



174- [JAWS] - Lab - Escalado y equilibrio de carga de una arquitectura

Datos Generales:

Nombre: Tomás Alfredo Villaseca Constantinescu

País: Chile

Fecha: 21/10/2023

Contacto: tomas.villaseca.c@gmail.com

Después de completar este laboratorio, usted podrá ser capaz de hacer lo siguiente:

- Crear una AMI a partir de una instancia EC2.
- Crear un equilibrador de carga.
- Crear una plantilla de lanzamiento y un grupo de Auto Scaling.
- Configurar un grupo de Auto Scaling para escalar nuevas instancias dentro de subredes privadas.
- Utilizar las alarmas de Amazon CloudWatch para monitorizar el rendimiento de la infraestructura.

Resumen Laboratorio:

En este laboratorio, utilizará Elastic Load Balancing (ELB) y Amazon EC2 Auto Scaling para equilibrar la carga y escalar automáticamente su infraestructura.

ELB distribuye automáticamente el tráfico entrante de las aplicaciones entre varias instancias de Amazon EC2.

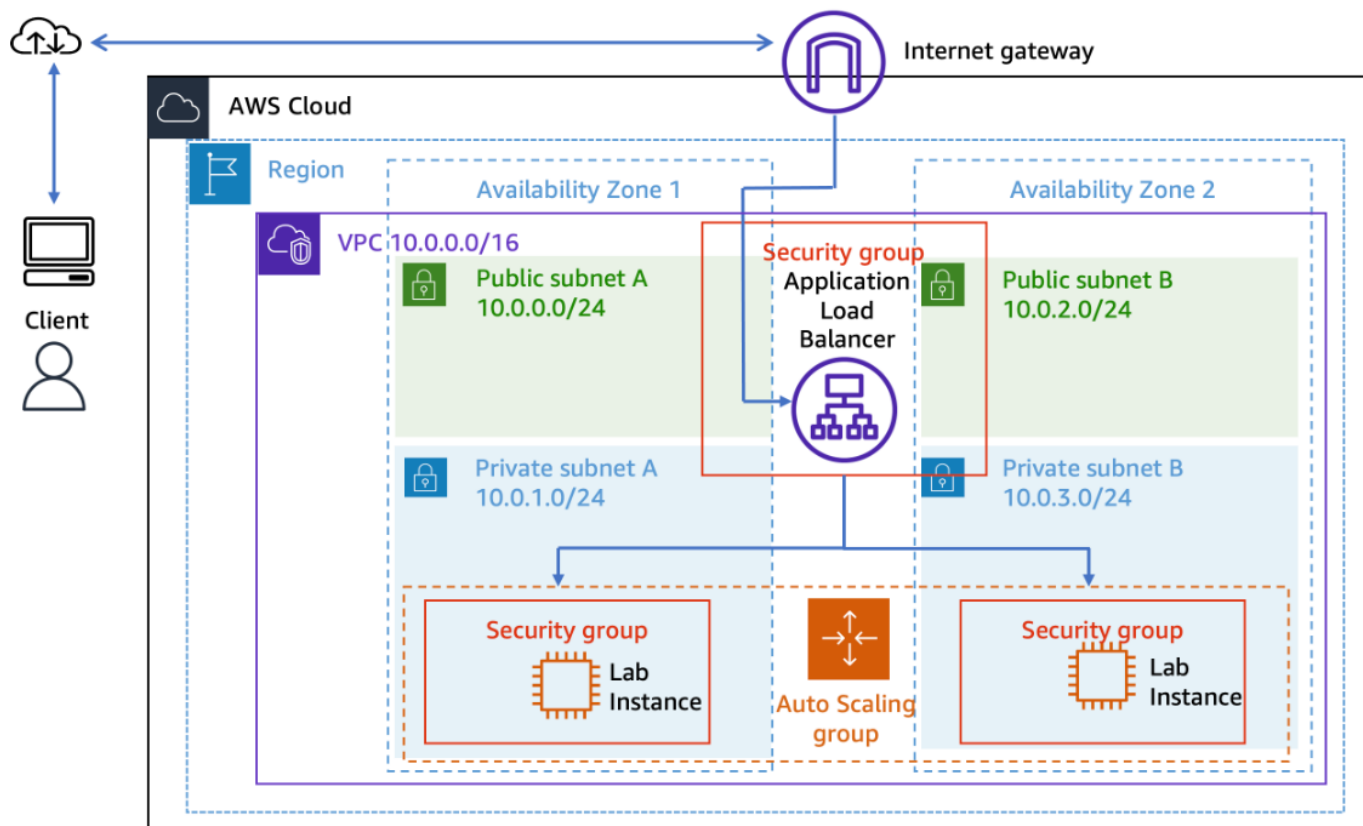
ELB proporciona la cantidad de capacidad de equilibrio de carga necesaria para enrutar el tráfico de las aplicaciones con el fin de conseguir tolerancia a fallos en sus aplicaciones.

