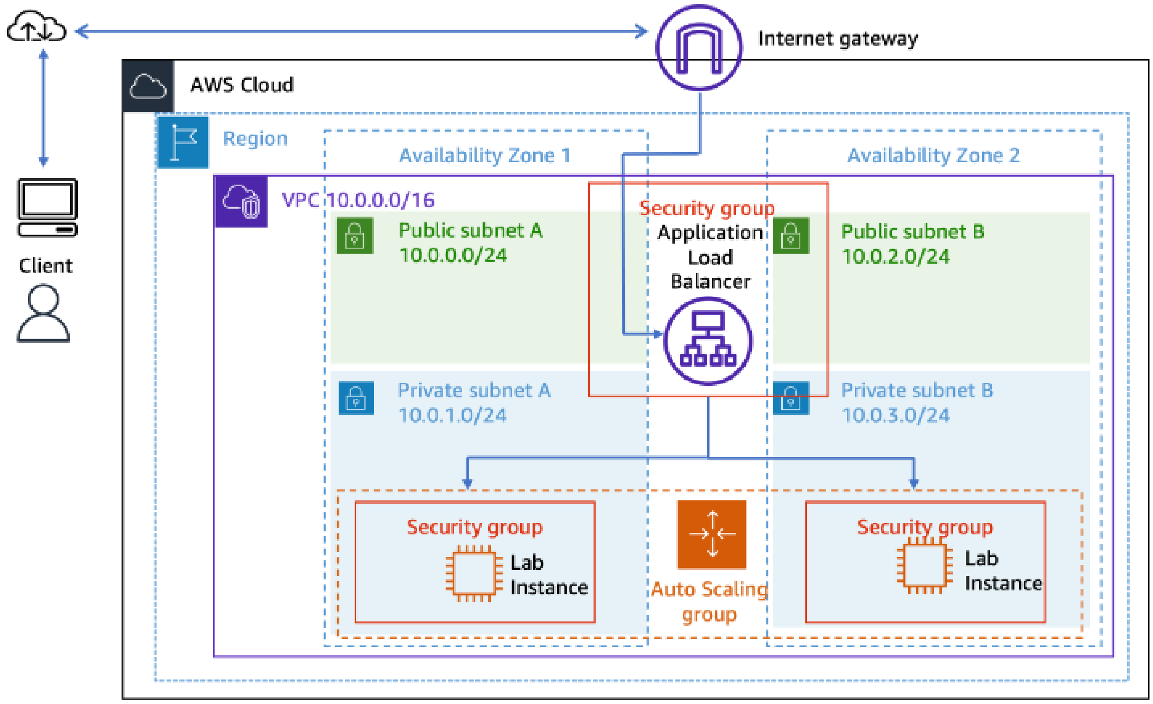
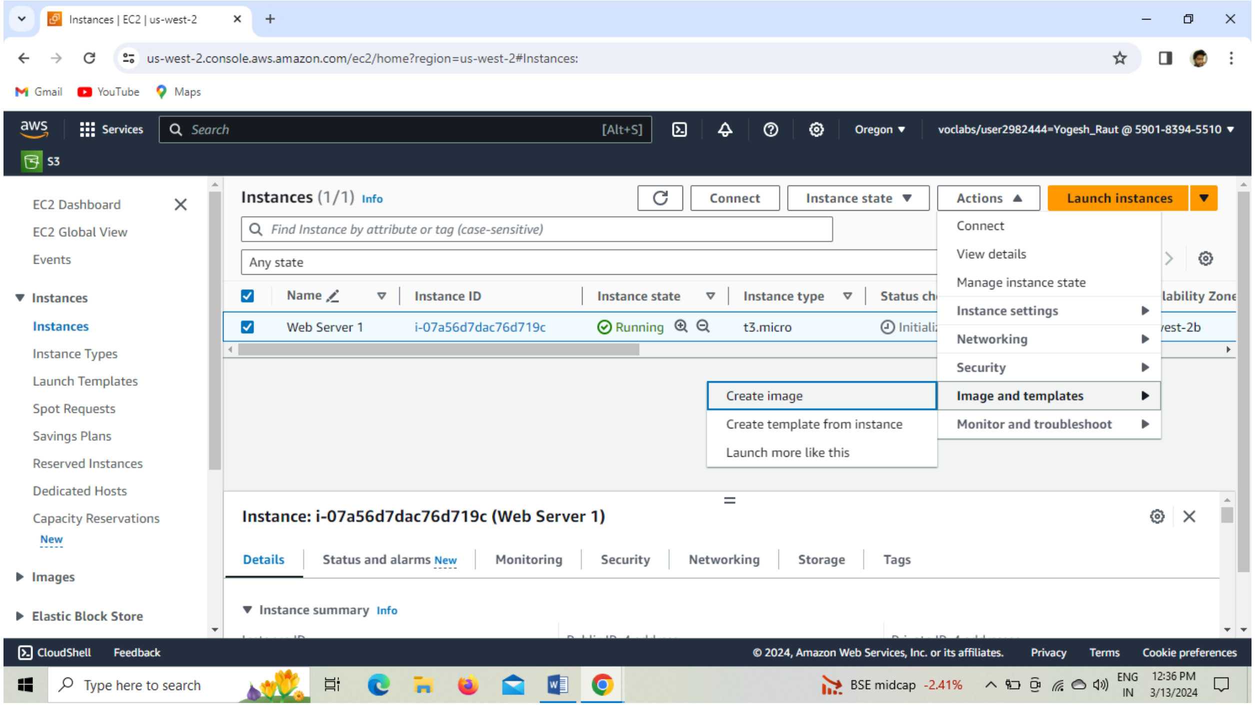


Fig 2. Final Architecture

STEPS TO FOLLOW:

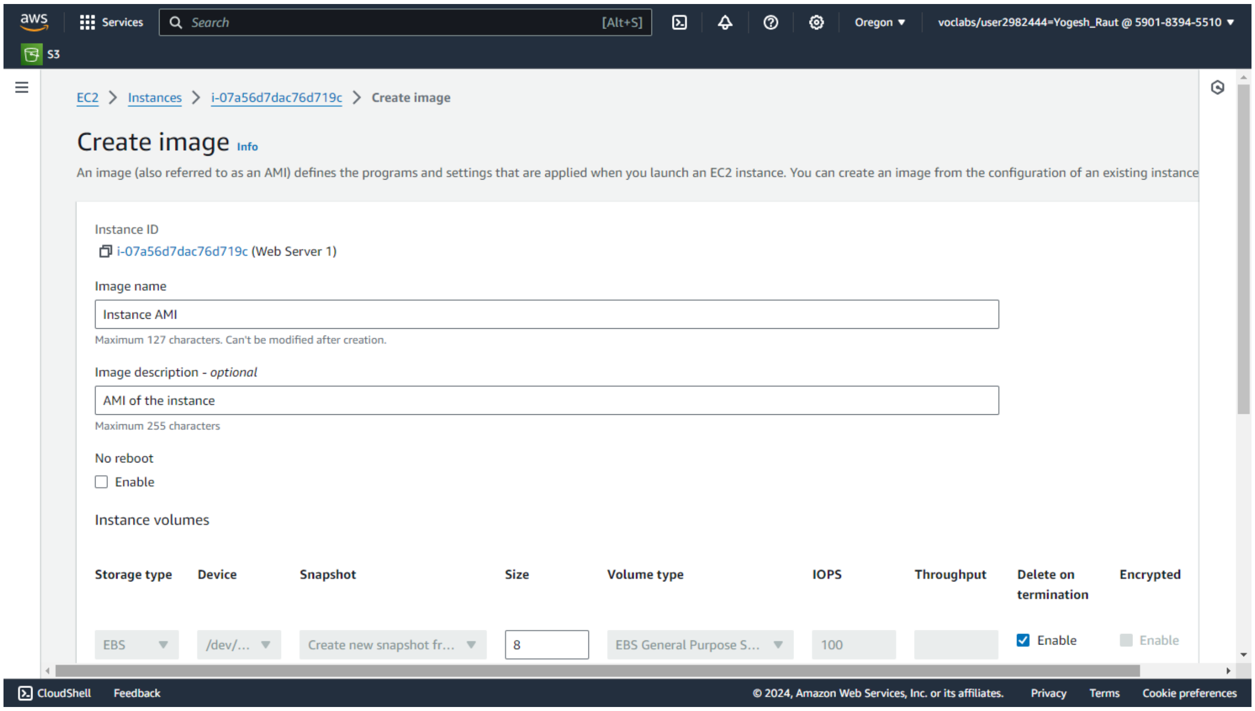
A. Creating an Amazon Machine Image (AMI) of an EC2 Instance

1. In **AWS Management Console**, search for **EC2** after which you select **Instances** from the left navigation pane to list the instances.
2. Select your instance, and from **Actions > Image and templates**, select **Create Image**. Now, configure the image as follows.
   - **I.** **Image name:** Instance AMI
   - **II.** **Description:** AMI of the instance



☑	Name	Instance ID	Instance state	Instance type	Status ch	lability Zone
☑	Web Server 1	i-07a56d7dac76d719c	Running	t3.micro	Initiali	est-2b

Instance: i-07a56d7dac76d719c (Web Server 1)


Services
Search [Alt+S]
Oregon
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510
EC2 > Instances > i-07a56d7dac76d719c > Create image
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Create image Info

An image (also referred to as an AMI) defines the programs and settings that are applied when you launch an EC2 instance. You can create an image from the configuration of an existing instance

Instance ID
i-07a56d7dac76d719c (Web Server 1)

Image name
Instance AMI
Maximum 127 characters. Can't be modified after creation.

