COLOMBIA POTENCIA DE LA VIDA

TIC

LECCIÓN 2: ESCALADO DE LOS RECURSOS INFORMÁTICOS.

TALENTO TECH | AZ PROYECTOS EDUCATIVOS | UTP Universidad Tecnológica de Pereira

COLOMBIA POTENCIA DE LA VIDA

TIC

¿QUÉ ES LA ELASTICIDAD?

Una infraestructura elástica puede ampliarse y contraerse a medida que cambian las necesidades de capacidad.

EJEMPLOS:

- **Aumento de la cantidad de servidores web si se incrementa el tráfico**
- **Reducción de la capacidad de escritura en la base de datos si el tráfico disminuye**
- **Manejo de la fluctuación diaria de la demanda en toda la arquitectura**

TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

Una característica de una arquitectura reactiva es la elasticidad. La elasticidad significa que la infraestructura puede ampliarse o contraerse si cambian las necesidades de capacidad.

Puede adquirir recursos cuando los necesite y liberar recursos cuando no los necesite. La elasticidad le permite lo siguiente:

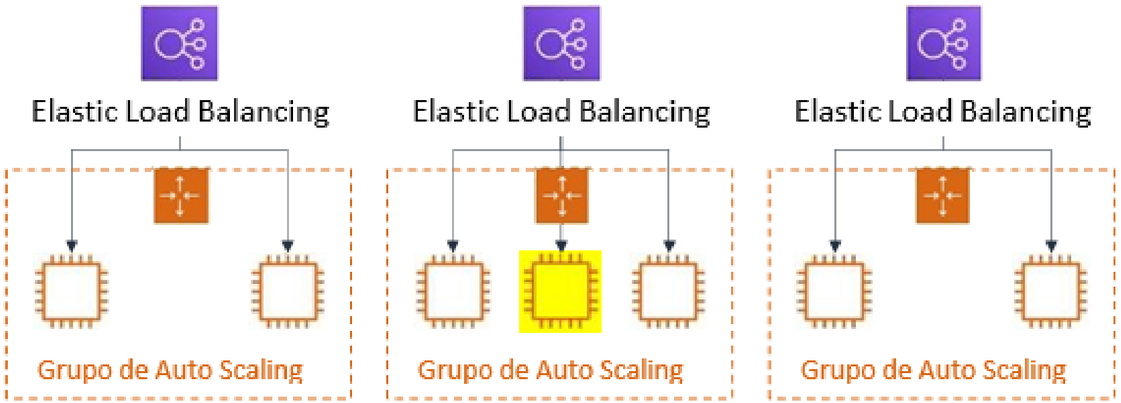
- **Aumentar la cantidad de servidores web si el tráfico a la aplicación se incrementa**
- **Reducir la capacidad de escritura en la base de datos si el tráfico disminuye**
- **Manejar la fluctuación diaria de la demanda en toda la arquitectura**

En el ejemplo de la cafetería, la elasticidad es importante porque, después de que se emita el programa de televisión, se podría constatar un aumento inmediato del tráfico del sitio web. El tráfico podría caer a niveles normales una semana después, o bien podría aumentar de nuevo durante las temporadas festivas.

¿QUÉ ES UN ESCALADO?

UNA TÉCNICA QUE SE UTILIZA PARA LOGRAR ELASTICIDAD

Escalado horizontal

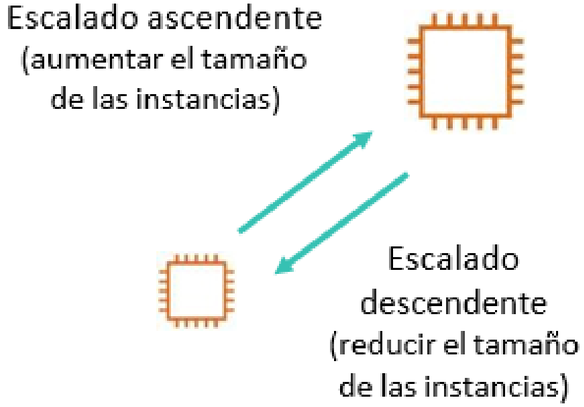


Configuración básica

Escalado ascendente (lanzar instancias)

Escalado descendente (terminar instancias)

Escalado vertical



El escalado es una técnica que se utiliza para lograr elasticidad. El escalado es la capacidad de aumentar o reducir la capacidad de cómputo de una aplicación.