Auto Scaling le ayuda a mantener la disponibilidad de las aplicaciones y le ofrece la posibilidad de reducir o aumentar automáticamente la capacidad de Amazon EC2 en función de las condiciones que defina.

Puede utilizar el escalado automático para asegurarse de que está ejecutando el número deseado de instancias EC2.

El escalado automático también puede aumentar automáticamente el número de instancias EC2 durante picos de demanda para mantener el rendimiento y puede reducir la capacidad durante periodos de inactividad para reducir costes.

Diagrama de la arquitectura final al completar el laboratorio:



Tarea 1: Crear una AMI

para Auto Scaling

En esta tarea, se crea una AMI a partir del Web Server 1 existente. Esta acción guarda el contenido del boot disk para que se puedan lanzar nuevas instancias con idéntico contenido.

Paso 1: AWS Management Console → Search → EC2 → Instances → Web Server 1

Instances (1/1) Info				
<input type="text" value="Find instance by attribute or tag (case-sensitive)"/>				
<input checked="" type="checkbox"/>	Name	Instance ID	Instance state	Instance type
<input checked="" type="checkbox"/>	Web Server 1	i-073ab12a6e2cf1462	Running	t3.micro

Paso 2: Web Server 1 → Actions → Image and templates → Create image

Connect

Instance state ▼

Actions ▲

Launch instances ▼

Public IPv4 DNS

Public I

54.186.

Connect

View details

Manage instance state

Instance settings ▶

Networking ▶

Security ▶

Image and templates ▶

Monitor and troubleshoot ▶

Create image

Create template from instance

Launch more like this

Paso 3: Create Image

- Image name = Web Server AMI
- Image description = Lab AMI for Web Server

Instance ID

i-073ab12a6e2cf1462 (Web Server 1)

Image name

Web Server AMI

Maximum 127 characters. Can't be modified after creation.

Image description - *optional*

Lab AMI for Web Server

Maximum 255 characters

<input checked="" type="checkbox"/>	Name	AMI ID	AMI name
<input checked="" type="checkbox"/>		ami-06beb6ca45dd0c50a	Web Server AMI

Tarea 2: Crear un balanceador de carga

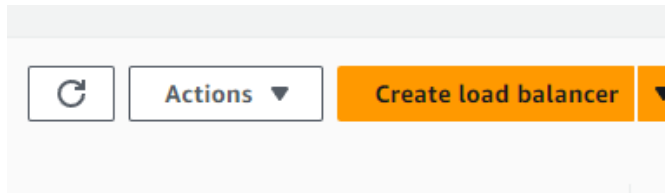
En esta tarea, se crea un equilibrador de carga que puede equilibrar el tráfico entre varias instancias EC2 y zonas de disponibilidad.

Paso 1: EC2 → Panel de navegación → Load Balancing → Load Balancers → Create Load Balancer

▼ Load Balancing

Load Balancers

Target Groups



Paso 2: Create Load Balancer → Load Balancer types

- Application Load Balancer → Create

Load balancer types

Application Load Balancer [Info](#)

Network Load Balancer [Info](#)

Gateway Load Balancer [Info](#)

Paso 3: Create Application Load Balancer → Basic Configuration

- Load Balancer Name = LabELB

Basic configuration

Load balancer name

Name must be unique within your AWS account and can't be changed after the load balancer is created.

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

An internet-facing load balancer routes requests from clients over the internet to targets. Requires a public subnet. [Learn more](#) [↗](#)

☐ Internal

An internal load balancer routes requests from clients to targets using private IP addresses.