Image description - *optional*
AMI of the instance
Maximum 255 characters

No reboot
☐ Enable

Instance volumes

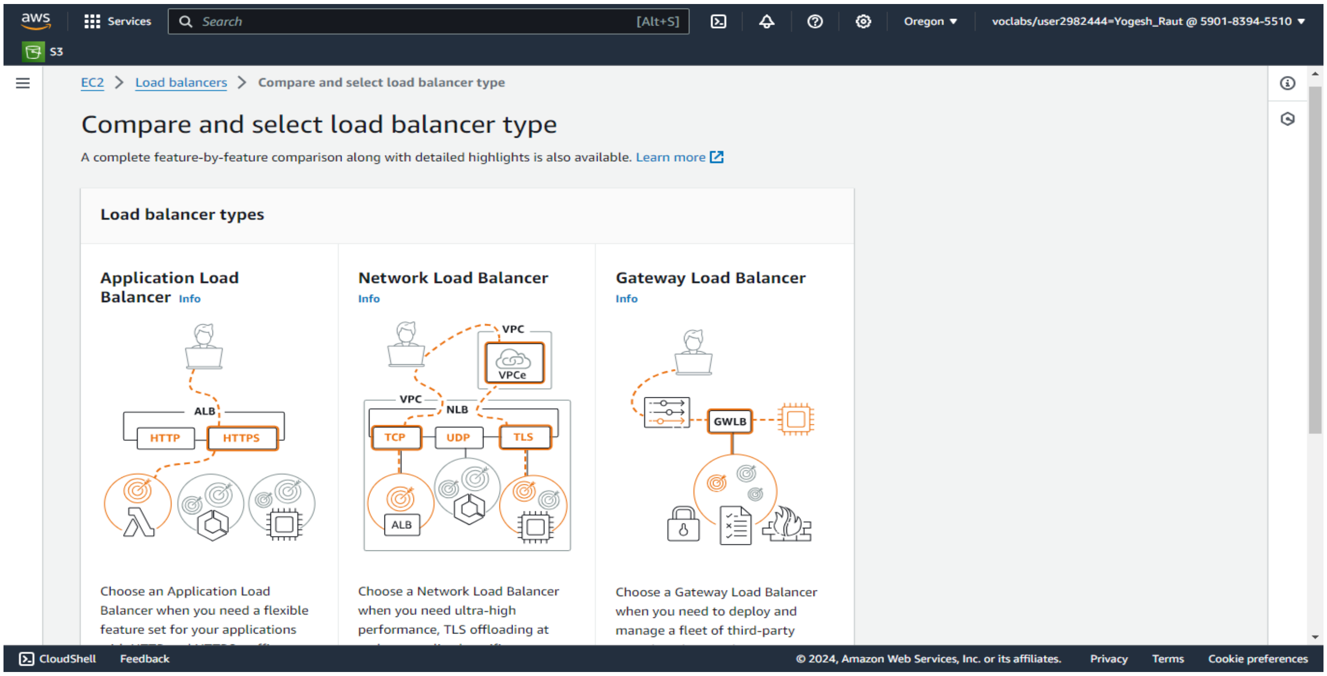
Storage type	Device	Snapshot	Size	Volume type	IOPS	Throughput	Delete on termination	Encrypted
EBS	/dev/...	Create new snapshot fr...	8	EBS General Purpose S...	100		☑ Enable	☐ Enable

3. Select **Create Image.**

B. Creating a load balancer

1. Select **Load Balancing** section, where you'll find **Load balancer** from left navigation pane.

2. Choose **Create Load balancer.** If asked for **Load balancer type**, select **Application Load Balancer**.



EC2 > Load balancers > Compare and select load balancer type

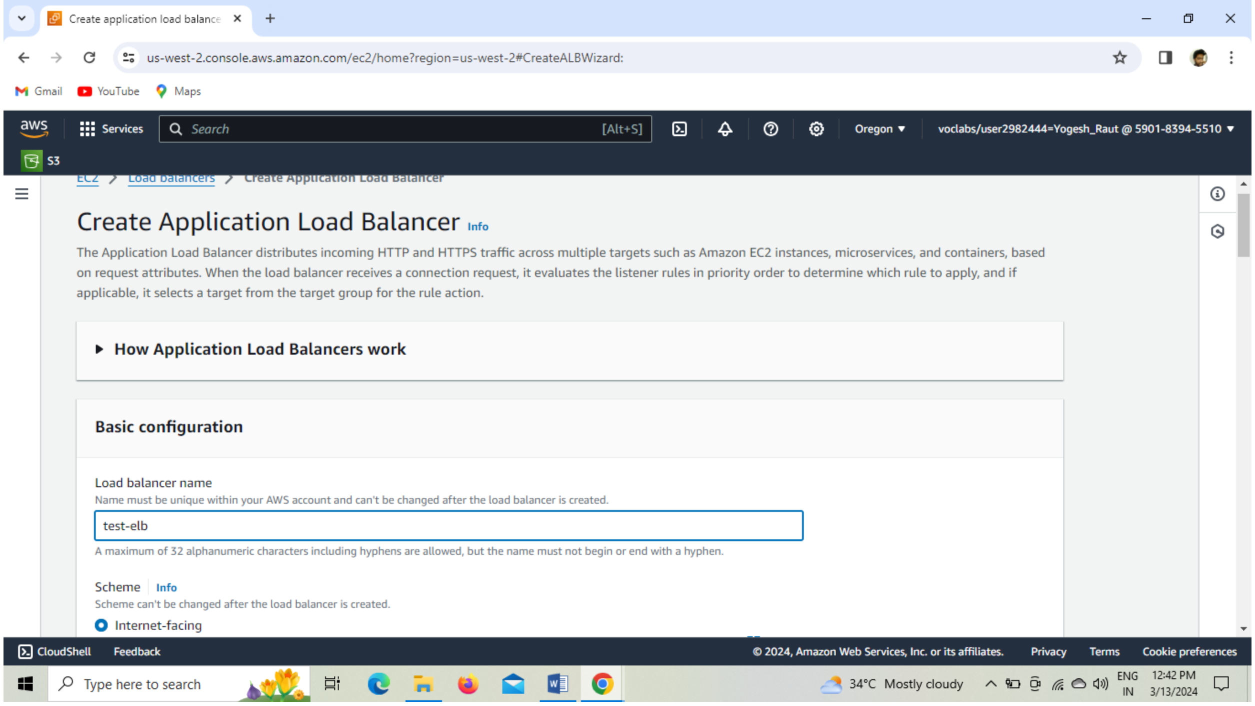
Compare and select load balancer type

A complete feature-by-feature comparison along with detailed highlights is also available. Learn more

Load balancer types

Application Load Balancer Info	Network Load Balancer Info	Gateway Load Balancer Info
Choose an Application Load Balancer when you need a flexible feature set for your applications	Choose a Network Load Balancer when you need ultra-high performance, TLS offloading at	Choose a Gateway Load Balancer when you need to deploy and manage a fleet of third-party

3. In **Basic Configuration** section, give **Load balancer name** as **test-elb**.


Create application load balance
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateALBWizard:
Gmail
YouTube
Maps
aws
Services
Search
[Alt+S]
Oregon ▼
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▼
S3
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
34°C Mostly cloudy
ENG
IN
12:42 PM
3/13/2024

EC2 > Load balancers > Create Application Load Balancer

Create Application Load Balancer Info

The Application Load Balancer distributes incoming HTTP and HTTPS traffic across multiple targets such as Amazon EC2 instances, microservices, and containers, based on request attributes. When the load balancer receives a connection request, it evaluates the listener rules in priority order to determine which rule to apply, and if applicable, it selects a target from the target group for the rule action.

▸ How Application Load Balancers work

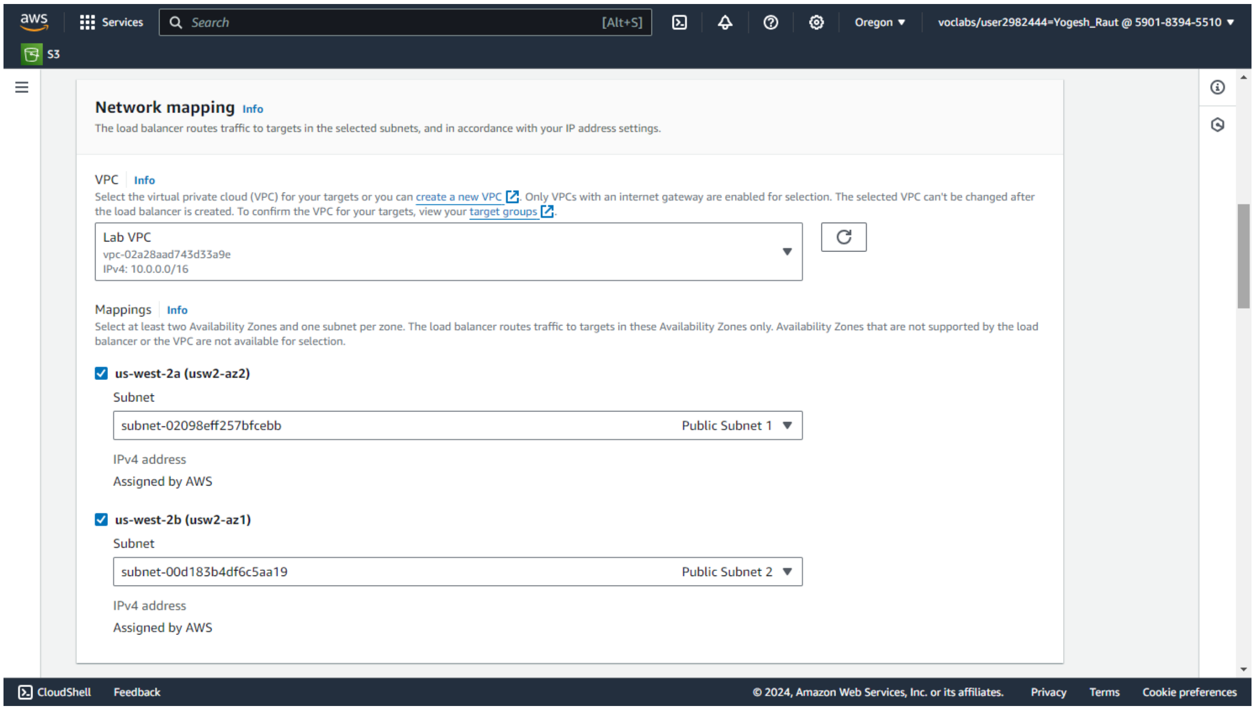
Basic configuration

Load balancer name
Name must be unique within your AWS account and can't be changed after the load balancer is created.
test-elb
A maximum of 32 alphanumeric characters including hyphens are allowed, but the name must not begin or end with a hyphen.

Scheme Info
Scheme can't be changed after the load balancer is created.
● Internet-facing

4. In **Network mapping** section, configure the settings as follows:
   I. **VPC:** Select your VPC.

II. **Mappings:** Choose both Availability Zones. For **AZ1** and **AZ2,** choose **Public Subnets 1 and 2** respectively.



