



AWS Academy Cloud Foundations (ES)
Module 10 Student Guide
Versión 2.0.1
100-ACCLFO-20-ES-SG

© 2020 Amazon Web Services, Inc. o sus empresas afiliadas.
Todos los derechos reservados.

Este contenido no puede reproducirse ni redistribuirse, total ni parcialmente, sin el permiso previo por escrito de Amazon Web Services, Inc. Queda prohibida la copia, el préstamo o la venta de carácter comercial.

Envíenos sus correcciones o comentarios relacionados con el curso a:
aws-course-feedback@amazon.com.

Si tiene cualquier otra duda, contacte con nosotros en:
<https://aws.amazon.com/contact-us/aws-training/>.

Todas las marcas comerciales pertenecen a sus propietarios.

Contenido


Módulo 10: Monitoreo y escalado automático

4




Bienvenido al Módulo 10: Monitoreo y escalado automático

Información general sobre el módulo



Temas <ul style="list-style-type: none">• Elastic Load Balancing• Amazon CloudWatch• Amazon EC2 Auto Scaling	Actividades <ul style="list-style-type: none">• Actividad de Elastic Load Balancing• Actividad de Amazon CloudWatch Laboratorio <ul style="list-style-type: none">• Ajuste de la escala y balanceo de la carga de su arquitectura
---	---

 **Revisión de conocimientos**

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.2

En este módulo, se abordarán los siguientes temas:

- Elastic Load Balancing
- Amazon CloudWatch
- Amazon EC2 Auto Scaling

El módulo también incluye dos actividades. Una actividad lo desafiará a indicar los casos de uso de Elastic Load Balancing. La otra actividad lo desafiará a identificar ejemplos de Amazon CloudWatch.

Además, el módulo incluye un laboratorio práctico en el que utilizará Amazon EC2 Auto Scaling, Elastic Load Balancing y Amazon CloudWatch juntos para crear una arquitectura escalable de forma dinámica.

Por último, se le solicitará que complete una revisión de conocimientos que probará su comprensión de los conceptos clave que se tratan en este módulo.

Objetivos del módulo



Después de completar este módulo, debería ser capaz de lo siguiente:

- Indicar cómo distribuir el tráfico entre las instancias de Amazon Elastic Compute Cloud (Amazon EC2) usando Elastic Load Balancing
- Identificar cómo Amazon CloudWatch permite monitorear recursos y aplicaciones de AWS en tiempo real
- Explicar cómo Amazon EC2 Auto Scaling lanza y publica servidores en respuesta a los cambios en las cargas de trabajo
- Realizar tareas de escalado y balanceo de carga para mejorar una arquitectura

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

5

Después de completar este módulo, debería ser capaz de lo siguiente:

- Indicar cómo distribuir el tráfico entre las instancias de Amazon Elastic Compute Cloud (Amazon EC2) usando Elastic Load Balancing
- Identificar cómo Amazon CloudWatch permite monitorear recursos y aplicaciones de AWS en tiempo real
- Explicar cómo Amazon EC2 Auto Scaling lanza y publica servidores en respuesta a los cambios en las cargas de trabajo
- Realizar tareas de escalado y balanceo de carga para mejorar una arquitectura

Módulo 10: Monitoreo y escalado automático

Sección 1: Elastic Load Balancing

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

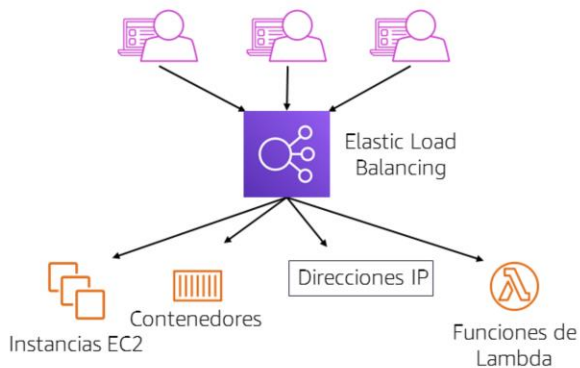


Sección 1: Elastic Load Balancing

Elastic Load Balancing



- Distribuye el tráfico entrante de las aplicaciones o de la red entre varios destinos en una única zona de disponibilidad o en varias zonas de disponibilidad.
- Escala el balanceador de carga a medida que el tráfico dirigido a la aplicación cambia con el tiempo.



© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

3

Los sitios web modernos de mucho tráfico atienden cientos de miles, si no millones, de solicitudes simultáneas de los usuarios o los clientes y, luego, devuelven el texto, las imágenes, los videos o los datos de las aplicaciones correctos de una manera rápida y confiable. Por lo general, se necesitan servidores adicionales para satisfacer estos altos volúmenes.

Elastic Load Balancing es un servicio de AWS que distribuye el tráfico entrante de las aplicaciones o la red entre varios destinos, como las instancias de Amazon Elastic Compute Cloud (Amazon EC2), los contenedores, las direcciones de protocolo de Internet (IP) y las funciones de Lambda, en una única zona de disponibilidad o en varias. Elastic Load Balancing escala el balanceador de carga a medida que el tráfico dirigido a la aplicación cambia con el tiempo. Puede escalar automáticamente a la mayoría de las cargas de trabajo.

Tipos de balanceadores de carga



Balanceador de carga de aplicaciones	Balanceador de carga de red	Balanceador de carga clásico (generación anterior)
<ul style="list-style-type: none"> Balanceo de carga del tráfico HTTP y HTTPS Dirige el tráfico a los destinos en función del contenido de la solicitud. Proporciona direccionamiento de solicitudes avanzadas a su entrega en arquitecturas modernas de aplicaciones, como los microservicios y los contenedores. Funciona en la capa de aplicación (capa 7 del modelo OSI). 	<ul style="list-style-type: none"> Balanceo de carga del tráfico TCP, UDP y TLS donde se requiere un rendimiento extremo Dirige el tráfico a los destinos en función de los datos del protocolo IP. Puede gestionar millones de solicitudes por segundo y, a la vez, mantener latencias muy bajas. Está optimizado para gestionar patrones de tráfico repentinos y volátiles. Funciona en la capa de transporte (capa 4 del modelo OSI). 	<ul style="list-style-type: none"> Balanceo de carga del tráfico HTTP, HTTPS, TCP y SSL Balanceo de carga entre varias instancias EC2 Funciona en las capas de aplicación y de transporte.

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

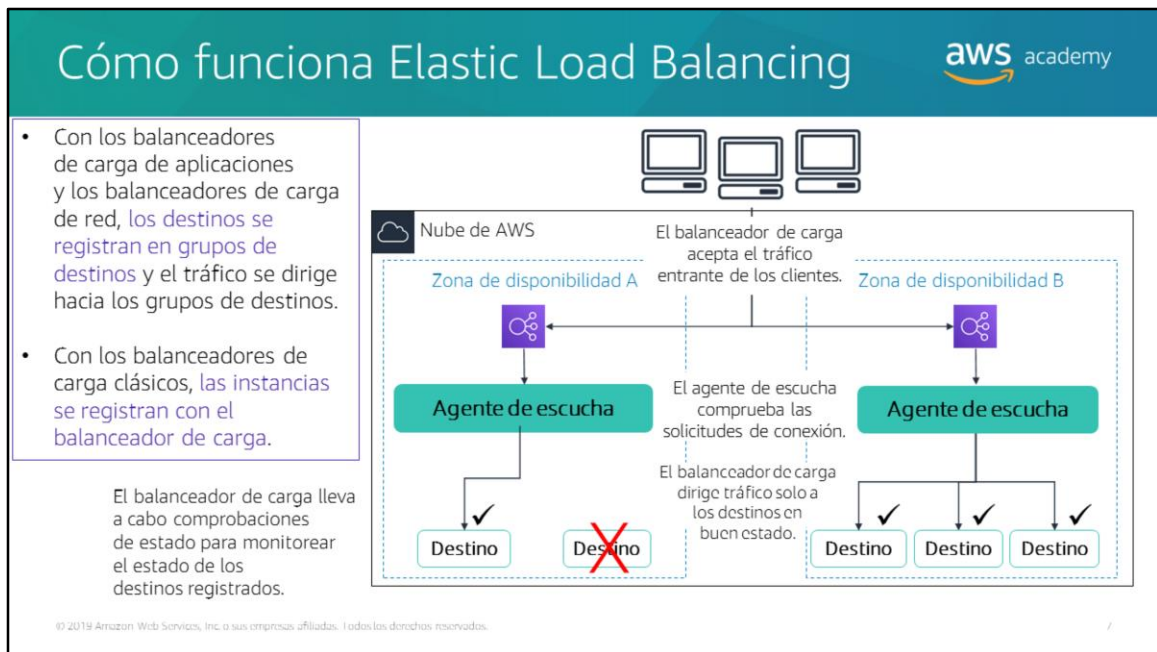
5

Elastic Load Balancing está disponible en tres tipos:

- El *balanceador de carga de aplicaciones* opera en el nivel de la aplicación (capa 7 del modelo de interconexión de sistemas abiertos u OSI). Dirige el tráfico a los destinos (instancias de Amazon Elastic Compute Cloud [Amazon EC2], contenedores, direcciones de protocolo de Internet [IP] y funciones de Lambda) en función del contenido de la solicitud. Es ideal para el balanceo de carga avanzado del tráfico del protocolo de transferencia de hipertexto (HTTP) y del HTTP seguro (HTTPS). El balanceador de carga de aplicaciones brinda direccionamiento de solicitudes avanzadas a su entrega en arquitecturas modernas de las aplicaciones, como los microservicios y las aplicaciones basadas en contenedores. El balanceador de carga de aplicaciones simplifica y mejora la seguridad de sus aplicaciones al garantizar que se utilicen en todo momento los protocolos y cifrados de la capa de conexión segura (SSL) o de la seguridad de la capa de transporte (TLS) más recientes.
- El *balanceador de carga de red* opera en el nivel de transporte de la red (capa 4 del modelo OSI) y dirige las conexiones a los destinos (instancias EC2, microservicios y contenedores) en función de los datos del protocolo IP. Funciona bien para balancear la carga del tráfico del protocolo de control de transmisión (TCP) y del protocolo de datagramas de usuario (UDP). El balanceador de carga de red es capaz de gestionar millones de solicitudes por segundo mientras mantiene latencias ultrabajas. El balanceador de carga de red está optimizado para gestionar patrones de tráfico de red repentinos y volátiles.

- El *balanceador de carga clásico* proporciona balanceo de carga básico en varias instancias EC2 y opera tanto en el nivel de aplicación como en el de transporte de red. Un balanceador de carga clásico admite el balanceo de carga de las aplicaciones que utilizan HTTP, HTTPS, TCP y SSL. El balanceador de carga clásico es una implementación más antigua. Cuando sea posible, AWS recomienda utilizar un balanceador de carga de aplicaciones o un balanceador de carga de red dedicados.

Para obtener más información acerca de las diferencias entre los tres tipos de balanceadores de carga, consulte *Comparaciones de productos* en la [página de características](#) de Elastic Load Balancing.



Un balanceador de carga acepta el tráfico entrante de los clientes y dirige las solicitudes a sus destinos registrados (como, por ejemplo, las instancias EC2) en una o más zonas de disponibilidad.


Puede configurar el balanceador de carga para que acepte el tráfico entrante especificando uno o más *agentes de escucha*. Un agente de escucha es un proceso que verifica las solicitudes de conexión. Se configura con un protocolo y un número de puerto para las conexiones de los clientes al balanceador de carga. De modo similar, se configura con un protocolo y un número de puerto para las conexiones del balanceador de carga a los destinos.


También puede configurar el balanceador de carga para que realice *comprobaciones de estado*, que se utilizan para monitorear el estado de los destinos registrados de manera que el balanceador de carga solo envíe solicitudes a las instancias que estén en buen estado. Cuando el balanceador de carga detecta que un destino no está en buen estado, deja de enviar tráfico a ese destino. Luego, cuando detecta que está nuevamente en buen estado, reanuda el tráfico hacia ese destino.

Hay una diferencia clave en el modo en que se configuran los tipos de balanceador de carga. Con los balanceadores de carga de aplicaciones y los balanceadores de

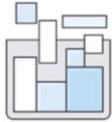
carga de red, los destinos se registran en *grupos de destinos* y el tráfico se dirige hacia los grupos de destinos. Con los balanceadores de carga clásicos, las instancias se registran con el balanceador de carga.

Casos de uso de Elastic Load Balancing







Aplicaciones de alta disponibilidad y tolerantes a errores




Aplicaciones en contenedores




Elasticidad y escalabilidad



Nube virtual privada (VPC)



Entornos híbridos



Invocación de las funciones de Lambda a través de HTTP(S)

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

Hay muchos motivos para utilizar un balanceador de carga, entre los que se incluyen los siguientes:

- **Lograr alta disponibilidad y mejor tolerancia a errores para sus aplicaciones:** Elastic Load Balancing balancea el tráfico en los destinos en buen estado en varias zonas de disponibilidad. Si uno o varios de los destinos de una sola zona de disponibilidad presentan errores, Elastic Load Balancing dirigirá el tráfico a los destinos en buen estado de otras zonas de disponibilidad. Una vez que los destinos vuelven a un estado correcto, el balanceador de carga reanuda de forma automática el tráfico dirigido hacia ellos.
- **Balancear automáticamente la carga de las aplicaciones en contenedores:** gracias a la compatibilidad mejorada de los contenedores con Elastic Load Balancing, ahora puede balancear la carga entre varios puertos de la misma instancia EC2. También puede aprovechar la integración profunda con Amazon Elastic Container Service (Amazon ECS), que ofrece contenedores completamente administrados. Solo debe registrar un servicio con un balanceador de carga y Amazon ECS administrará de forma transparente el registro y la anulación del registro de los contenedores de Docker. El balanceador de carga detecta el puerto de manera automática y se reconfigura dinámicamente.

- *Escalar automáticamente sus aplicaciones:* Elastic Load Balancing trabaja con Amazon CloudWatch y Amazon EC2 Auto Scaling para ayudarlo a escalar las aplicaciones según las demandas de sus clientes. Las alarmas de Amazon CloudWatch pueden activar el escalado automático para su flota de instancias EC2 cuando la latencia de cualquiera de sus instancias EC2 supere un límite preconfigurado. Luego, Amazon EC2 Auto Scaling aprovisiona nuevas instancias y, así, sus aplicaciones estarán listas para atender la siguiente solicitud del cliente. El balanceador de carga registrará la instancia EC2 y dirigirá el tráfico hacia ella según sea necesario.
- *Utilizar Elastic Load Balancing en su nube virtual privada (VPC):* puede utilizar Elastic Load Balancing para crear un punto de entrada público en su VPC o para dirigir el tráfico de solicitudes entre las capas de su aplicación dentro de la VPC. Puede asignar grupos de seguridad al balanceador de carga a fin de controlar qué puertos están abiertos para una lista de fuentes permitidas. Como Elastic Load Balancing trabaja con su VPC, todas las listas existentes de control de acceso a la red (ACL a la red) y las tablas de direccionamiento continúan ofreciendo controles de red adicionales. En el momento de crear un balanceador de carga en su VPC, puede especificar si será público (opción predeterminada) o interno. Si selecciona interno, no necesitará disponer de una gateway de Internet para llegar al balanceador de carga y sus direcciones IP privadas se utilizarán en el registro de su sistema de nombres de dominio (DNS).
- *Habilitar el balanceo de carga híbrido:* Elastic Load Balancing le permite balancear la carga entre los recursos en las instalaciones y en AWS con el mismo balanceador de carga. Por ejemplo, si debe distribuir el tráfico de las aplicaciones entre los recursos de AWS y los que están en las instalaciones, puede registrar todos los recursos en el mismo grupo de destino y asociarlo con el balanceador de carga. De manera alternativa, puede usar el balanceo de cargas ponderado basado en DNS entre los recursos de AWS y en las instalaciones mediante el uso de dos balanceadores de carga, uno para los recursos de AWS y el otro para los recursos en las instalaciones. También puede utilizar el balanceo de carga híbrido para beneficiar aplicaciones independientes en las que una se encuentra en una VPC y la otra en las instalaciones. Coloque los destinos de la VPC en un grupo de destino y los destinos en las instalaciones en otro grupo y, a continuación, utilice el direccionamiento basado en el contenido para dirigir el tráfico a cada grupo de destinos.
- *Invocar las funciones de Lambda a través de HTTP(S):* Elastic Load Balancing admite la invocación de funciones de Lambda para atender solicitudes HTTP(S). Esto permite a los usuarios acceder a aplicaciones sin servidor desde cualquier cliente HTTP, incluidos los navegadores web. Puede registrar las funciones de Lambda como destinos y utilizar el soporte para las reglas de direccionamiento basado en el contenido en los balanceadores de carga de aplicaciones a fin de