IP address type [Info](#)

Select the type of IP addresses that your subnets use.

☒ IPv4

Recommended for internal load balancers.

☐ Dualstack

Includes IPv4 and IPv6 addresses.

Paso 4: Create Application Load Balancer → Network mapping

- VPC → Lab VPC
- Mappings → Seleccionar las dos AZs listadas.
- AZ 1 → Public Subnet 1
- AZ 2 → Public Subnet 2

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-0cd6bb38c92cca536
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-01780faffacf30cf8Public Subnet 1 ▼

IPv4 address

Assigned by AWS

☒ **us-west-2b (usw2-az1)**

Subnet

subnet-05cfdb81bb2f28dbcPublic Subnet 2 ▼

IPv4 address

Assigned by AWS


Estas opciones configuran el equilibrador de carga para que funcione en varias zonas de disponibilidad.

Paso 5: Create Application Load Balancer → Security Groups

- Desplegar lista → Web Security Group (permite conexión HTTP)

Security groups

Select up to 5 security groups

Web Security Group 
sg-07a905745150be2a4 VPC: vpc-0cd6bb38c92cca536

Paso 6: Create Application Load Balancer → Listeners and routing

- Create target group → Se abre una nueva pestaña en el navegador.

▼ Listener HTTP:80

Protocol	Port	Default action	Info
HTTP ▼	: 80 1-65535	Forward to	Select a target group

Create target group [↗](#)

Paso 7: Target groups → Basic configuration

- Choose target type → Instances

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.

- Target group name = lab-target-group

Target group name

lab-target-group

A maximum of 32 alphanumeric characters including hyphens are allowed, but the name must not begin or end with a hyphen.

Paso 8: Target groups → Register targets → Create target group

- Volver a la pestaña de Load Balancers
- Refrescar la sección Listeners and routing.
- Forward to → lab-target-group

[Previous](#) [Create target group](#)

<input checked="" type="checkbox"/>	Name ▼	ARN ▼	Port ▼	Protocol
<input checked="" type="checkbox"/>	lab-target-group	arn:aws:elasticloadbalanci...	80	HTTP

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

Create target group [↗](#)

Paso 9: Create Application Load Balancer → Create Load Balancer

- View Load Balancer
- Copiar el DNS name de LabELB → LabELB-996166595.us-west-2.elb.amazonaws.com


Cancel



Create load balancer

✔

Successfully created load balancer: **LabELB**
Note: It might take a few minutes for your load balancer

DNS name [Info](#)

 LabELB-996166595.us-west-2.elb.amazonaws.com (A Record)

<input checked="" type="checkbox"/>	Name	DNS name	State	VPC ID	Availability Zones	Type
<input checked="" type="checkbox"/>	LabELB	 LabELB-996166595.us-we...	 Provisioning	vpc-0cd6bb38c92cca536	2 Availability Zones	application

Tarea 3: Crear una plantilla de lanzamiento

En esta tarea, se creará una plantilla de lanzamiento (launch template) para el grupo de Auto Scaling. Una plantilla de lanzamiento es una plantilla que un grupo de Auto Scaling utiliza para lanzar instancias EC2.

Al crear una plantilla de lanzamiento, se especifica información para las instancias, como la AMI, el tipo de instancia, el par de claves, el grupo de seguridad y los discos.

Paso 1: EC2 → Instances → Launch Templates → Create Launch Template

▼ Instances

Instances

Instance Types

Launch Templates

Spot Requests

New launch template

Create launch template

Paso 2: Create Launch Template → Launch template name and description

- Launch template name = lab-app-launch-template
- Template version description = a web server for the load test app
- Auto scaling guidance → Seleccionar casilla “Provide guidance to help me set up a template that I can use with EC2 Auto Scaling”.

Launch template name and description

Launch template name - *required*

lab-app-launch-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

Paso 3: Create Launch Template → Application and OS Images

- My AMIs → Web Server AMI


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)

Web Server AMI

ami-06beb6ca45dd0c50a

2023-10-21T20:59:32.000Z Virtualization: hvm ENA enabled: true Root device type: ebs

Paso 4: Create Launch Template → Instance type

- Desplegar lista → t3.micro

Instance type

t3.micro

Family: t3 2 vCPU 1 GiB Memory Current generation: true

On-Demand SUSE base pricing: 0.0104 USD per Hour

On-Demand Windows base pricing: 0.0196 USD per Hour

On-Demand RHEL base pricing: 0.0704 USD per Hour

On-Demand Linux base pricing: 0.0104 USD per Hour

Paso 5: Create Launch Template → Key pair (login)


- Desplegar lista → Don't include in launch template.