Network mapping Info

The load balancer routes traffic to targets in the selected subnets, and in accordance with your IP address settings.

VPC Info

Select the virtual private cloud (VPC) for your targets or you can create a new VPC. Only VPCs with an internet gateway are enabled for selection. The selected VPC can't be changed after the load balancer is created. To confirm the VPC for your targets, view your target groups.

Lab VPC
vpc-02a28aad743d33a9e
IPv4: 10.0.0.0/16

Mappings Info

Select at least two Availability Zones and one subnet per zone. The load balancer routes traffic to targets in these Availability Zones only. Availability Zones that are not supported by the load balancer or the VPC are not available for selection.

☑ us-west-2a (usw2-az2)

Subnet: subnet-02098eff257bfcebb Public Subnet 1

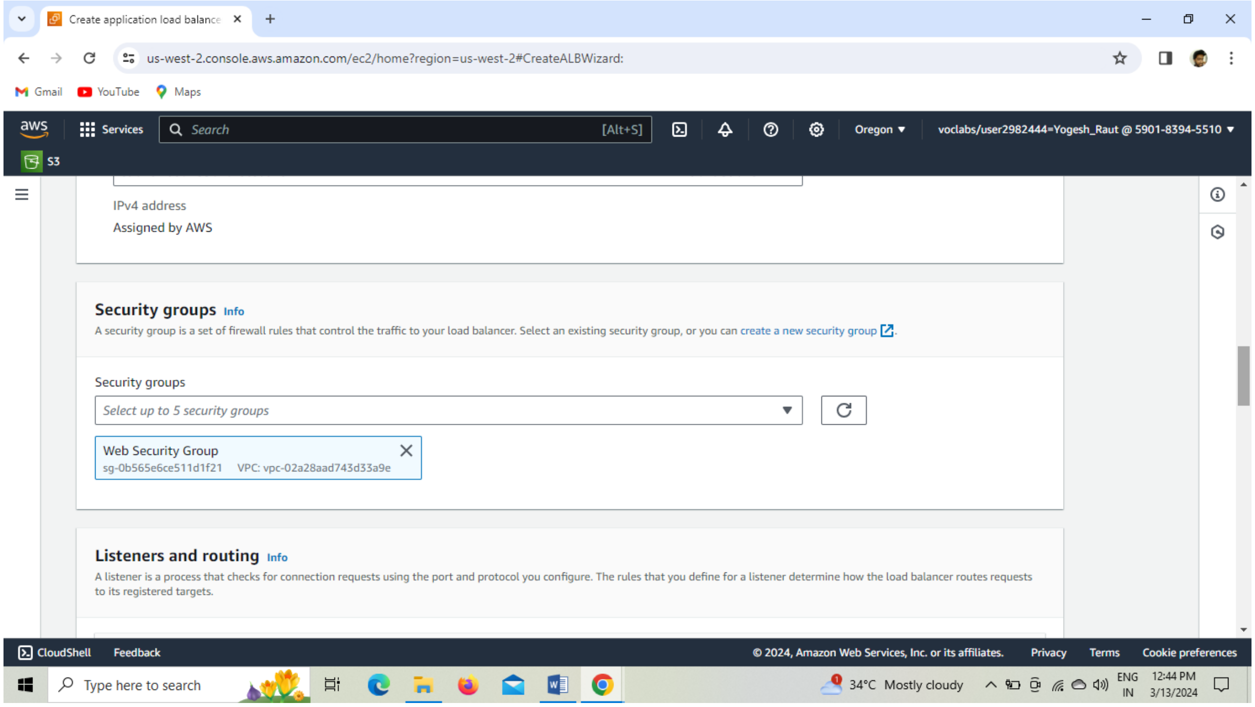
IPv4 address: Assigned by AWS

☑ us-west-2b (usw2-az1)

Subnet: subnet-00d183b4df6c5aa19 Public Subnet 2

IPv4 address: Assigned by AWS

5. **Security Groups:** Web Security Group (**HTTP permitted**)



Create application load balance
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateALBWizard:
Gmail
YouTube
Maps
aws
Services
Search [Alt+S]
Oregon ▼
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▼
S3

IPv4 address: Assigned by AWS

Security groups Info

A security group is a set of firewall rules that control the traffic to your load balancer. Select an existing security group, or you can create a new security group.

Security groups: Select up to 5 security groups

Web Security Group
sg-0b565e6ce511d1f21 VPC: vpc-02a28aad743d33a9e

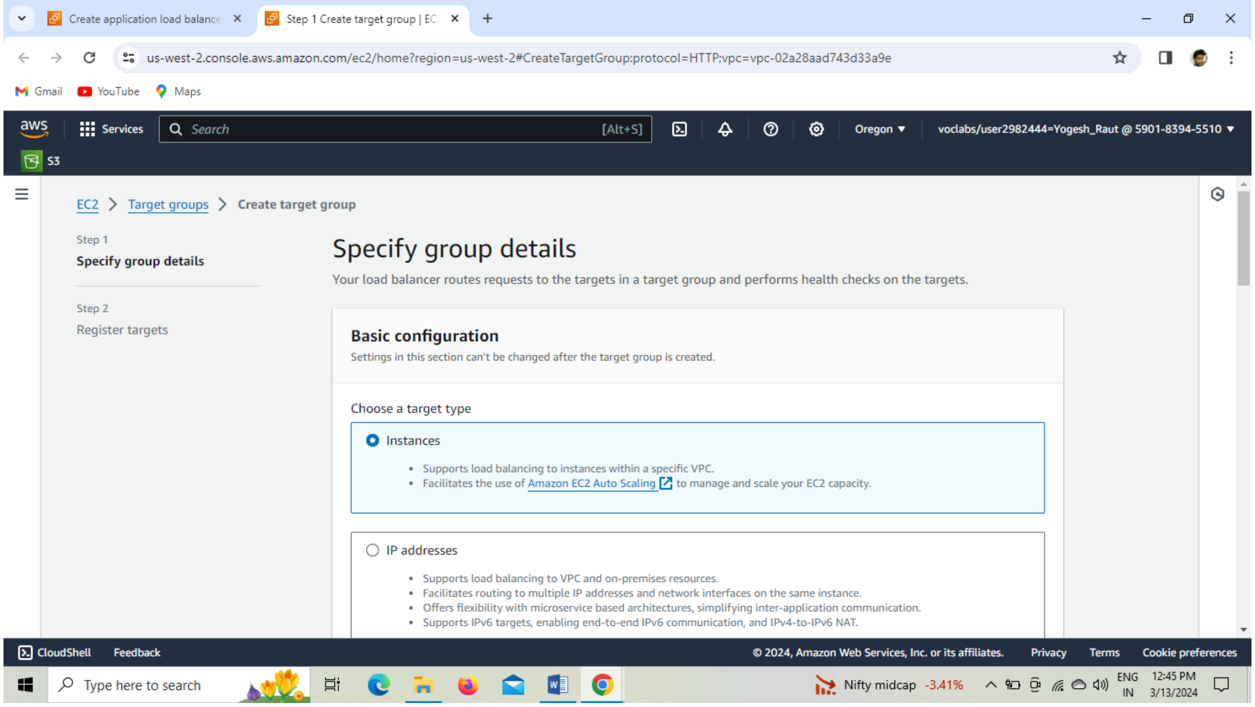
Listeners and routing Info

A listener is a process that checks for connection requests using the port and protocol you configure. The rules that you define for a listener determine how the load balancer routes requests to its registered targets.

6. In the **Listeners and routing** section, choose the **Create target group** link.

7. On the new **Target groups browser tab,** in the **Basic configuration** section, configure the following:
   - I. **Target type:** Instances
   - II. **Target Group Name:** test-target-group

Click **Next.**



us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateTargetGroup:protocol=HTTP;vpc=vpc-02a28aad743d33a9e

EC2 > Target groups > Create target group

Step 1
Specify group details

Step 2
Register targets

Specify group details

Your load balancer routes requests to the targets in a target group and performs health checks on the targets.

Basic configuration

Settings in this section can't be changed after the target group is created.

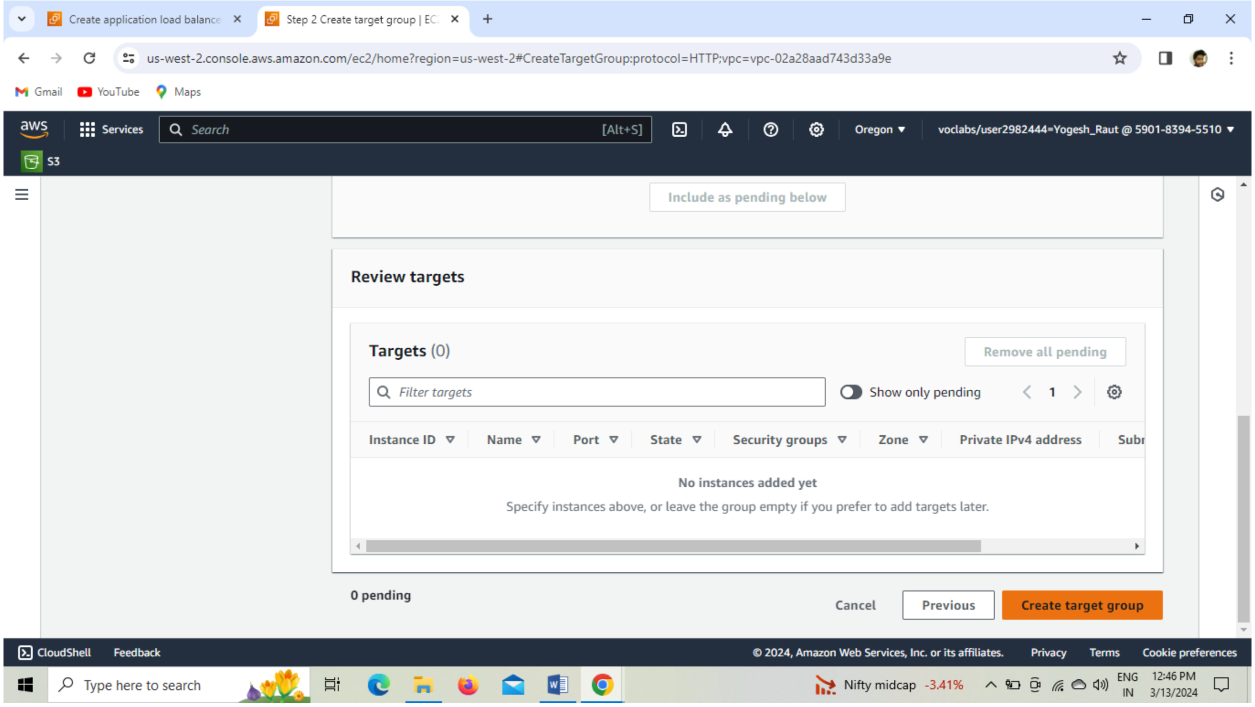
Choose a target type

● Instances
- Supports load balancing to instances within a specific VPC.
- Facilitates the use of Amazon EC2 Auto Scaling to manage and scale your EC2 capacity.