Hay dos tipos de escalado:

ESCALADO HORIZONTAL

ESCALADO VERTICAL

EL ESCALADO HORIZONTAL

Ocurre si se agregan o se eliminan recursos. Por ejemplo, es posible que necesite agregar más discos duros a una matriz de almacenamiento o agregar más servidores para admitir una aplicación. Agregar recursos se denomina escalado ascendente y terminar recursos se denomina escalado descendente. El escalado horizontal es una buena manera de crear aplicaciones a escala de Internet que aprovechen la elasticidad de la informática en la nube.

COLOMBIA POTENCIA DE LA VIDA

TIC

EL ESCALADO VERTICAL

Ocurre si aumenta o disminuye las especificaciones de un recurso determinado. Por ejemplo, podría actualizar un servidor para que tenga un disco duro más grande o una CPU más rápida. Con Amazon Elastic Compute Cloud (Amazon EC2), puede detener una instancia y cambiar su tamaño a un tipo de instancia que tenga más capacidad de RAM, CPU, E/S o redes. El escalado vertical puede llegar en algún momento a un límite, y no siempre es un enfoque eficiente o de disponibilidad alta. Sin embargo, es fácil de implementar y puede ser suficiente para muchos casos de uso, sobre todo a corto plazo.

AMAZON EC2 AUTO SCALING

Lanza o termina instancias en función de las condiciones especificadas

De forma automática, registra nuevas instancias con balanceadores de carga si así se especifica

Puede llevar a cabo lanzamientos en las zonas de disponibilidad

COLOMBIA POTENCIA DE LA VIDA

TIC

En la nube, el escalado se puede manejar automáticamente. 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 políticas definidas, las programaciones y las comprobaciones de estado. Si especifica las políticas de escalado, Amazon EC2 Auto Scaling puede lanzar o terminar instancias en función del aumento o la disminución de la demanda de la aplicación.

Amazon EC2 Auto Scaling se integra a Elastic Load Balancing: registra automáticamente nuevas instancias con balanceadores de carga para distribuir el tráfico entrante entre las instancias.

TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

Amazon EC2 Auto Scaling le permite crear arquitecturas de disponibilidad alta que abarcan varias zonas de disponibilidad en una región. Más adelante en esta unidad obtendrá más información acerca de la disponibilidad alta. Si una zona de disponibilidad pasa a estar en mal estado o no está disponible, Amazon EC2 Auto Scaling lanza nuevas instancias en una zona de disponibilidad que no se haya visto afectada. Cuando la zona de disponibilidad en mal estado vuelve a tener un estado correcto, Amazon EC2 Auto Scaling redistribuye automáticamente las instancias de la aplicación de manera uniforme entre todas las zonas de disponibilidad designadas.

OPCIONES DE ESCALADO

PROGRAMADO

Adecuado para cargas de trabajo predecibles

Caso de uso: Desactivación las instancias de desarrollo y durante la noche

ESCALADO SEGÚN LA FECHA Y LA HORA

ADMITE EL SEGUIMIENTO DE VALORES OBJETIVO

DINÁMICO

Adecuado para cambios en las condiciones

Caso de uso: Escalado según el uso de la CPU

PREDICTIVO

Adecuado para la demanda prevista

Caso de uso: **Manejo de un aumento en la carga de trabajo para el sitio web de comercio electrónico durante un evento importante de ventas**

ESCALADO SEGÚN EL APRENDIZAJE AUTOMÁTICO (ML)

CON AMAZON EC2 AUTO SCALING SE OFRECEN VARIAS OPCIONES DE ESCALADO PARA ATENDER LAS NECESIDADES DE LAS APLICACIONES DE LA MEJOR MANERA.