▼ Key pair (login) Info

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance.

Key pair name

Don't include in launch template

 Create new key pair

Paso 6: Create Launch Template → Network settings

- Security groups → Web Security Group

▼ Network settings

Info

Subnet

Info

Don't include in launch template

▼

↻

Create new subnet

↗

When you specify a subnet, a network interface is automatically added to your template.

Firewall (security groups)

Info

A security group is a set of firewall rules that control the traffic for your instance. Add rules to allow specific traffic to reach your instance.

☒ Select existing security group

☐ Create security group

Security groups

Info

Select security groups

▼

↻

Compare security group rules

Web Security Group

sg-07a905745150be2a4

✕

VPC: vpc-0cd6bb38c92cca536

► Advanced network configuration

Paso 7: Create Launch Template → Create

- View Launch Templates

Cancel

Create launch template

✓ Success

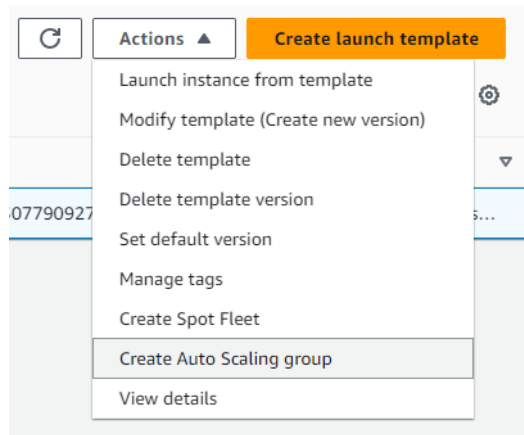
Successfully created lab-app-launch-template(lt-0904792e6d76cb7cf).

	Launch Template ID	Launch Template Name
<input checked="" type="radio"/>	lt-0904792e6d76cb7cf	lab-app-launch-template

Tarea 4: Crear un Auto Scaling Group

En esta tarea, se utiliza la plantilla de lanzamiento para crear un Auto Scaling group.

Paso 1: Launch Templates → lab-app-launch-template → Actions → Create Auto Scaling group



Paso 2: Create Auto Scaling group → Name

- Auto Scaling group name = Lab Auto Scaling Group

Name

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

Lab Auto Scaling Group

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

Paso 3: Create Auto Scaling group → Instance launch options → Network

- VPC → Lab VPC
- Availability Zones and subnets → Private Subnet 1 / Private Subnet 2.

VPC

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

vpc-0cd6bb38c92cca536 (Lab VPC) ▼

10.0.0.0/16

↻

[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 ▼

↻

us-west-2a | subnet-0d1ad3f1598f5fd21 (Private Subnet 1) ✕

10.0.1.0/24

us-west-2b | subnet-00534fdcc643940cc (Private Subnet 2) ✕

10.0.3.0/24

Paso 4: Create Auto Scaling group → Configure advanced options

- Load Balancing → Attach to an existing load balancer

Load balancing [Info](#)

Use the options below to attach your Auto Scaling group to an existing load balancer, or to a new load balancer that you define.

- ☐ No load balancer
Traffic to your Auto Scaling group will not be fronted by a load balancer.
- ☒ Attach to an existing load balancer
Choose from your existing load balancers.
- ☐ Attach to a new load balancer
Quickly create a basic load balancer to attach to your Auto Scaling group.

- Choose from your load balancer target groups
- Existing load balancer target groups → lab-target-group | HTTP

- ☒ Choose from your load balancer target groups
This option allows you to attach Application, Network, or Gateway Load Balancers.
- ☐ Choose from Classic Load Balancers

Existing load balancer target groups

Only instance target groups that belong to the same VPC as your Auto Scaling group are available for selection.

Select target groups ▼



lab-target-group | HTTP X
Application Load Balancer: LabELB

- Health checks → Health check type → Seleccionar casilla “ELB”.