dirigir las solicitudes a distintas funciones de Lambda. Puede utilizar el balanceador de carga de aplicaciones como un punto de enlace HTTP común para las aplicaciones que utilizan servidores e informática sin servidor. Puede crear un sitio web completo utilizando funciones de Lambda o combinar instancias EC2, contenedores, servidores en las instalaciones y funciones de Lambda para crear aplicaciones.

Actividad: Elastic Load Balancing



Debe admitir el tráfico a una aplicación en contenedores.

Balanceador de carga de aplicaciones

El tráfico TCP es extremadamente irregular e impredecible.

Balanceador de carga de red

Necesita balanceo de carga simple con múltiples protocolos.

Balanceador de carga clásico

Debe admitir una dirección IP estática o elástica o un destino IP fuera de la VPC.

Balanceador de carga de red

Necesita un balanceador de carga que pueda gestionar millones de solicitudes por segundo mientras mantiene latencias bajas.

Balanceador de carga de red

Debe admitir solicitudes HTTPS.

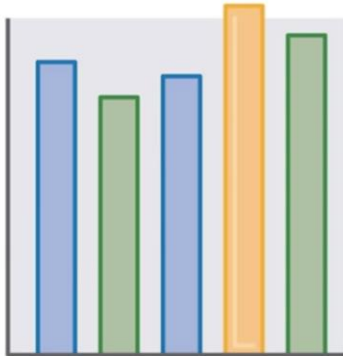
Balanceador de carga de aplicaciones

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

9

Para esta actividad, mencione el balanceador de carga que usaría para la situación presentada.

Monitoreo del balanceador de carga



- **Métricas de Amazon CloudWatch:** estas métricas sirven para verificar que el sistema funcione según lo previsto y, además, este servicio crea una alarma para iniciar una acción si alguna métrica pasa a estar fuera de un rango aceptable.
- **Registros de acceso:** estos registros guardan información detallada acerca de las solicitudes enviadas a su balanceador de carga.
- **Registros de AWS CloudTrail:** estos registros guardan información acerca del quién, qué, cuándo y dónde de las interacciones de la API en los servicios de AWS.

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.


10

Puede utilizar las siguientes características para monitorear los balanceadores de carga, analizar los patrones de tráfico y solucionar los problemas de los balanceadores de carga y los destinos:


- **Métricas de Amazon CloudWatch:** Elastic Load Balancing publica puntos de datos en Amazon CloudWatch para los balanceadores de carga y los destinos. CloudWatch le permite recuperar las estadísticas acerca de esos puntos de datos en conjuntos ordenados de datos de serie temporal, denominados métricas. Puede utilizar las métricas para comprobar si el sistema funciona según lo esperado. Por ejemplo, puede crear una alarma de CloudWatch para monitorear una métrica determinada e iniciar una acción (como, por ejemplo, enviar una notificación a una dirección de email) si la métrica no está comprendida dentro del rango que considera aceptable.
- **Registros de acceso:** puede utilizar registros de acceso para registrar información detallada acerca de las solicitudes que se realizaron al balanceador de carga y almacenarla como archivos de registro en Amazon Simple Storage Service (Amazon S3). Puede utilizar estos registros de acceso para analizar los patrones de tráfico y solucionar los problemas en los destinos o en las aplicaciones de backend.

- *Registros de AWS CloudTrail:* puede utilizar AWS CloudTrail para registrar información detallada acerca de las llamadas que se realizaron a la interfaz de programación de aplicaciones (API) de Elastic Load Balancing y almacenarla como archivos de registro en Amazon S3. Puede utilizar estos registros de CloudTrail para determinar quién realizó la llamada, qué llamadas se efectuaron, cuándo se realizó la llamada, la dirección IP de origen de donde procedió la llamada, etcétera.

Aprendizajes clave de la sección 1



11



- Elastic Load Balancing distribuye el tráfico entrante de las aplicaciones o de la red entre varios destinos en una o más zonas de disponibilidad.
- Elastic Load Balancing admite tres tipos de balanceadores de carga:
 - Balanceador de carga de aplicaciones
 - Balanceador de carga de red
 - Balanceador de carga clásico
- ELB ofrece comprobaciones de estado, seguridad y monitoreo de las instancias.

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

Estos son algunos de los aprendizajes clave de esta sección del módulo:

- Elastic Load Balancing distribuye el tráfico entrante de las aplicaciones o de la red entre varios destinos (como las instancias de Amazon EC2, los contenedores, las direcciones IP y las funciones de Lambda) en una o más zonas de disponibilidad.
- Elastic Load Balancing admite tres tipos de balanceadores de carga:
 - Balanceador de carga de aplicaciones
 - Balanceador de carga de red
 - Balanceador de carga clásico
- Elastic Load Balancing ofrece varias herramientas de monitoreo para el registro y monitoreo continuos destinados a las auditorías y los análisis.

Módulo 10: Monitoreo y escalado automático

Sección 2: Amazon CloudWatch

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.



Sección 2: Amazon CloudWatch

Monitoreo de los recursos de AWS



Para utilizar AWS de manera eficiente, necesita información acerca de sus recursos de AWS:

- ¿Cómo se sabe cuándo se debe **lanzar más instancias de Amazon EC2**?
- ¿Se está viendo afectado el **rendimiento o la disponibilidad de su aplicación** por insuficiencia de capacidad?
- En realidad, ¿cuánto **se está usando** de su infraestructura?

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.15


Para utilizar AWS de forma eficiente, necesita información sobre sus recursos de AWS.

Por ejemplo, podría ser necesario saber lo siguiente:

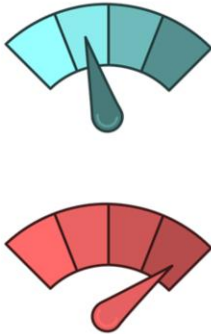
- ¿Cuándo debería lanzar más instancias de Amazon EC2?
- ¿Se está viendo afectado el rendimiento o la disponibilidad de su aplicación por insuficiencia de capacidad?
- En realidad, ¿cuánto se está usando de su infraestructura?

¿Cómo se registra esta información?

Amazon CloudWatch



Amazon CloudWatch



- Monitoreo:
 - Recursos de AWS
 - Aplicaciones que se ejecutan en AWS
- Recopilación y seguimiento:
 - Métricas estándar
 - Métricas personalizadas
- Alarmas:
 - Enviar notificaciones a un tema de Amazon SNS
 - Efectuar acciones de Amazon EC2 Auto Scaling o Amazon EC2
- Eventos:
 - Definir reglas para que coincidan con los cambios en el entorno de AWS y dirigir estos eventos a uno o más flujos o funciones de destino para su procesamiento

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

14

Puede registrar esta información con Amazon CloudWatch.

Amazon CloudWatch es un servicio de monitoreo y observabilidad creado para los ingenieros de DevOps, los desarrolladores, los ingenieros de fiabilidad de sitio (SRE) y los administradores de TI. CloudWatch monitorea sus recursos de AWS (y las aplicaciones que ejecuta en AWS) en tiempo real. Puede utilizar CloudWatch para recopilar y hacer un seguimiento de las métricas, que son las variables referidas a sus recursos y aplicaciones que puede medir.