○ IP addresses
- Supports load balancing to VPC and on-premises resources.
- Facilitates routing to multiple IP addresses and network interfaces on the same instance.
- Offers flexibility with microservice based architectures, simplifying inter-application communication.
- Supports IPv6 targets, enabling end-to-end IPv6 communication, and IPv4-to-IPv6 NAT.

Create application load balance
Step 1 Create target group | EC
Gmail
YouTube
Maps
aws
Services
Search
[Alt+S]
Oregon ▼
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▼
S3
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
Nifty midcap -3.41%
ENG
IN
12:45 PM
3/13/2024

8. On the **Register targets** page, choose **Create target group.** Once the target group has been created successfully, close the **Target groups browser tab.**


Create application load balance
Step 2 Create target group | EC
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateTargetGroup:protocol=HTTP;vpc=vpc-02a28aad743d33a9e
Gmail
YouTube
Maps
Services
Search
[Alt+S]
Oregon
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510
S3
Include as pending below

Review targets

Targets (0)

Remove all pending
Filter targets
Show only pending
1

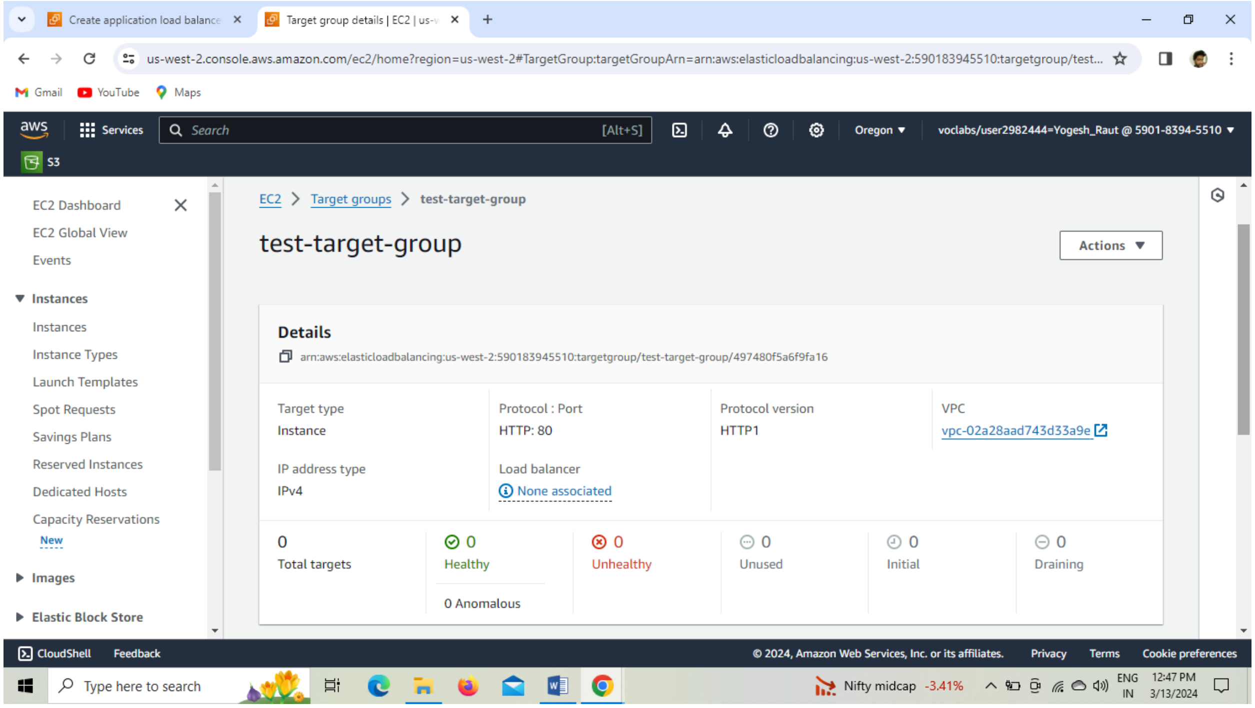
Instance ID	Name	Port	State	Security groups	Zone	Private IPv4 address	Subr

No instances added yet

Specify instances above, or leave the group empty if you prefer to add targets later.

0 pending

Cancel
Previous
Create target group
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
Nifty midcap -3.41%
ENG IN
12:46 PM
3/13/2024


Create application load balance
Target group details | EC2 | us-
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#TargetGroup:targetGroupArn=arn:aws:elasticloadbalancing:us-west-2:590183945510:targetgroup/test...
Gmail
YouTube
Maps
Services
Search
[Alt+S]
Oregon
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510
S3
EC2 Dashboard
EC2 Global View
Events
Instances
Instances
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances
Dedicated Hosts
Capacity Reservations
New
Images
Elastic Block Store
EC2 > Target groups > test-target-group
Actions

test-target-group

Details

arn:aws:elasticloadbalancing:us-west-2:590183945510:targetgroup/test-target-group/497480f5a6f9fa16

Target type: Instance

Protocol : Port: HTTP: 80

Protocol version: HTTP1

VPC: vpc-02a28aad743d33a9e

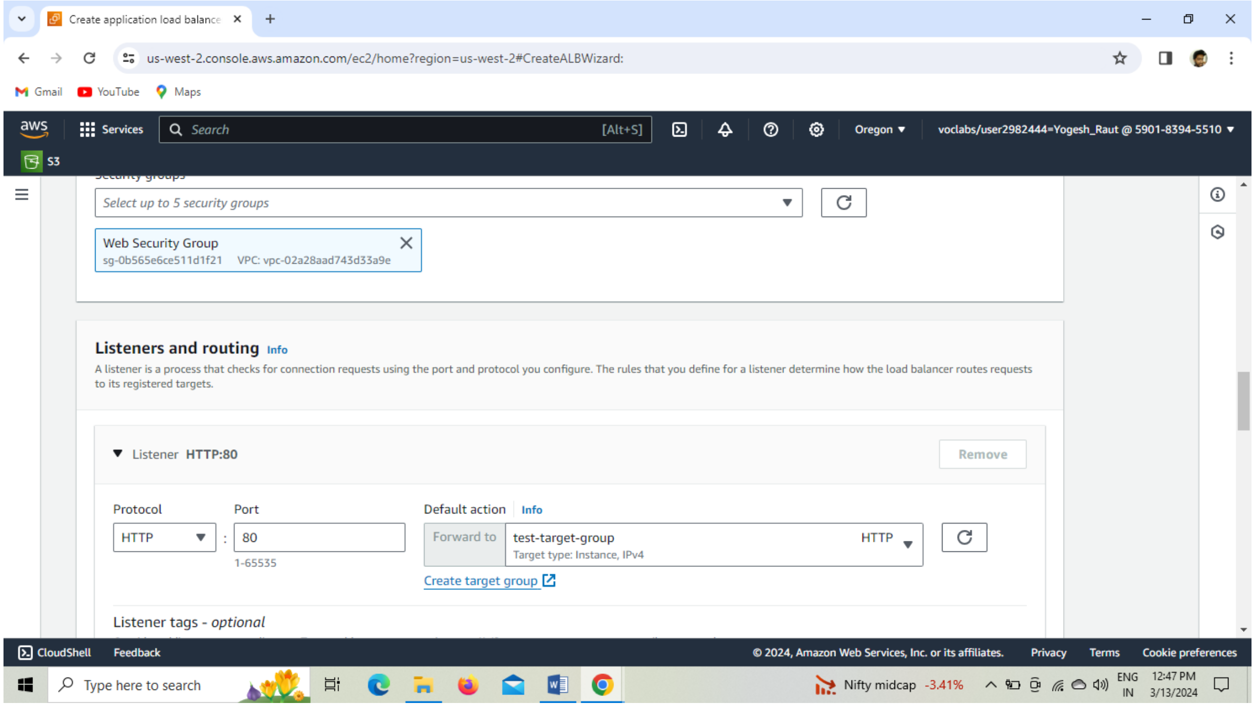
IP address type: IPv4

Load balancer: None associated

0	0	0	0	0	0
Total targets	Healthy 0 Anomalous	Unhealthy	Unused	Initial	Draining

CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
Nifty midcap -3.41%
ENG IN
12:47 PM
3/13/2024

9. Return to the **Load balancers browser tab**. In the **Listeners and routing** section, choose **Refresh.** Now from the **Forward to dropdown list,** choose **test-target-group** and choose **Create Load Balancer.**


Create application load balance
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateALBWizard:
Gmail
YouTube
Maps
Services
Search
[Alt+S]
Oregon
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510
S3