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 **Recommended**

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.

Paso 5: Create Auto Scaling group → Configure group size and scaling policies

- Group size → Desired capacity = 2, Minimum capacity = 2, Maximum capacity = 4.

Group size - *optional* [Info](#)

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

Minimum capacity

Maximum capacity

- Scaling policies → Seleccionar “Target tracking scaling policy “
- Scaling policies → Metric type → Average CPU utilization
- Scaling policies → Target Value = 50

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

Scaling policy name

Metric type [Info](#)

Monitored metric that determines if resource utilization is too low or high. If using EC2 metrics, consider enabling detailed monitoring for better scaling performance.



Target value

Este cambio indica a Auto Scaling Group que mantenga una utilización media de la CPU en todas las instancias del 50 por ciento. Auto Scaling agrega o elimina capacidad automáticamente según sea necesario para mantener la métrica en o cerca del valor objetivo especificado. Se ajusta a las fluctuaciones en la métrica debido a un patrón de carga fluctuante.

Paso 6: Create Auto Scaling group → Add tags → Add tag

- Key = Name
- Value = Lab Instance

Tags (1)

Key	Value - optional	Tag new instances
<input type="text" value="Name"/>	<input type="text" value="Lab Instance"/>	<input checked="" type="checkbox"/>

49 remaining

Paso 6: Create Auto Scaling group → Create

✔ Lab Auto Scaling Group, 1 Scaling policy created successfully

Estas opciones lanzan instancias EC2 en subredes privadas en ambas zonas de disponibilidad.

Su Auto Scaling Group muestra inicialmente un recuento de instancias de cero, pero se lanzarán nuevas instancias para alcanzar el recuento deseado de dos instancias.

Auto Scaling groups (1/1) Info			
<input type="text" value="Search your Auto Scaling groups"/>			
<input checked="" type="checkbox"/>	Name	Launch template/configuration ↗	Instances
<input checked="" type="checkbox"/>	Lab Auto Scaling Group	lab-app-launch-template Version Defal	2

Tarea 5: Verificar el funcionamiento del balanceador de carga

En esta tarea, se verifica que el equilibrio de carga funciona correctamente.

Paso 1: EC2 → Instances

- Se pueden ver disponibles dos instancias de nombre Lab Instance que fueron lanzadas por Auto Scaling.

Instances (2/3) Info						
<input type="text" value="Find instance by attribute or tag (case-sensitive)"/>						
<input type="checkbox"/>	Name	Instance ID	Instance state	Instance type	Status check	
<input checked="" type="checkbox"/>	Lab Instance	i-0c35a005236a5a387	Running	t3.micro	2/2 checks passed	
<input checked="" type="checkbox"/>	Lab Instance	i-02d7bb946897a2957	Running	t3.micro	2/2 checks passed	
<input type="checkbox"/>	Web Server 1	i-073ab12a6e2cf1462	Running	t3.micro	2/2 checks passed	

Paso 2: EC2 → Load Balancing → Target Groups → lab-target-group

- Registered targets → Aparecen dos Lab Instances
- Health Status → healthy
- healthy = Indica que una instancia ha superado la comprobación de salud del equilibrador de carga (Esta comprobación significa que el equilibrador de carga enviará tráfico a la instancia).


<input checked="" type="checkbox"/>	Name	ARN	Port	Protocol	Target type	Load balancer
<input checked="" type="checkbox"/>	lab-target-group	arn:aws:elasticloadbalanci...	80	HTTP	Instance	LabELB
<div>Details Targets Monitoring Health checks Attributes Tags</div>						
Registered targets (2)						
<input type="text" value="Filter targets"/>						
<input type="checkbox"/>	Instance ID	Name	Port	Zone	Health status	
<input type="checkbox"/>	i-02d7bb946897a2957	Lab Instance	80	us-west-2a	healthy	
<input type="checkbox"/>	i-0c35a005236a5a387	Lab Instance	80	us-west-2b	healthy	

Paso 3: Pegar el DNS name copiado anteriormente en una nueva pestaña de navegador web.