Puede crear una alarma para monitorear cualquier métrica de Amazon CloudWatch en su cuenta y utilizar la alarma para enviar automáticamente una notificación a un tema de Amazon Simple Notification Service (Amazon SNS) o efectuar una acción de Amazon EC2 Auto Scaling o de Amazon EC2. Por ejemplo, puede crear alarmas acerca del uso de la CPU de una instancia EC2, la latencia de solicitudes de Elastic Load Balancing, el rendimiento de la tabla de Amazon DynamoDB, la longitud de la cola de Amazon Simple Queue Service (Amazon SQS) o, incluso, los cargos de su factura de AWS. También puede crear una alarma para las métricas personalizadas que sean específicas de su infraestructura o aplicaciones personalizadas.

Además, puede utilizar Amazon CloudWatch Events para definir reglas que coincidan con los eventos entrantes (o con los cambios en su entorno de AWS) para dirigirlos a

los destinos para su procesamiento. Los destinos pueden incluir instancias de Amazon EC2, funciones de AWS Lambda, transmisiones de Kinesis, tareas de Amazon ECS, máquinas de estado de Step Functions, temas de Amazon SNS, colas de Amazon SQS y destinos integrados. CloudWatch Events toma conocimiento de los cambios operativos a medida que se producen. CloudWatch Events responde a estos cambios operativos y toma medidas correctivas según sea necesario, enviando mensajes para responder al entorno, activando funciones, realizando cambios y captando información del estado.

Con CloudWatch, obtiene visibilidad de todo el sistema respecto de la utilización de los recursos, el rendimiento de las aplicaciones y el estado operativo. No existen compromisos iniciales ni tarifas mínimas, y solo paga por lo que usa. El uso se le cobra al final del mes.

Las alarmas de CloudWatch aws academy

- Cree alarmas en función de lo siguiente:
 - Límite estático
 - Detección de anomalías
 - Expresión matemática de métricas
- Especifique lo siguiente:
 - Espacio de nombres
 - Métrica
 - Estadística
 - Periodo
 - Condiciones
 - Configuración adicional
 - Acciones

Statistic
Average

Period
5 minutes

Conditions

Threshold type
☒ Static Use a value as a threshold
☐ Anomaly detection Use a band as a threshold

Whenever CPUUtilization is...
 Define the alarm condition
☒ Greater > threshold
☐ Greater/Equal >= threshold
☐ Lower/Equal <= threshold
☐ Lower < threshold

than...
 Define the threshold value
 100
 Must be a number

▶ Additional configuration

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados. 15

Puede crear una alarma de CloudWatch que supervise una sola métrica de CloudWatch o el resultado de una expresión matemática que se base en las métricas de CloudWatch. Puede crear una alarma de CloudWatch que se base en un límite estático, en la detección de anomalías o en una expresión matemática métrica.

Cuando se crea una alarma basada en un límite estático, se elige una métrica de CloudWatch para que la alarma supervise y el límite para dicha métrica. La alarma pasa al estado de ALARMA cuando la métrica supera el límite en una cantidad especificada de periodos de evaluación.

Para una alarma basada en un límite estático, debe especificar lo siguiente:


- *Espacio de nombres*: un espacio de nombres contiene la métrica de CloudWatch que desee. Por ejemplo, *AWS/EC2*.
- *Métrica*: la métrica es la variable que desea medir. Por ejemplo, *el uso de la CPU*.
- *Estadística*: una estadística puede ser un promedio, una suma, un mínimo, un máximo, un recuento de muestra, un percentil predefinido o un percentil personalizado.
- *Periodo*: el periodo es el periodo de evaluación de la alarma. Cuando se evalúa la






alarma, cada periodo se acumula en un punto de datos.

- *Condiciones*: cuando se especifican las condiciones de un límite estático, se debe especificar si la métrica debe ser *Mayor*, *Mayor o igual*, *Menor o igual* o *Menor* que el valor del límite, y, también, se debe especificar el valor del límite.
- *Información de configuración adicional*: incluye la cantidad de puntos de datos dentro del periodo de evaluación que debe infringirse para activar la alarma y cómo CloudWatch debe tratar los datos faltantes cuando evalúa la alarma.
- *Acciones*: puede elegir enviar una notificación a un tema de Amazon SNS o efectuar una acción de Amazon EC2 Auto Scaling o de Amazon EC2.

Para obtener más información acerca de la creación de alarmas de CloudWatch, consulte los temas de [Uso de alarmas](#) en la documentación de AWS.

Actividad: Amazon CloudWatch



 Amazon EC2	Si el uso promedio de la CPU es > 60 % durante 5 minutos...	Correcto.
 Amazon RDS	Si la cantidad de conexiones simultáneas es > 10 durante 1 minuto...	Correcto.
 Amazon S3	Si el tamaño máximo del bucket en bytes es de alrededor de 3 durante 1 día...	Incorrecto. <i>Alrededor</i> no es una opción de límite. Debe especificar un límite de >, >=, <= o <.
 Elastic Load Balancing	Si la cantidad de hosts en buen estado es < 5 durante 10 minutos...	Correcto.
 Amazon Elastic Block Store	Si el volumen de operaciones de lectura es > 1000 durante 10 segundos...	Incorrecto. Debe especificar una estadística (por ejemplo, el <i>volumen promedio</i>).

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

1b

Para esta actividad, trate de identificar cuáles son alarmas correctas de CloudWatch. Para las que son incorrectas, trate de identificar el error.

Aprendizajes clave de la sección 2



1 /



- Amazon CloudWatch lo ayuda a monitorear sus recursos de AWS y las aplicaciones que ejecuta en AWS en tiempo real.
- CloudWatch le permite realizar lo siguiente:
 - Recopilar y hacer un seguimiento de las métricas estándar y personalizadas.
 - Establecer alarmas para enviar notificaciones automáticas a los temas de SNS o efectuar acciones de Amazon EC2 Auto Scaling o Amazon EC2.
 - Definir reglas que coincidan con los cambios en su entorno de AWS y dirigir estos eventos a los destinos para su procesamiento.

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

Estos son algunos de los aprendizajes clave de esta sección del módulo:

- Amazon CloudWatch lo ayuda a monitorear sus recursos de AWS y las aplicaciones que ejecuta en AWS en tiempo real.
- CloudWatch le permite realizar lo siguiente:
 - Recopilar y hacer un seguimiento de las métricas estándar y personalizadas.
 - Establecer alarmas para enviar notificaciones automáticas a los temas de SNS o efectuar acciones de Amazon EC2 Auto Scaling o Amazon EC2 en función del valor de la métrica o expresión referido a un límite durante cierta cantidad de periodos.
 - Definir reglas que coincidan con los cambios en su entorno de AWS y dirigir estos eventos a los destinos para su procesamiento.