Select up to 5 security groups

Web Security Group
sg-0b565e6ce511d1f21 VPC: vpc-02a28aad743d33a9e

Listeners and routing Info

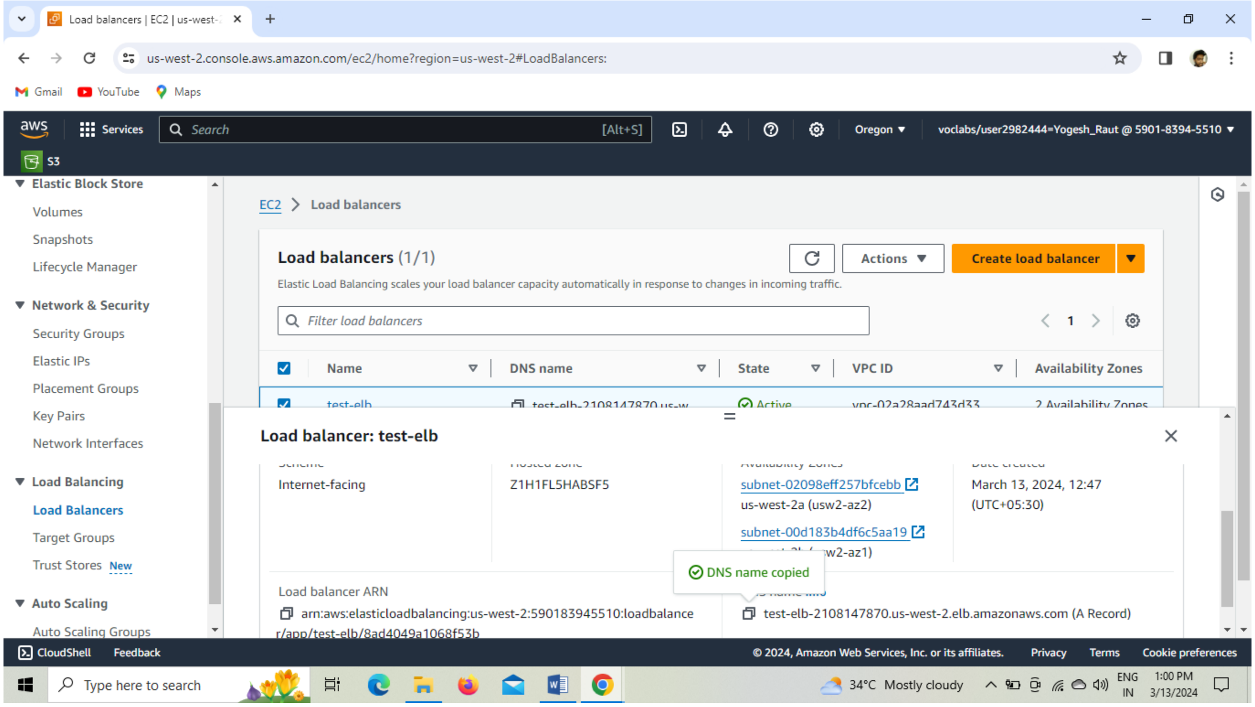
A listener is a process that checks for connection requests using the port and protocol you configure. The rules that you define for a listener determine how the load balancer routes requests to its registered targets.

▼ Listener HTTP:80 Remove

Protocol	Port	Default action Info
HTTP	80 1-65535	Forward to test-target-group HTTP Target type: Instance, IPv4 Create target group

Listener tags - *optional*

10. To view the load balancer that you created, choose **View load balancer** and copy **DNS Name.**



EC2 > Load balancers

Load balancers (1/1)

Actions ▼ Create load balancer

Elastic Load Balancing scales your load balancer capacity automatically in response to changes in incoming traffic.

Filter load balancers

☑	Name	DNS name	State	VPC ID	Availability Zones
☑	test-elb	test-elb-2108147870.us-w...	Active	vpc-02a28aad743d33...	2 Availability Zones

Load balancer: test-elb

Scheme	Hosted zone	Availability Zones	Date created
Internet-facing	Z1H1FL5HABSF5	subnet-02098eff257bfcebb us-west-2a (usw2-az2) subnet-00d183b4df6c5aa19 ...w2-az1)	March 13, 2024, 12:47 (UTC+05:30)

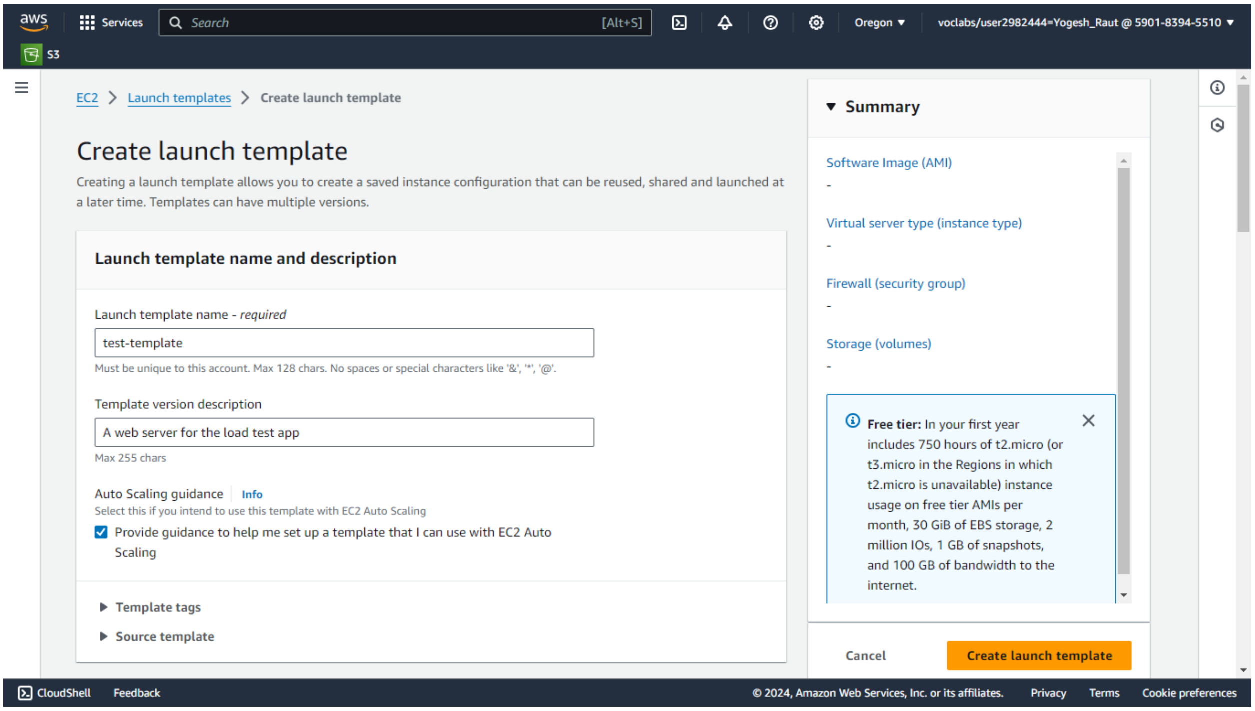
DNS name copied

Load balancer ARN
arn:aws:elasticloadbalancing:us-west-2:590183945510:loadbalancer/app/test-elb/8ad4049a1068f53b

test-elb-2108147870.us-west-2.elb.amazonaws.com (A Record)

C. Create a launch template

1. At the top of the **AWS Management Console**, in the search bar, enter and choose **EC2.**
2. In the left navigation pane, locate the **Instances** section, and choose **Launch Templates** and select **Create launch template**.


aws
Services
Search
[Alt+S]
Oregon ▼
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▼
S3
EC2 > Launch templates > Create launch template
Cancel
Create launch template
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Create launch template

Creating a launch template allows you to create a saved instance configuration that can be reused, shared and launched at a later time. Templates can have multiple versions.

Launch template name and description

Launch template name - *required*

test-template

Must be unique to this account. Max 128 chars. No spaces or special characters like '&', '*', '@'.

Template version description

A web server for the load test app

Max 255 chars

Auto Scaling guidance Info

Select this if you intend to use this template with EC2 Auto Scaling

☑ Provide guidance to help me set up a template that I can use with EC2 Auto Scaling

▶ Template tags

▶ Source template

▼ Summary

Software Image (AMI)
-

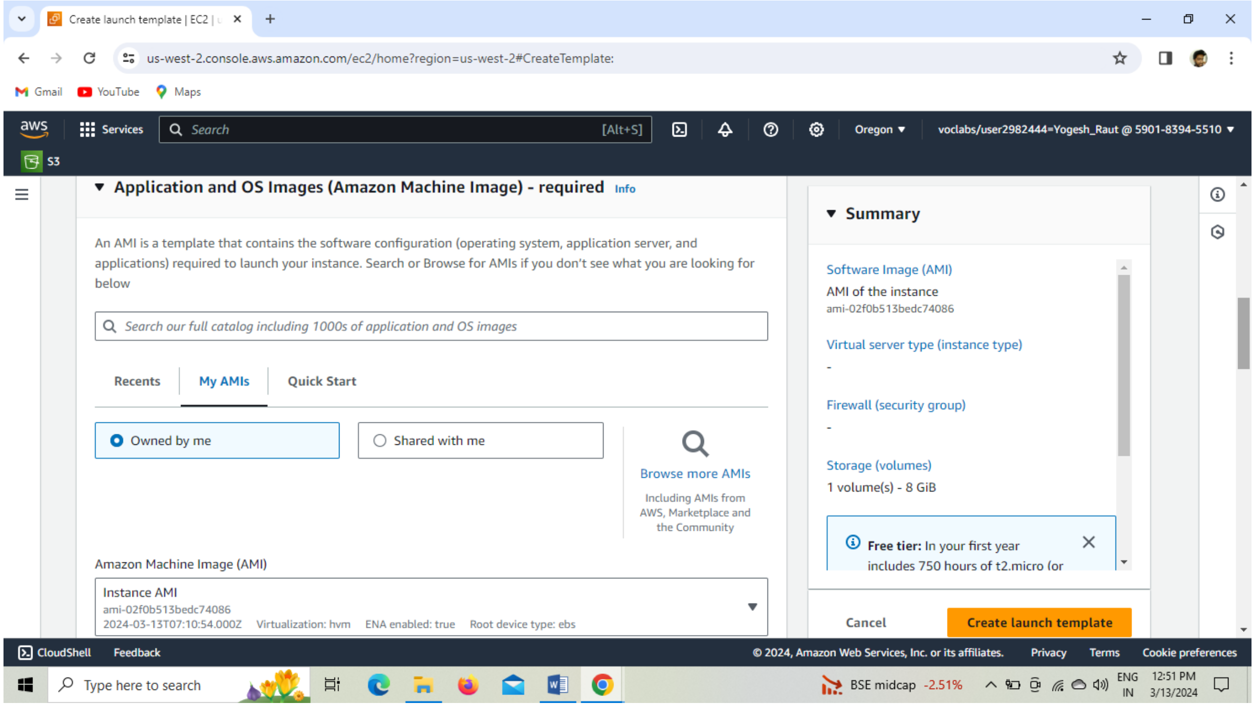
Virtual server type (instance type)
-

Firewall (security group)
-

Storage (volumes)
-

Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million IOs, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

3. Now **configure the launch template** as follows:
   - **I.** **Launch template name:** test-template
   - **II.** **Template version description:** A web server for the load test app
   - **III.** **Auto Scaling guidance:** Provide guidance to help me set up a template that I can use with EC2 Auto Scaling.
   - **IV.** **Application and OS Images (Amazon Machine Image):** Instance AMI
   - **V.** **Instance Type:** t3.micro
   - **VI.** **Key pair name:** Don't include in launch template**.**
   - **VII.** **Network Settings > Security Group:** Web Security Group.