- La Load Test Application debería aparecer en su navegador, lo que significa que el equilibrador de carga recibió la solicitud, la envió a una de las instancias de EC2 y, a continuación, devolvió el resultado.

labelb-996166595.us-west-2.elb.amazonaws.com

RRSS Anime Política Pagos Trabajo Cyber Security

 Load Test

Meta-Data	Value
InstanceId	i-02d7bb946897a2957
Availability Zone	us-west-2a

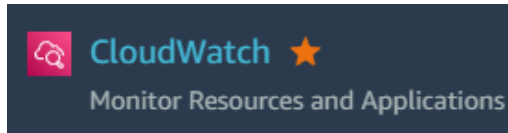
Current CPU Load: 0%

Tarea 6: Probar el Auto Scaling Group

Ha creado un Auto Scaling Group con un mínimo de dos instancias y un máximo de cuatro instancias.

Actualmente, dos instancias están funcionando porque el tamaño mínimo es dos y el grupo no está bajo ninguna carga. Ahora aumentará la carga para que el autoescalado añada instancias adicionales.

Paso 1: AWS Management Console → Search → CloudWatch



Paso 2: CloudWatch → Panel de navegación → Alarms → All Alarms

- Hay 2 alarmas desplegadas que fueron creadas automáticamente por el Auto Scaling Group.
- Estas alarmas mantienen automáticamente la carga media de la CPU cerca del 50 por ciento, mientras se mantienen dentro de la limitación de tener de 2 a 4 instancias.

▼ Alarms ⚠️ 0 ✅ 1 ⋮ 1

In alarm

All alarms

Alarms (2)		<input type="checkbox"/> Hide Auto Scaling alarms	Clear selection	↻	Create composite alarm
<input type="text" value="Search"/>		Any state ▼	Any type ▼	Any actions ... ▼	
<input type="checkbox"/>	Name ▼	State ▼	Last state update ▼	Conditions	Actions
<input type="checkbox"/>	TargetTracking-Lab Auto Scaling Group-AlarmHigh-a23d063a-12fd-4224-a87f-ee44d6d01299	✅ OK	2023-10-21 21:29:34	CPUUtilization > 50 for 3 datapoints within 3 minutes	✅ Actions enabled
<input type="checkbox"/>	TargetTracking-Lab Auto Scaling Group-AlarmLow-f325c53f-f4a6-4059-8d3f-ba5004fc8f45	⋮ Insufficient data	2023-10-21 21:28:06	CPUUtilization < 45 for 15 datapoints within 15 minutes	✅ Actions enabled

Paso 3: CloudWatch → Panel de navegación → Alarms → All Alarms → AlarmHigh

- State → OK
- OK = Indica que la alarma no ha sido iniciada.

<input type="checkbox"/>	Name ▼	State
<input checked="" type="checkbox"/>	TargetTracking-Lab Auto Scaling Group-AlarmHigh-a23d063a-12fd-4224-a87f-ee44d6d01299	✅ OK

Paso 4: Ir al navegador que tiene la Load Test Application y seleccionar Load Test.

- Esto hace que la aplicación genere cargas elevadas. La página del navegador se actualiza automáticamente para que todas las instancias del grupo Auto Scaling generen cargas.