Módulo 10: Monitoreo y escalado automático

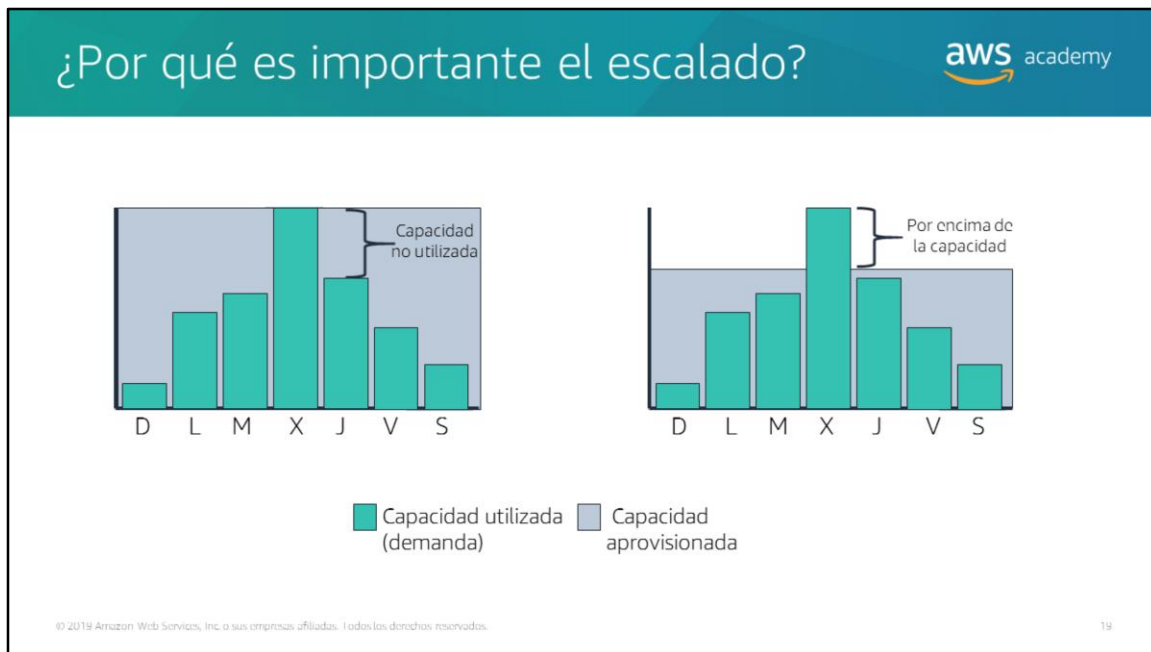
Sección 3: Amazon EC2 Auto Scaling

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.



Sección 3: Amazon EC2 Auto Scaling

Cuando ejecute sus aplicaciones en AWS, es conveniente que se asegure de que su arquitectura pueda escalar a fin de gestionar los cambios en la demanda. En esta sección, aprenderá a escalar de forma automática sus instancias EC2 con Amazon EC2 Auto Scaling.

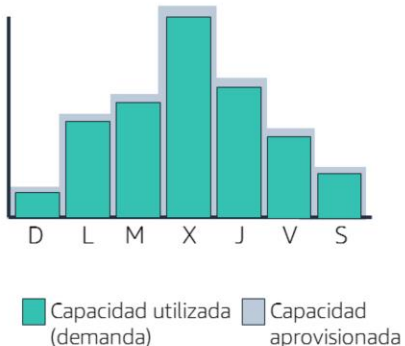


El escalado es la capacidad de aumentar o reducir la capacidad de cómputo de una aplicación. Para comprender por qué es importante el escalado, considere este ejemplo de una carga de trabajo que tiene diferentes requisitos de recursos. En este ejemplo, la mayor capacidad de recursos se requiere el miércoles y, la menor capacidad de recursos, los domingos.

Una opción consiste en asignar más capacidad de la necesaria de manera de poder satisfacer siempre la demanda más alta, que, en este caso, es la de los miércoles. Sin embargo, esta situación implica que está ejecutando recursos que no se utilizarán por completo la mayoría de los días de la semana. Con esta opción, los costos no están optimizados.

Otra opción consiste en asignar menos capacidad para reducir los costos. Esta situación implica que ciertos días no dispondrá de la capacidad suficiente. Si no resuelve el problema de la capacidad, la aplicación podría tener un rendimiento inferior o, incluso, podría dejar de estar disponible para los usuarios.

Amazon EC2 Auto Scaling



Día	Capacidad utilizada (demanda)	Capacidad aprovisionada
D	Baja	Baja
L	Medio-Baja	Medio-Baja
M	Medio	Medio
X	Alta	Alta
J	Medio-Alta	Medio-Alta
V	Medio-Baja	Medio-Baja
S	Baja	Baja

- Lo ayuda a mantener la disponibilidad de las aplicaciones
- Le permite agregar o eliminar automáticamente instancias EC2 de acuerdo con las condiciones que defina
- Detecta las instancias EC2 dañadas y las aplicaciones en mal estado, y reemplaza las instancias sin su intervención
- Ofrece varias opciones de escalado: manual, programado, dinámico o bajo demanda y predictivo

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

En la nube, como la capacidad de cómputo es un recurso programático, es posible enfocar el tema del escalado de una manera flexible. Amazon EC2 Auto Scaling es un servicio de AWS que lo ayuda a mantener la disponibilidad de la aplicación y le permite agregar o eliminar instancias EC2 de forma automática según las condiciones que defina. Puede utilizar las características de administración de flotas de EC2 Auto Scaling para mantener el estado y la disponibilidad de la suya.

Amazon EC2 Auto Scaling ofrece varias formas de ajustar el escalado para satisfacer las necesidades de sus aplicaciones de la mejor manera. Puede agregar o eliminar instancias EC2 manualmente, según una programación, en respuesta a los cambios en la demanda o en combinación con AWS Auto Scaling para el escalado predictivo. El escalado dinámico y el escalado predictivo se pueden utilizar juntos para escalar con mayor rapidez.

Para obtener más información acerca de Amazon EC2 Auto Scaling, consulte la página de producto de [Amazon EC2 Auto Scaling](#).

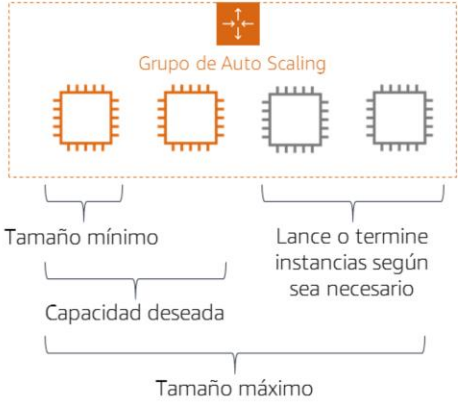


El escalado automático resulta útil para las cargas de trabajo predecibles, como, por ejemplo, el tráfico semanal en la empresa de venta minorista Amazon.com.



El escalado automático también es útil para el escalado dinámico bajo demanda. Amazon.com experimenta un pico de tráfico estacional en noviembre (el Black Friday y el Cyber Monday, que son días a finales de noviembre en los que los vendedores minoristas estadounidenses tienen ventas importantes). Si Amazon aprovisiona una capacidad fija que se ajuste al uso máximo, no se utilizará el 76 % de los recursos durante la mayor parte del año. El escalado de la capacidad es necesario para admitir las demandas de servicio fluctuantes. Sin el escalado, los servidores podrían bloquearse debido a la saturación y la empresa perdería la confianza de los clientes.

Grupos de Auto Scaling



Un **grupo de Auto Scaling** consiste en una colección de instancias EC2 que se tratan como una agrupación lógica a efectos de la administración y el escalado automático.

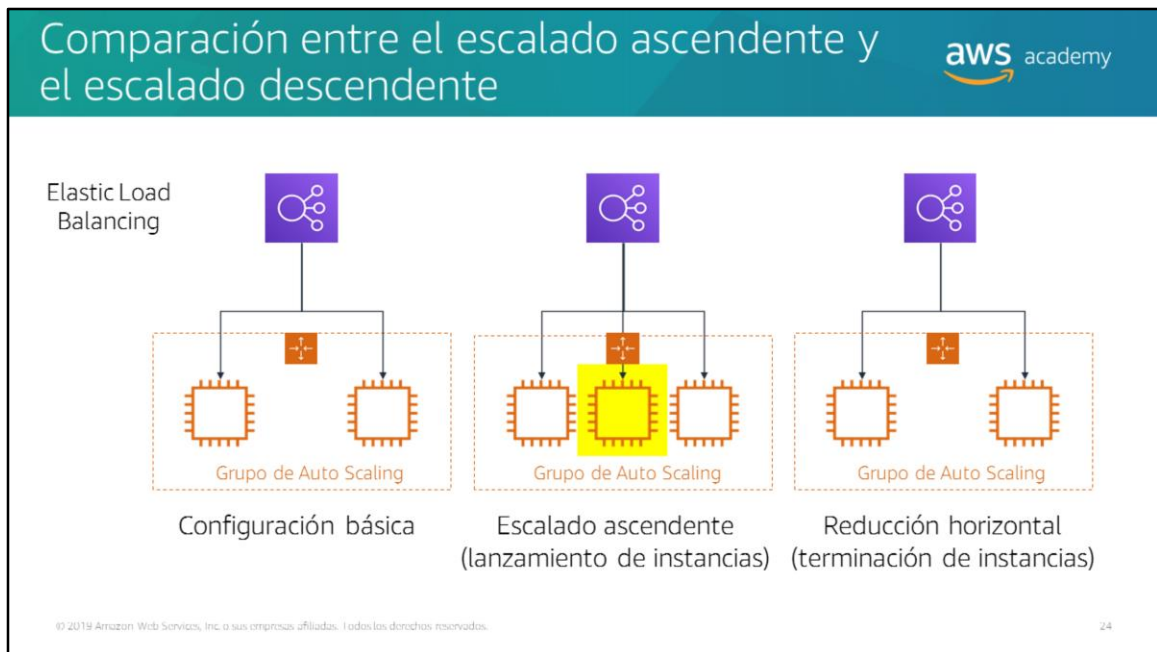
© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