Create launch template | EC2

us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#CreateTemplate:

Oregon ▾ voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▾

▾ Application and OS Images (Amazon Machine Image) - required Info

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. Search or Browse for AMIs if you don't see what you are looking for below

Search our full catalog including 1000s of application and OS images

Recents | My AMIs | Quick Start

● Owned by me ○ Shared with me

Browse more AMIs
Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)
Instance AMI
ami-02f0b513bedc74086
2024-03-13T07:10:54.000Z Virtualization: hvm ENA enabled: true Root device type: ebs

▾ Summary

Software Image (AMI)
AMI of the instance
ami-02f0b513bedc74086

Virtual server type (instance type)
-

Firewall (security group)
-

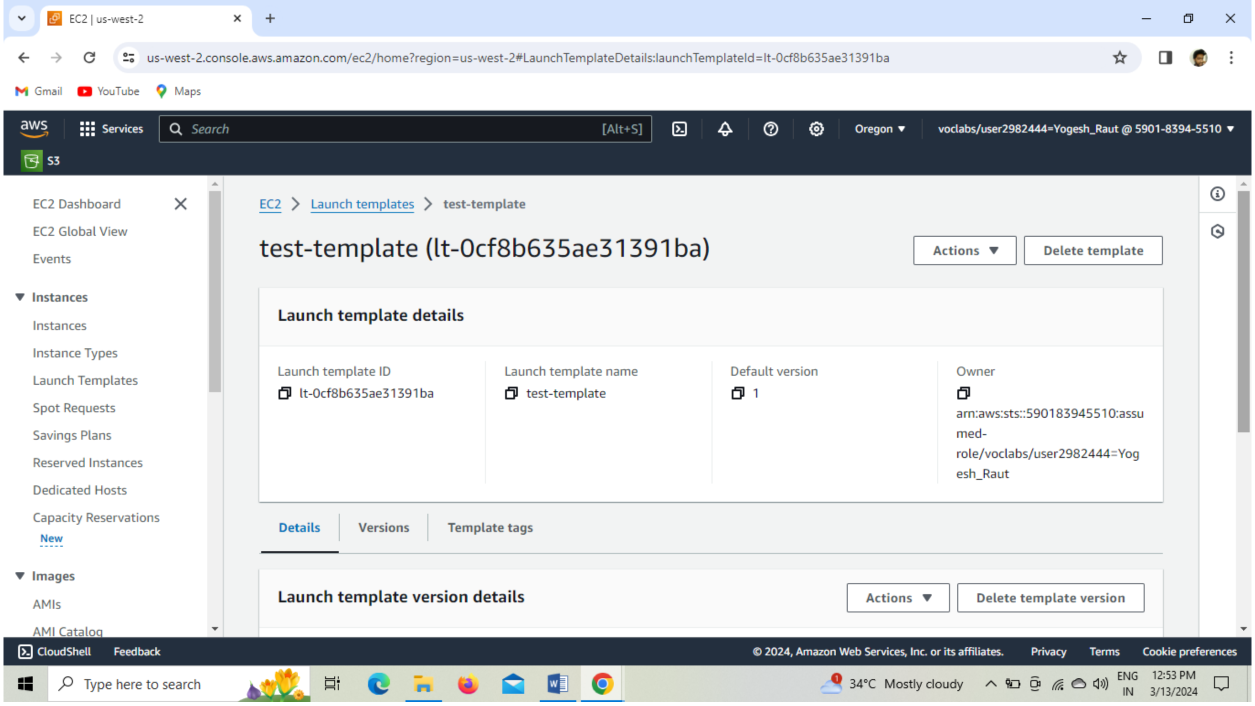
Storage (volumes)
1 volume(s) - 8 GiB

Free tier: In your first year includes 750 hours of t2.micro (or

Cancel | Create launch template

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

4. Select **Create launch template.**



EC2 | us-west-2

us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#LaunchTemplateDetails:launchTemplateId=lt-0cf8b635ae31391ba

Oregon ▾ voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510 ▾

EC2 > Launch templates > test-template

test-template (lt-0cf8b635ae31391ba)

Actions ▾ | Delete template

Launch template details

Launch template ID	Launch template name	Default version	Owner
lt-0cf8b635ae31391ba	test-template	1	arn:aws:sts::590183945510:assumed-role/voclabs/user2982444=Yogesh_Raut

Details | Versions | Template tags

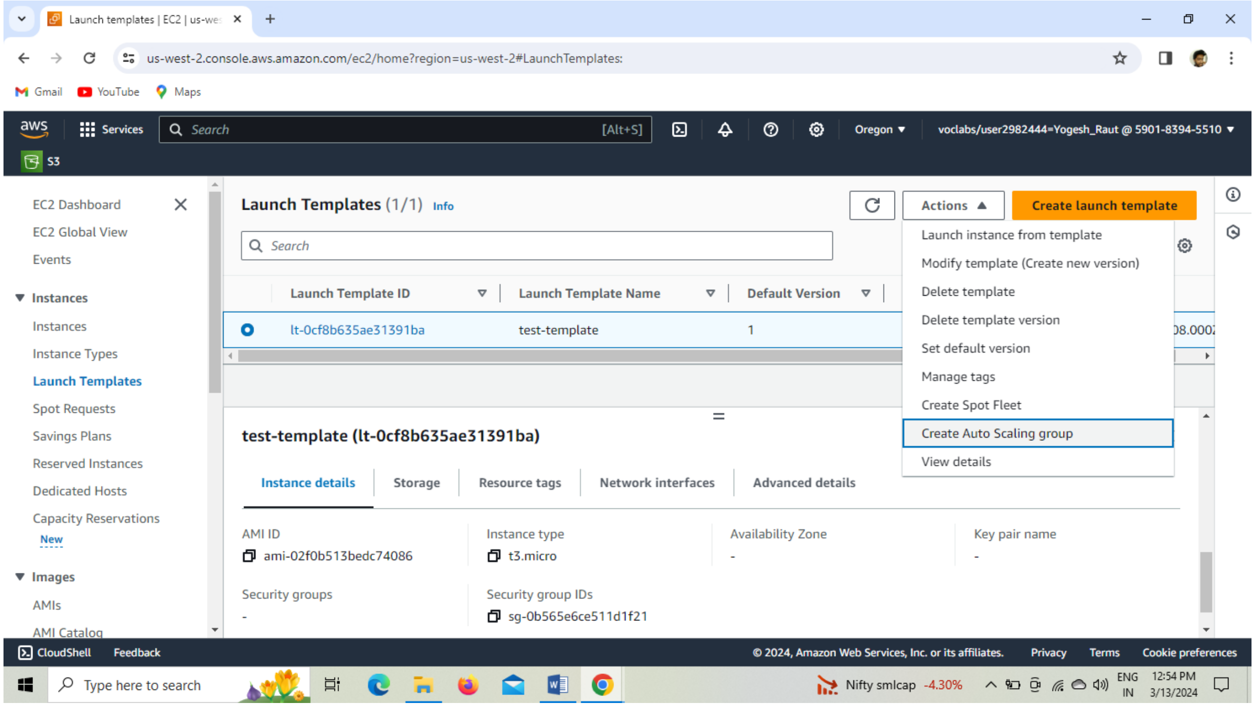
Launch template version details

Actions ▾ | Delete template version

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

D. Creating an Auto Scaling group

1. Select the **launch template** you created, and choose **Create Auto Scaling group** from **Actions** tab.



Launch Template ID	Launch Template Name	Default Version
● lt-0cf8b635ae31391ba	test-template	1

test-template (lt-0cf8b635ae31391ba)

AMI ID: ami-02f0b513bedc74086
Instance type: t3.micro
Availability Zone: -
Key pair name: -
Security groups: -
Security group IDs: sg-0b565e6ce511d1f21

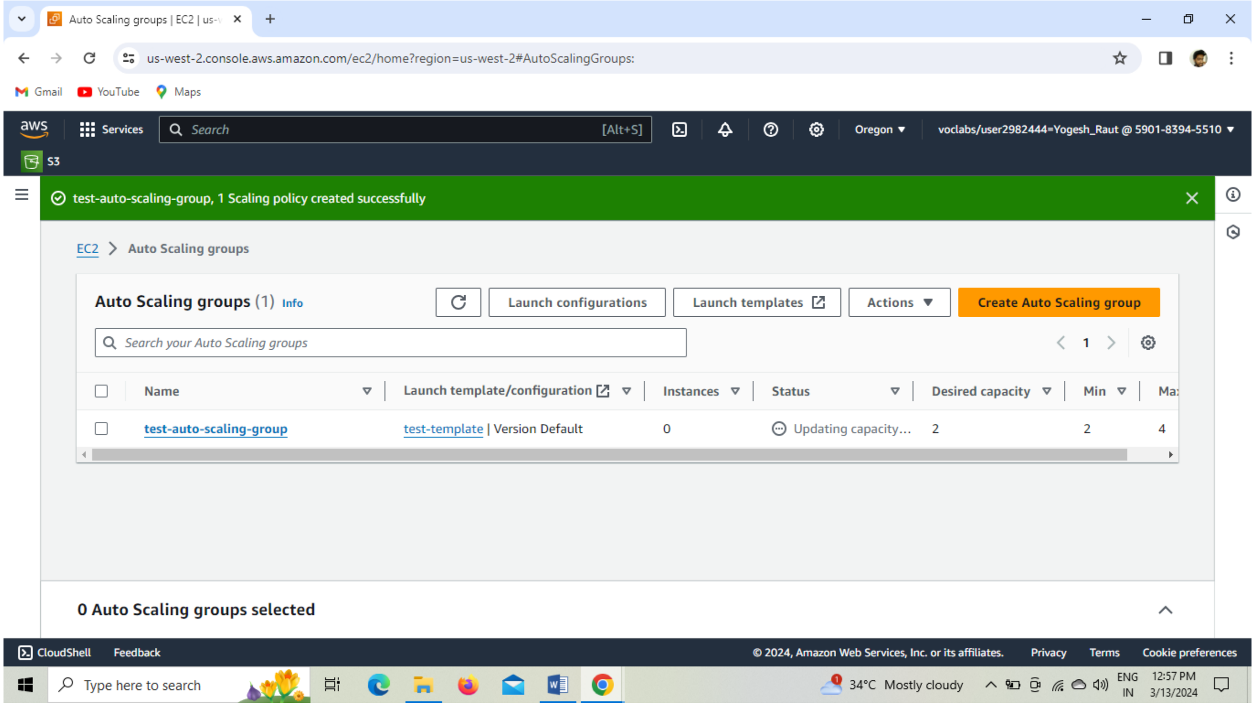
2. Configure the group as follows:

   I. **Name:** test-auto-scaling-group
   II. **Network:**
      a. **VPC:** Lab VPC
      b. **Availability Zones and Subnets:** Private Subnets 1 and 2.