aws

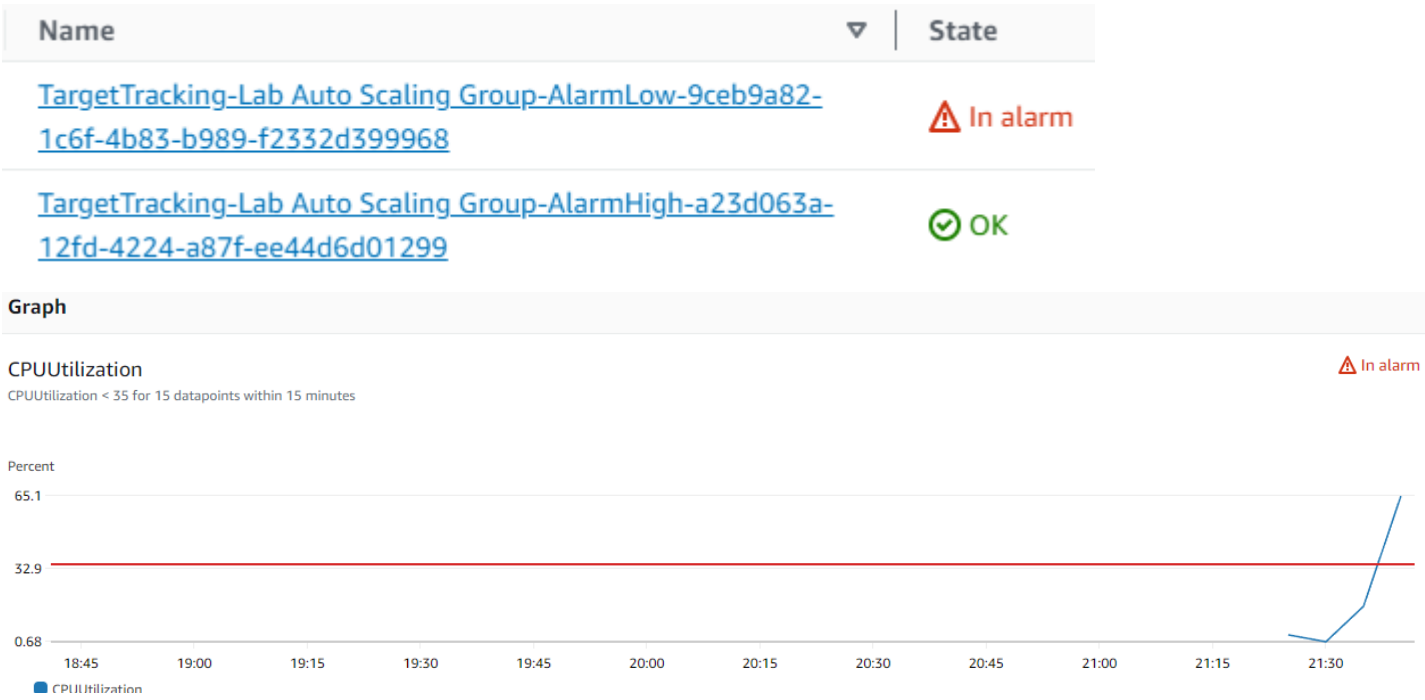
Load Test

Generating CPU Load! (auto refresh in 5 seconds)

Current CPU Load: 66%









Paso 5: CloudWatch → Panel de navegación → Alarms → All Alarms

- AlarmLow → OK
- AlarmHigh → In Alarm
- Se puede ver el gráfico AlarmHigh indicando un porcentaje creciente de CPU. Una vez que cruza la línea del 50% durante más de 3 minutos, se inicia el escalado automático para añadir instancias adicionales.



Paso 6: EC2 → Instances

- Se pueden ver más de dos instancias de nombre Lab Instance en ejecución.
- Auto Scaling Group lanzó nuevas instancias en respuesta a la alarma.

	Name	Instance ID	Instance state	Instance type	Status check
<input checked="" type="checkbox"/>	Lab Instance	i-015ebcdb763c03567	 Running	t3.micro	 Initializing
<input checked="" type="checkbox"/>	Lab Instance	i-0c35a005236a5a387	 Running	t3.micro	 2/2 checks passed
<input checked="" type="checkbox"/>	Lab Instance	i-02d7bb946897a2957	 Running	t3.micro	 2/2 checks passed
<input type="checkbox"/>	Web Server 1	i-073ab12a6e2cf1462	 Running	t3.micro	 2/2 checks passed

Tarea 7: Terminar la instancia

Web Server 1

En esta tarea, se finaliza la instancia Web Server1.

Esta instancia se utilizó para crear la AMI que utilizó su Auto Scaling Group, pero ya no es necesaria.

Paso 1: EC2 → Instances → Web Server 1

	Name	Instance ID	Instance state	Instance type	Status check
<input checked="" type="checkbox"/>	Web Server 1	i-073ab12a6e2cf1462	Running	t3.micro	2/2 checks passed

Paso 2: Web Server 1 → Instance State → Terminate Instance

Availability Zone

Pub

west-2b

west-2b

west-2b

Stop instance

Start instance

Reboot instance

Hibernate instance

Terminate instance

< 1 >

Public IP

54.186.1

-

Successfully terminated i-073ab12a6e2cf1462

	Name	Instance ID	Instance state
<input checked="" type="checkbox"/>	Web Server 1	i-073ab12a6e2cf1462	Terminated

Laboratorio Completado