25

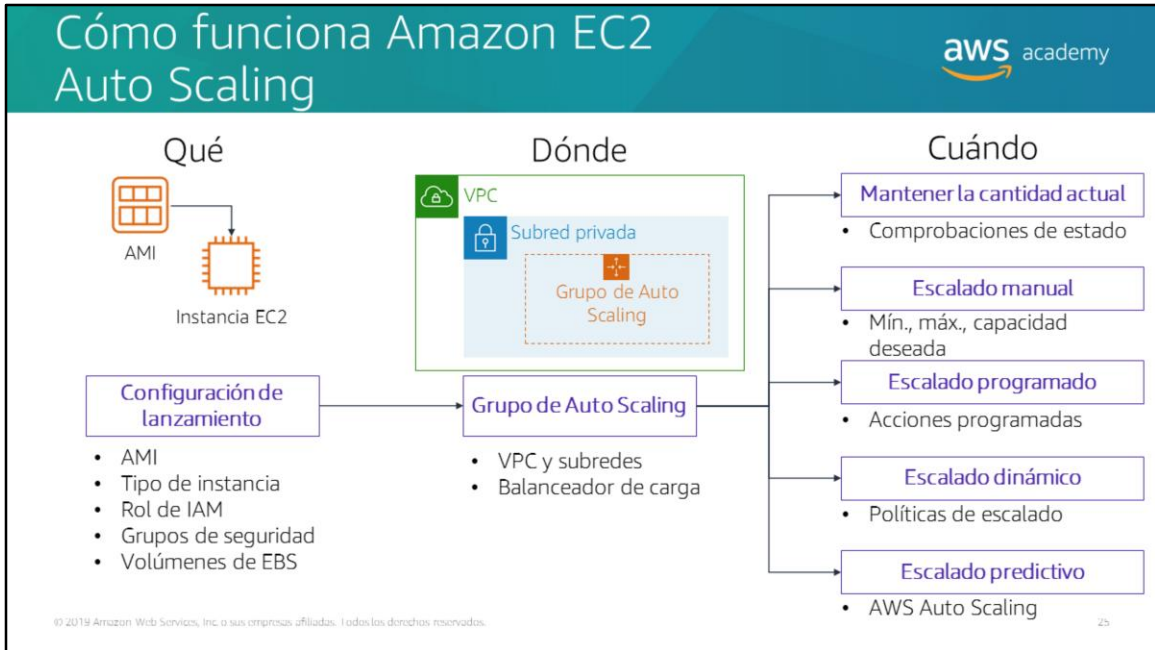
Un [grupo de Auto Scaling](#) es una colección de instancias de Amazon EC2 que se tratan como una agrupación lógica a efectos de la administración y el escalado automático. El tamaño de un grupo de Auto Scaling depende de la cantidad de instancias que se establecen como la *capacidad deseada*. Puede ajustar su tamaño para satisfacer la demanda, ya sea de forma manual o mediante el escalado automático.

Puede especificar la cantidad mínima de instancias en cada grupo de Auto Scaling y, al estar diseñado para ello, Amazon EC2 Auto Scaling evitará que su grupo se reduzca por debajo de este tamaño. Puede especificar la cantidad máxima de instancias en cada grupo de Auto Scaling y, al estar diseñado para ello, Amazon EC2 Auto Scaling evitará que su grupo supere este tamaño. Si especifica la capacidad deseada, ya sea al crear el grupo o en cualquier momento posterior, Amazon EC2 Auto Scaling está diseñado para ajustar el tamaño de su grupo de manera que tenga la cantidad especificada de instancias. Si especifica las políticas de escalado, Amazon EC2 Auto Scaling puede lanzar o terminar instancias en función de los aumentos o las disminuciones en las demandas de la aplicación.

Por ejemplo, este grupo de Auto Scaling tiene un tamaño mínimo de una instancia, una capacidad deseada de dos instancias y un tamaño máximo de cuatro instancias. Las políticas de escalado que defina ajustan la cantidad de instancias dentro del número mínimo y máximo de instancias, en función de los criterios que especifique.



Con Amazon EC2 Auto Scaling, el lanzamiento de instancias se denomina *escalado ascendente* y la terminación de instancias se denomina *escalado descendente*.



Para lanzar instancias EC2, un grupo de Auto Scaling utiliza una [configuración de lanzamiento](#), que es una plantilla de configuración de instancias. Puede considerar la configuración de lanzamiento como *qué* es lo que está escalando. Cuando se crea una configuración de lanzamiento, se especifica información sobre las instancias. La información que especifica incluye el ID de la Imagen de Amazon Machine (AMI), el tipo de instancia, el rol de AWS Identity and Access Management (IAM), el almacenamiento adicional, uno o más grupos de seguridad y cualquier volumen de Amazon Elastic Block Store (Amazon EBS).

Usted define las cantidades mínima y máxima de instancias y la capacidad deseada de su grupo de Auto Scaling. A continuación, lo lanza a una subred dentro de una VPC (puede considerar esto como *dónde* está escalando). Amazon EC2 Auto Scaling se integra con Elastic Load Balancing a fin de permitirle asociar uno o más balanceadores de carga a un grupo de Auto Scaling existente. Tras asociar el balanceador de carga, registra automáticamente las instancias del grupo y distribuye el tráfico entrante entre las instancias.

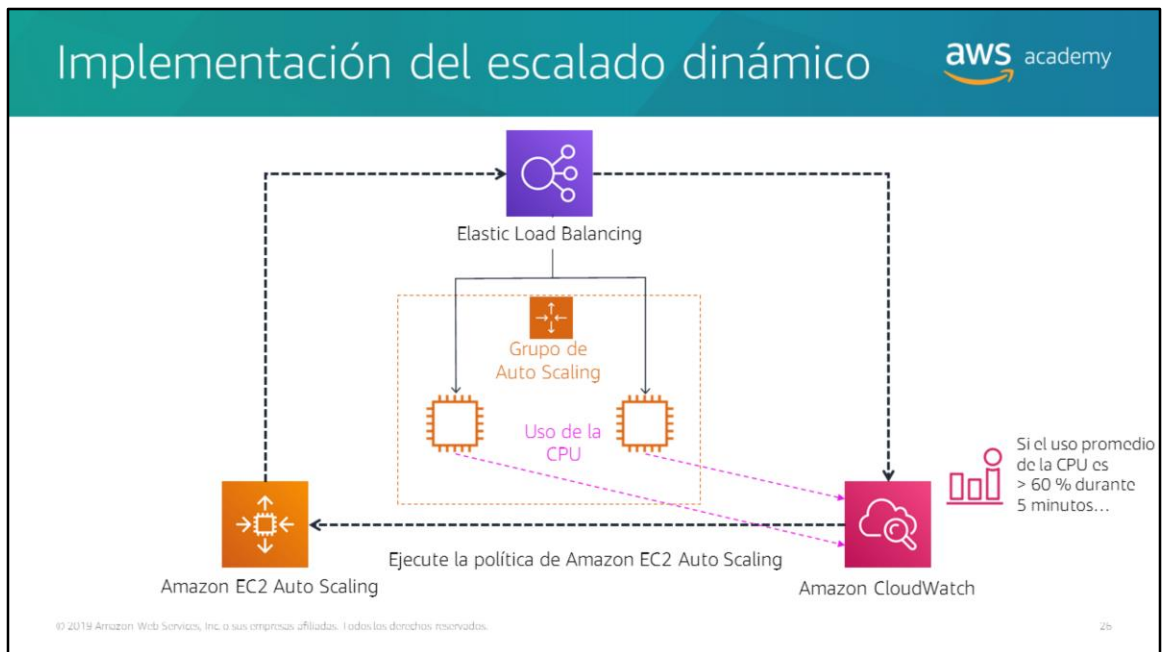
Por último, debe especificar *cuándo* desea que se produzca el evento de escalado. Existen muchas opciones de escalado:

- *Mantener los niveles actuales de la instancia en todo momento:* puede configurar su grupo de Auto Scaling para mantener un número específico de instancias en

ejecución en todo momento. Para mantener los niveles de instancia actuales, Amazon EC2 Auto Scaling efectúa una comprobación de estado periódica de las instancias en ejecución dentro un grupo de Auto Scaling. Cuando Amazon EC2 Auto Scaling encuentra una instancia en mal estado, la termina y lanza una nueva.

- [Escalado manual](#): con el escalado manual solo debe especificar el cambio en la capacidad máxima, mínima o deseada de su grupo de Auto Scaling.
- [Escalado programado](#): con el escalado programado, las acciones de escalado se efectúan automáticamente en función de la fecha y la hora. Esto resulta útil para las cargas de trabajo predecibles cuando sabe exactamente cuándo aumentar o reducir la cantidad de instancias de su grupo. Por ejemplo, consideremos que cada semana, el tráfico de su aplicación web comienza a aumentar el miércoles, permanece alto el jueves y comienza a disminuir el viernes. Puede programar las acciones de escalado en función de los patrones de tráfico predecibles de su aplicación web. Para implementar el escalado programado, cree una [acción programada](#).
- [Escalado dinámico bajo demanda](#): una forma más avanzada de escalar sus recursos le permite definir los parámetros que controlan el proceso de escalado. Por ejemplo, tiene una aplicación web que actualmente se ejecuta en dos instancias y desea que el uso de la CPU del grupo de Auto Scaling permanezca cerca del 50 % cuando la carga de la aplicación cambia. Esta opción resulta útil cuando el escalado responde a los cambios en las condiciones, pero no se sabe en qué momento cambiarán estas condiciones. El escalado dinámico le brinda capacidad adicional para manejar los picos de tráfico sin tener que mantener una cantidad excesiva de recursos sin utilizar. Puede configurar el grupo de Auto Scaling a fin de que el escalado se realice automáticamente para cubrir esta necesidad. El [tipo de política de escalado](#) determina cómo se implementa la acción de escalado. Puede utilizar Amazon EC2 Auto Scaling con Amazon CloudWatch para activar la política de escalado en respuesta a una alarma.
- [Escalado predictivo](#): puede usar Amazon EC2 Auto Scaling con AWS Auto Scaling para implementar el escalado predictivo, según el cual su capacidad escala en función de la demanda prevista. El escalado predictivo utiliza datos que se recopilan a partir del uso real de EC2, y los datos además se informan mediante miles de millones de puntos de datos extraídos de nuestras propias observaciones. Luego, AWS utiliza modelos de aprendizaje automático bien entrenados para predecir el tráfico esperado (y el uso de EC2), incluidos los patrones diarios y semanales. El modelo necesita datos históricos de al menos 1 día para comenzar a hacer predicciones. El modelo se revalúa cada 24 horas para crear una predicción de las siguientes 48 horas. El proceso de predicción genera un plan de escalado que puede impulsar a uno o más grupos de instancias EC2 escaladas automáticamente.

Para obtener más información acerca de estas opciones, consulte [Escalar el tamaño de su grupo de Auto Scaling](#) en la documentación de AWS.




Una configuración común para implementar el escalado dinámico consiste en crear una alarma de CloudWatch basada en la información de rendimiento de sus instancias EC2 o del balanceador de carga. Cuando se infringe un límite de rendimiento, una alarma de CloudWatch desencadena un evento de escalado automático que genera un escalado ascendente o descendente en las instancias EC2 del grupo de Auto Scaling.


Para comprender cómo funciona, considere este ejemplo:

- Se crea una alarma de Amazon CloudWatch para monitorear el uso de la CPU en su flota de instancias EC2 y ejecutar políticas de escalado automático si el uso promedio de la CPU en la flota supera el 60 % durante 5 minutos.
- Amazon EC2 Auto Scaling inicia una nueva instancia EC2 en su grupo de Auto Scaling en función de la configuración de lanzamiento que usted cree.
- Una vez agregada la nueva instancia, Amazon EC2 Auto Scaling realiza una llamada a Elastic Load Balancing para registrar la nueva instancia EC2 de ese grupo de Auto Scaling.
- Luego, Elastic Load Balancing lleva a cabo las comprobaciones de estado necesarias y comienza a distribuir tráfico a dicha instancia. Elastic Load Balancing dirige el tráfico entre las instancias EC2 y envía métricas a Amazon CloudWatch.

Amazon CloudWatch, Amazon EC2 Auto Scaling y Elastic Load Balancing funcionan bien individualmente. Sin embargo, juntos se vuelven más potentes y aumentan el control sobre cómo su aplicación gestiona la demanda de los clientes y la flexibilidad con la que la aplicación efectúa esta gestión.

AWS Auto Scaling





AWS Auto Scaling

- Monitorea sus aplicaciones y ajusta automáticamente la capacidad para mantener un rendimiento estable y predecible al menor costo posible.
- Proporciona una interfaz de usuario simple y potente que le permite crear planes de escalado para los siguientes recursos:
 - Instancias de Amazon EC2 y flotas de spot
 - Tareas de Amazon Elastic Container Service (Amazon ECS)
 - Tablas e índices de Amazon DynamoDB
 - Réplicas de Amazon Aurora

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

2 /

Hasta ahora, ha aprendido a escalar instancias EC2 con Amazon EC2 Auto Scaling. También ha aprendido que puede usar Amazon EC2 Auto Scaling con AWS Auto Scaling para efectuar el escalado predictivo.

AWS Auto Scaling es un servicio independiente que monitorea sus aplicaciones. Ajusta automáticamente la capacidad para mantener un rendimiento estable y predecible al menor costo posible. El servicio proporciona una interfaz de usuario sencilla y potente que le permite crear planes de escalado para los recursos, entre los que se incluyen los siguientes:


- Instancias de Amazon EC2 y flotas de spot
- Tareas de Amazon Elastic Container Service (Amazon ECS)
- Tablas e índices de Amazon DynamoDB
- Réplicas de Amazon Aurora

Si ya utiliza Amazon EC2 Auto Scaling para escalar las instancias EC2 de manera dinámica, ahora puede combinarlo con AWS Auto Scaling a fin de escalar recursos adicionales para otros servicios de AWS.


Para obtener más información acerca de AWS Auto Scaling, consulte [AWS](#)

[Auto Scaling.](#)

Aprendizajes clave de la sección 3



28



- El escalado le permite responder rápidamente a los cambios en las necesidades de recursos.
- Amazon EC2 Auto Scaling mantiene la disponibilidad de las aplicaciones agregando o eliminando instancias EC2 de manera automática.
- Un grupo de Auto Scaling es una colección de instancias EC2.
- Una configuración de lanzamiento es una plantilla de configuración de instancias.
- El escalado dinámico utiliza Amazon EC2 Auto Scaling, CloudWatch y Elastic Load Balancing.
- AWS Auto Scaling es un servicio independiente de Amazon EC2 Auto Scaling.

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

Estos son algunos de los aprendizajes clave de esta sección del módulo:

- El escalado le permite responder rápidamente a los cambios en las necesidades de recursos.
- Amazon EC2 Auto Scaling lo ayuda a mantener la disponibilidad de la aplicación y le permite agregar o eliminar instancias EC2 de forma automática según las cargas de trabajo.
- Un grupo de Auto Scaling es una colección de instancias EC2.
- Una configuración de lanzamiento es una plantilla de configuración de instancias.
- Puede implementar el escalado dinámico con Amazon EC2 Auto Scaling, Amazon CloudWatch y Elastic Load Balancing.