   III. **Configure advanced options:**
      a. **Load balancing:** Attach to an existing load balancer.
         - **Choose from your load balancer target groups:** test-target-group
   IV. **Health check type:** ELB
   V. **Group size:**
      a. **Desired capacity:** 2
      b. **Minimum capacity:** 2
      c. **Maximum capacity:** 4

   VI. **Scaling policies:** Target tracking scaling policy
   VII. **Metric type:** Average CPU utilization.
   VIII. **Target Value:** 50

3. Choose **Create Auto Scaling group.**


Auto Scaling groups | EC2 | us-
us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#AutoScalingGroups:
Gmail
YouTube
Maps
aws
Services
Search
[Alt+S]
Oregon
voclabs/user2982444=Yogesh_Raut @ 5901-8394-5510
S3

test-auto-scaling-group, 1 Scaling policy created successfully

EC2 > Auto Scaling groups

Auto Scaling groups (1) Info

Launch configurations
Launch templates
Actions
Create Auto Scaling group
Search your Auto Scaling groups
1

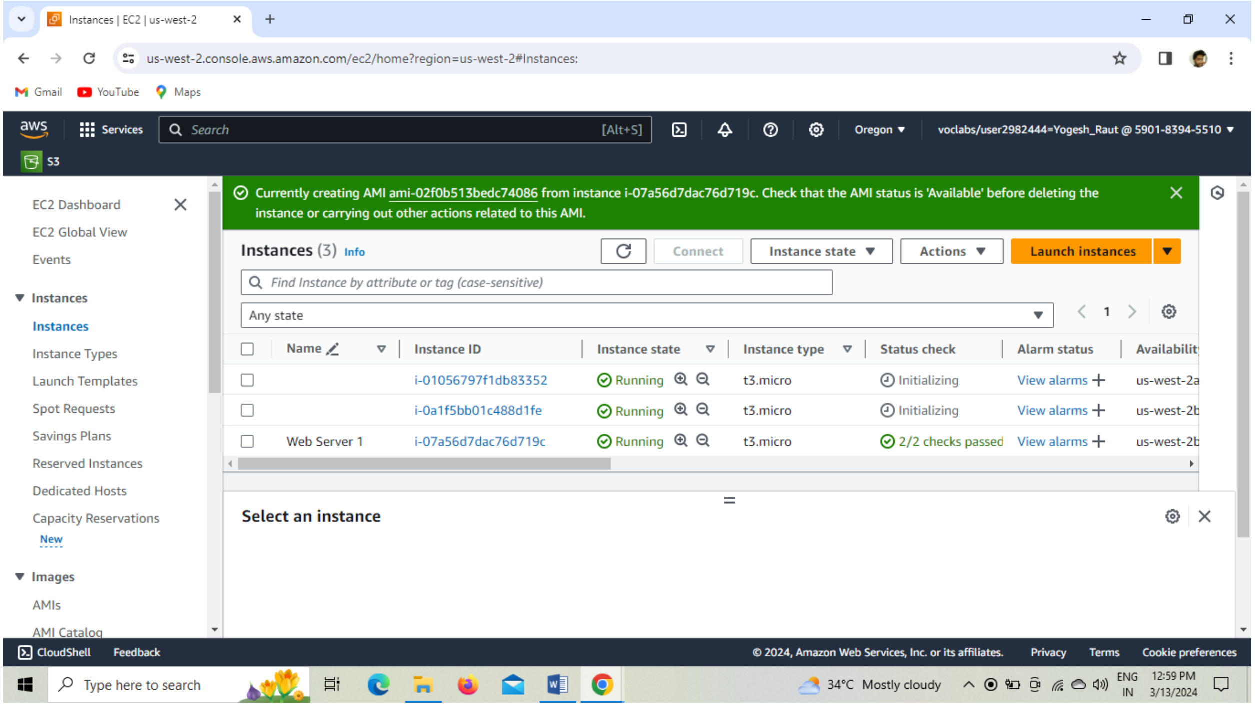
□	Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Ma:
□	test-auto-scaling-group	test-template | Version Default	0	Updating capacity...	2	2	4

0 Auto Scaling groups selected

CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
Type here to search
34°C Mostly cloudy
ENG
IN
12:57 PM
3/13/2024

E. Verifying if Load Balancer works:

1. To verify load balancer, **check Instances** if additional instances are added.



Currently creating AMI ami-02f0b513bedc74086 from instance i-07a56d7dac76d719c. Check that the AMI status is 'Available' before deleting the instance or carrying out other actions related to this AMI.

Instances (3) Info

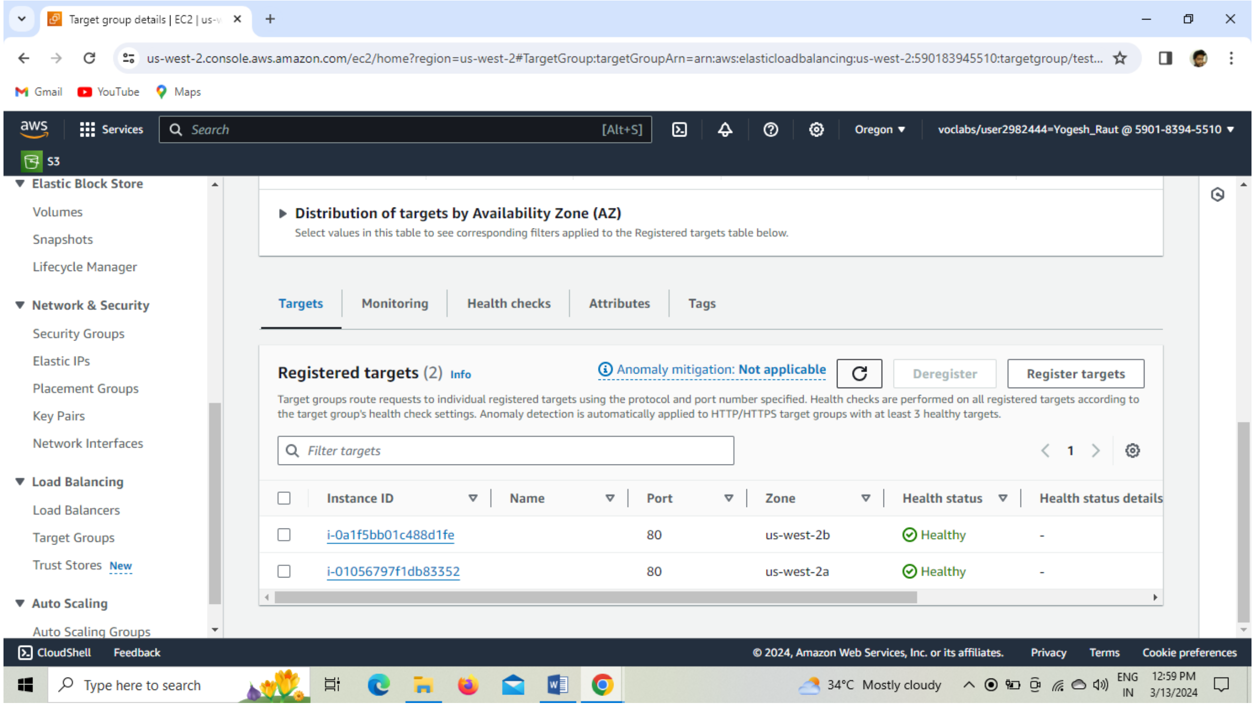
Find Instance by attribute or tag (case-sensitive)

Any state

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availabilit
	i-01056797f1db83352	Running	t3.micro	Initializing	View alarms +	us-west-2a
	i-0a1f5bb01c488d1fe	Running	t3.micro	Initializing	View alarms +	us-west-2b
Web Server 1	i-07a56d7dac76d719c	Running	t3.micro	2/2 checks passed	View alarms +	us-west-2b

Select an instance

2. In left navigation pane, in **Load Balancing** section, check **if new targets are registered in your target group** and that they are **healthy.**



Distribution of targets by Availability Zone (AZ)

Select values in this table to see corresponding filters applied to the Registered targets table below.

Registered targets (2) Info

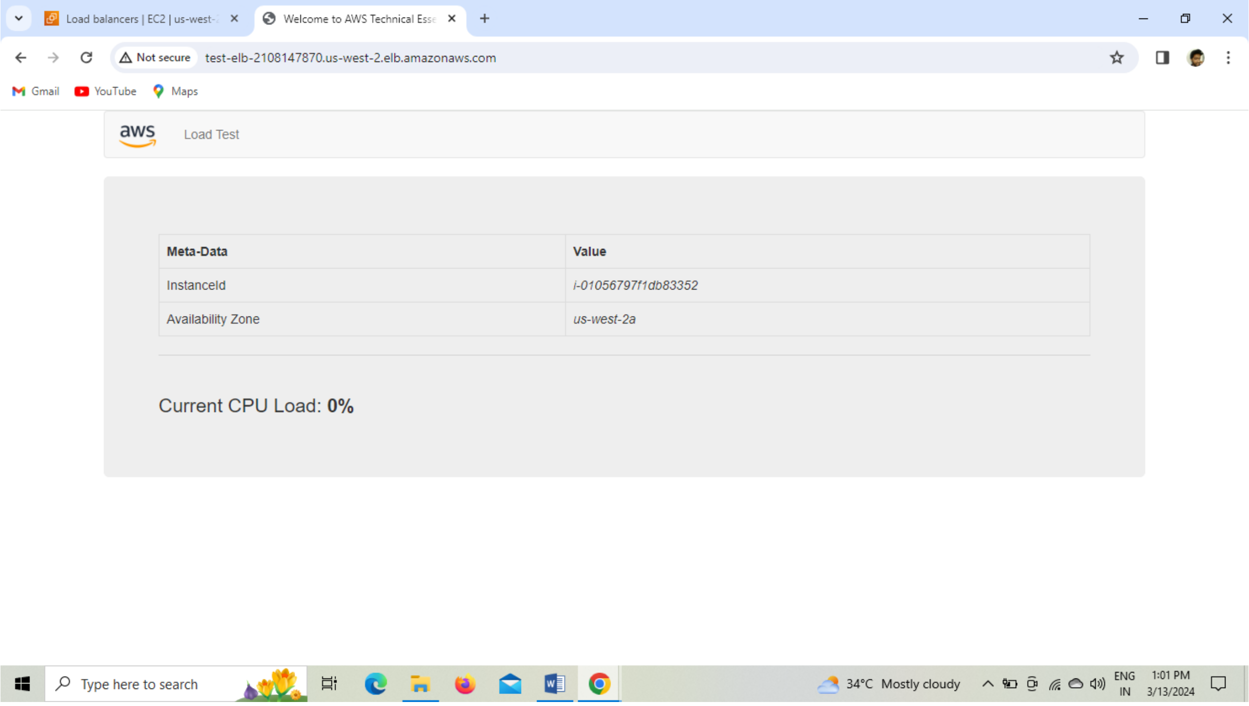
Target groups route requests to individual registered targets using the protocol and port number specified. Health checks are performed on all registered targets according to the target group's health check settings. Anomaly detection is automatically applied to HTTP/HTTPS target groups with at least 3 healthy targets.

Filter targets

Instance ID	Name	Port	Zone	Health status	Health status details
i-0a1f5bb01c488d1fe		80	us-west-2b	Healthy	-
i-01056797f1db83352		80	us-west-2a	Healthy	-

3. Open a new web browser tab, paste the **DNS name** that you copied before, and press Enter.

The Load Test application should appear in your browser, which means that the load balancer received the request, sent it to one of the EC2 instances, and then passed back the result**.**



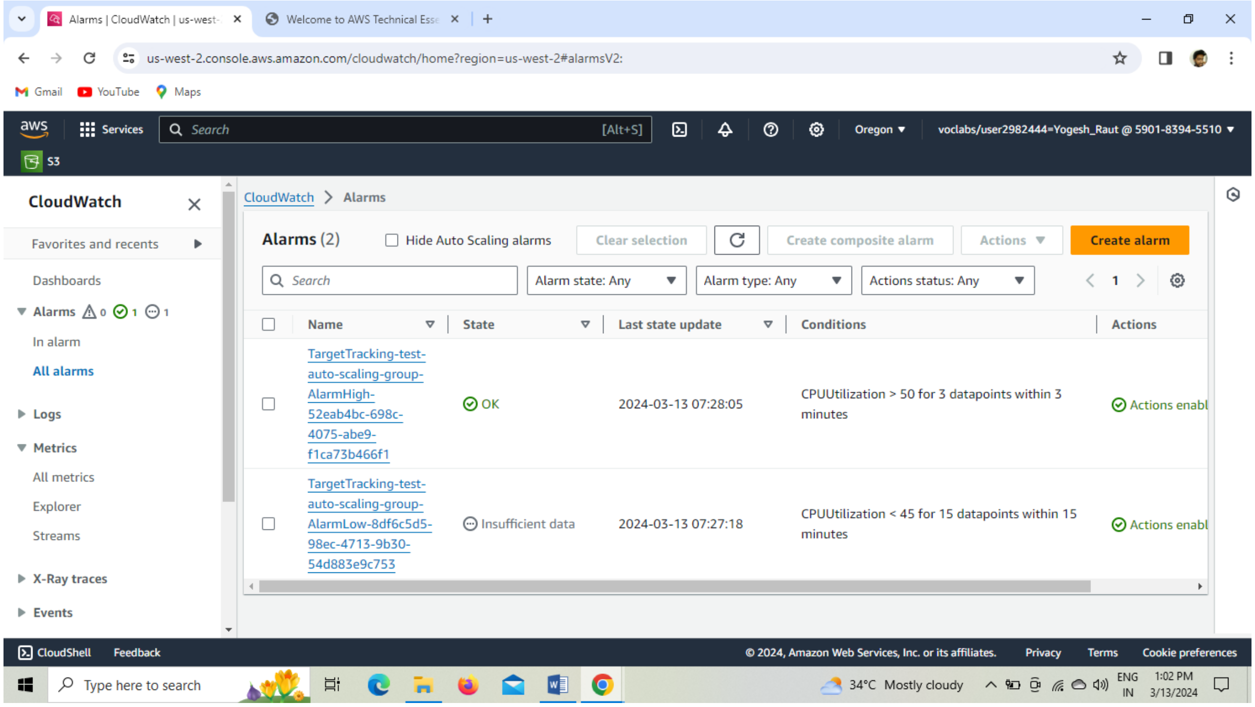
Load balancers | EC2 | us-west-
Welcome to AWS Technical Esse
Not secure
test-elb-2108147870.us-west-2.elb.amazonaws.com
Gmail
YouTube
Maps
aws
Load Test
Type here to search
34°C Mostly cloudy
ENG
IN
1:01 PM
3/13/2024

Meta-Data	Value
InstanceId	*i-01056797f1db83352*
Availability Zone	*us-west-2a*

Current CPU Load: **0%**

F. Testing Auto Scaling

1. Return to the **AWS Management Console**, but keep the Load Test application tab open.
2. In the AWS Management Console, in the search bar, enter and choose **CloudWatch**
3. In left navigation pane, choose **Alarms.** In **All Alarms**, check if **AlarmHigh** state is **OK.**
4. Choose **Load Test** from Load Test application tab to **increase system load.**
5. Go back to **CloudWatch,** and check **if Alarm Status changes.**
6. Once the **Alarm Status** changes to **In alarm,** new **instances will be launched**.



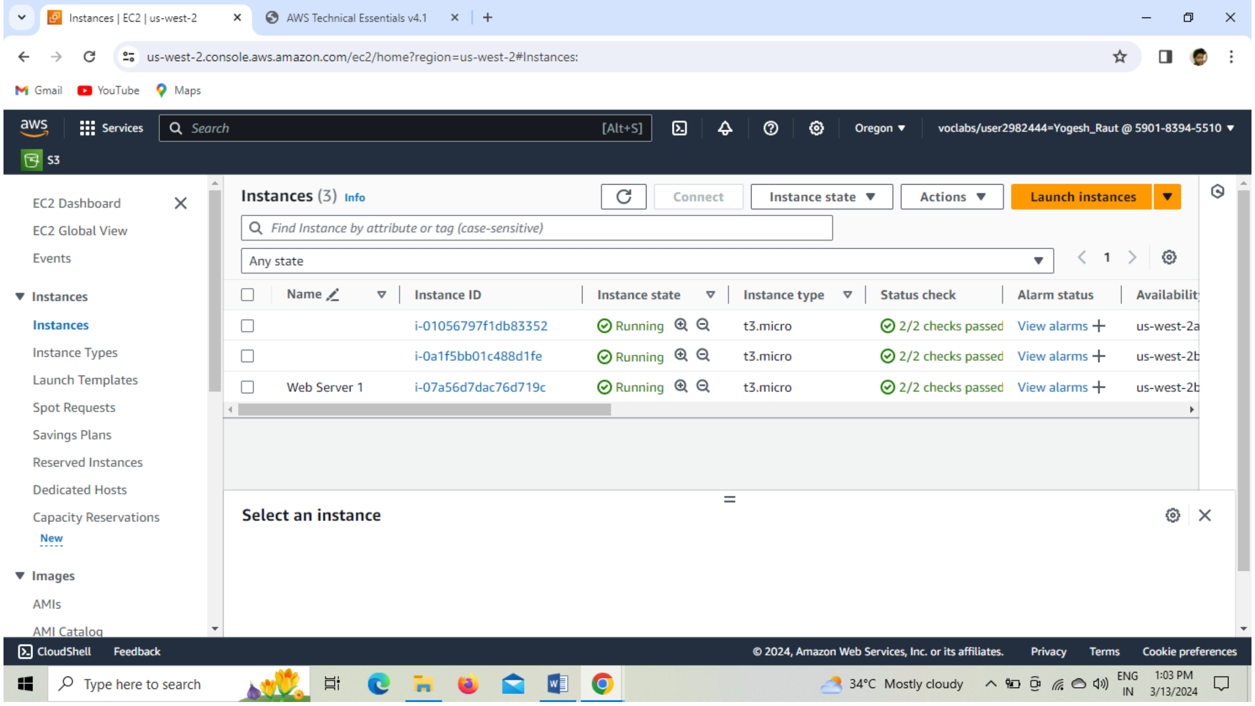
us-west-2.console.aws.amazon.com/cloudwatch/home?region=us-west-2#alarmsV2:

CloudWatch > Alarms

Alarms (2)

□ Hide Auto Scaling alarms

	Name	State	Last state update	Conditions	Actions
□	TargetTracking-test-auto-scaling-group-AlarmHigh-52eab4bc-698c-4075-abe9-f1ca73b466f1	✓ OK	2024-03-13 07:28:05	CPUUtilization > 50 for 3 datapoints within 3 minutes	✓ Actions enabl
□	TargetTracking-test-auto-scaling-group-AlarmLow-8df6c5d5-98ec-4713-9b30-54d883e9c753	Insufficient data	2024-03-13 07:27:18	CPUUtilization < 45 for 15 datapoints within 15 minutes	✓ Actions enabl



us-west-2.console.aws.amazon.com/ec2/home?region=us-west-2#Instances:

Instances (3) Info

Any state

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availabilit
□		i-01056797f1db83352	✓ Running	t3.micro	✓ 2/2 checks passed	View alarms +	us-west-2a
□		i-0a1f5bb01c488d1fe	✓ Running	t3.micro	✓ 2/2 checks passed	View alarms +	us-west-2b
□	Web Server 1	i-07a56d7dac76d719c	✓ Running	t3.micro	✓ 2/2 checks passed	View alarms +	us-west-2b

Select an instance

Congrats!! Your system is now automatically scalable and balancing the load too.