AWS Auto Scaling es un servicio independiente que monitorea sus aplicaciones y ajusta automáticamente la capacidad para los siguientes recursos:

- Instancias de Amazon EC2 y flotas de spot
- Tareas de Amazon ECS
- Tablas e índices de Amazon DynamoDB
- Réplicas de Amazon Aurora

Laboratorio 6: Ajuste de la escala y balanceo de la carga de su arquitectura

29

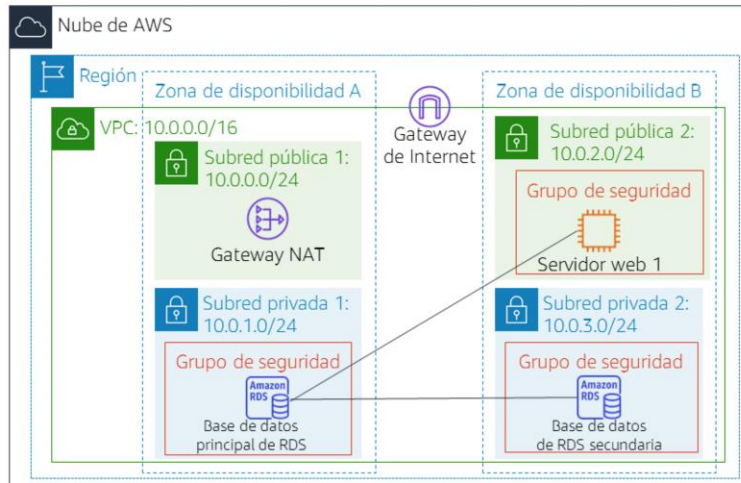
aws academy



© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

Ahora completará el Laboratorio 6: Ajuste de la escala y balanceo de la carga de su arquitectura.

Laboratorio 6: situación



© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

50

En este laboratorio, utilizará Elastic Load Balancing y Amazon EC2 Auto Scaling para balancear la carga y escalar su infraestructura. Comenzará con la infraestructura proporcionada.

Laboratorio 6: tareas



- Cree una imagen de Amazon Machine (AMI) a partir de una instancia en ejecución.
- Cree un balanceador de carga de aplicaciones.
- Cree una configuración de lanzamiento y un grupo de Auto Scaling.
- Escale automáticamente instancias nuevas dentro de una subred privada.
- Cree alarmas de Amazon CloudWatch y monitoree el rendimiento de la infraestructura.

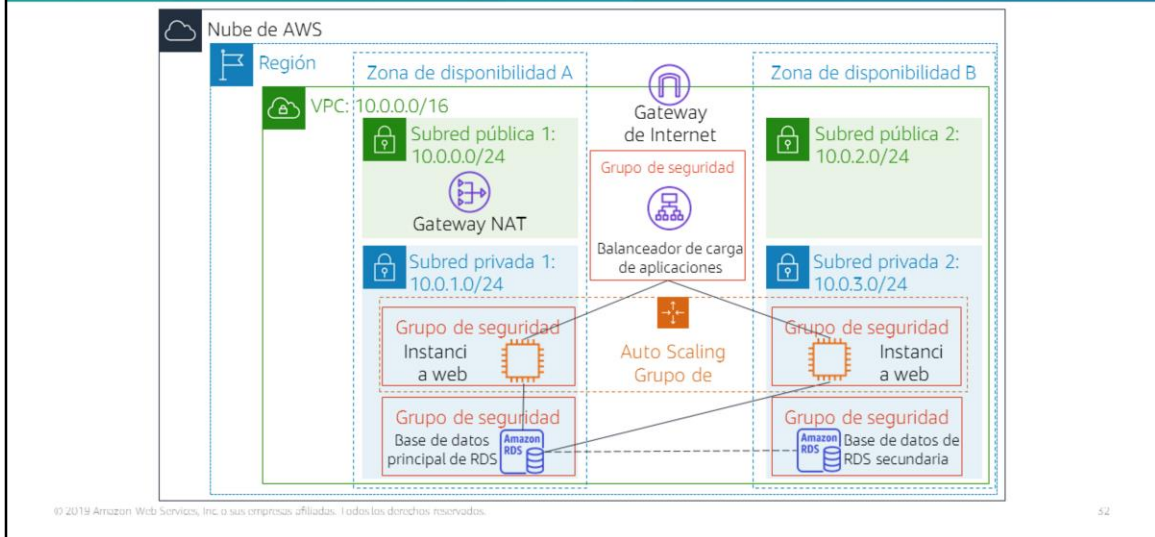
© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

51


En este laboratorio, realizará las siguientes tareas:

- Cree una imagen de Amazon Machine (AMI) a partir de una instancia en ejecución.
- Cree un balanceador de carga de aplicaciones.
- Cree una configuración de lanzamiento y un grupo de Auto Scaling.
- Escale automáticamente instancias nuevas dentro de una subred privada.
- Cree alarmas de Amazon CloudWatch y monitoree el rendimiento de la infraestructura.

Laboratorio 6: producto final




El diagrama resume lo que habrá creado después de completar el laboratorio.



Aprox. 30 minutos






Comience el Laboratorio 6:
Ajuste de la escala
y balanceo de la carga
de su arquitectura


© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

55

Ha llegado el momento de iniciar el laboratorio. Debería tardar aproximadamente 30 minutos en completarlo.



Análisis posterior
del laboratorio:
Aprendizajes
clave



© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

54

En este laboratorio, hizo lo siguiente:

- Crear una Imagen de Amazon Machine (AMI) a partir de una instancia en ejecución.
- Crear un balanceador de carga.
- Crear una configuración de lanzamiento y un grupo de Auto Scaling
- Escalar automáticamente nuevas instancias dentro de una subred privada.
- Crear alarmas de Amazon CloudWatch y monitorear el rendimiento de su infraestructura.

Módulo 10: Monitoreo y escalado automático

Conclusión del módulo

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.



Ha llegado el momento de hacer un repaso del módulo y concluir con una revisión de conocimientos y un debate sobre una pregunta del examen de certificación de prueba.

Resumen del módulo



En resumen, en este módulo, aprendió a hacer lo siguiente:

- Indicar cómo distribuir el tráfico entre las instancias de Amazon Elastic Compute Cloud (Amazon EC2) usando Elastic Load Balancing
- Identificar cómo Amazon CloudWatch le permite monitorear los recursos y las aplicaciones de AWS en tiempo real
- Explicar cómo Amazon EC2 Auto Scaling lanza y publica servidores en respuesta a los cambios en las cargas de trabajo
- Efectuar tareas de escalado y balanceo de carga para mejorar una arquitectura

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

5b

En resumen, en este módulo, aprendió a hacer lo siguiente:

- Indicar cómo distribuir el tráfico entre las instancias de Amazon Elastic Compute Cloud (Amazon EC2) usando Elastic Load Balancing
- Identificar cómo Amazon CloudWatch permite monitorear recursos y aplicaciones de AWS en tiempo real
- Explicar cómo Amazon EC2 Auto Scaling lanza y publica servidores en respuesta a los cambios en las cargas de trabajo
- Efectuar tareas de escalado y balanceo de carga para mejorar una arquitectura

Complete la revisión de conocimientos



© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

5/

Complete la revisión de conocimientos de este módulo.

Pregunta del examen de muestra



¿Qué servicio utilizaría para enviar alertas basadas en las alarmas de Amazon CloudWatch?

- A. Amazon Simple Notification Service
- B. AWS CloudTrail
- C. AWS Trusted Advisor
- D. Amazon Route 53

© 2019 Amazon Web Services, Inc. o sus empresas afiliadas. Todos los derechos reservados.

58

Observe las opciones de respuesta y descarte algunas en función de las palabras clave que se destacaron previamente.



¡Gracias por participar!