PROGRAMADO

DINÁMICO

PREDICTIVO

COLOMBIA POTENCIA DE LA VIDA

TIC

ESCALADO PROGRAMADO

Con el escalado programado, las acciones de escalado se efectúan automáticamente en función de la fecha y la hora. Esta característica resulta útil para las cargas de trabajo predecibles si sabe exactamente cuándo aumentar o reducir la cantidad de instancias del grupo.

Por ejemplo, imaginemos que todas las semanas el tráfico a la 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 la aplicación web. Para implementar el escalado programado, cree una acción programada.

ESCALADO DINÁMICO BAJO DEMANDA

Este enfoque es una forma más avanzada de escalar los recursos. Permite definir parámetros con que controlar el proceso de escalado. Por ejemplo, suponga que tiene una aplicación web que se ejecuta en dos instancias EC2. Desea que la utilización de la CPU del grupo de Auto Scaling permanezca cerca del 50 % si se modifica la carga de la aplicación.

Esta opción resulta útil si el escalado responde a cambios de las condiciones, pero no se sabe en qué momento cambiarán esas condiciones. El escalado dinámico le da capacidad adicional para manejar los picos de tráfico sin tener que mantener una cantidad excesiva de recursos inactivos. Puede configurar el grupo de Auto Scaling a fin de que el escalado se efectúe automáticamente para atender esa necesidad.

COLOMBIA POTENCIA DE LA VIDA

TIC

ESCALADO PREDICTIVO

Puede utilizar Amazon EC2 Auto Scaling con AWS Auto Scaling para implementar el escalado predictivo, por el cual la capacidad escala en función de la demanda prevista. El escalado predictivo utiliza datos que se recopilan a partir del uso real de Amazon EC2, y los datos además incorporan miles de millones de puntos de datos que se extraen de las observaciones de AWS.

Después AWS utiliza modelos de aprendizaje automático bien entrenados para predecir el tráfico por esperar (y el uso de Amazon EC2), incluidos los patrones diarios y semanales. El modelo necesita datos históricos de hace 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 con el que se puede impulsar uno o más grupos de instancias EC2 de escalado automático.

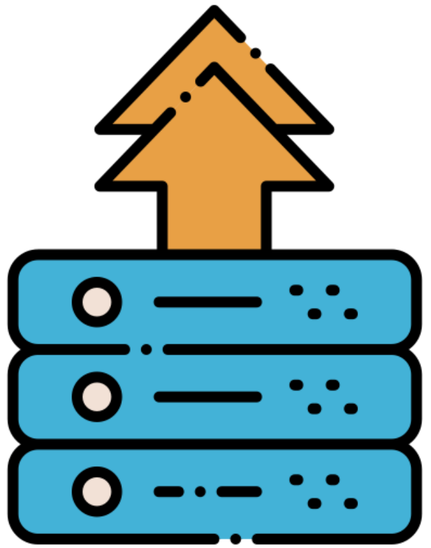
TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

El escalado dinámico y el escalado predictivo se pueden utilizar juntos para escalar la infraestructura con más rapidez.

Por último, también puede agregar o eliminar instancias EC2 manualmente. Con el escalado manual, solo debe especificar el cambio en la capacidad máxima, mínima o deseada de su grupo de Auto Scaling.



TIPOS DE POLÍTICAS DE ESCALADO DINÁMICO

ESCALADO SENCILLO: AJUSTE DE ESCALADO SENCILLO

Ejemplos de casos de uso: cargas de trabajo nuevas, picos de cargas de trabajo

ESCALADO POR PASOS: EL AJUSTE DEPENDE DEL TAMAÑO DE LA INTERRUPCIÓN DE ALARMA

Ejemplo de caso de uso: cargas de trabajo previsibles

ESCALADO DE SEGUIMIENTO DE VALORES OBJETIVO: VALOR OBJETIVO PARA UNA MÉTRICA DETERMINADA

Ejemplo de caso de uso: aplicaciones escalables horizontalmente, como aplicaciones de carga balanceada y aplicaciones de procesamiento de datos por lotes

Con las políticas de escalado por pasos y escalado sencillo, puede elegir las métricas de escalado y los valores de límite de las alarmas de CloudWatch que activan el proceso de escalado. También puede definir cómo debe escalarse el grupo de Auto Scaling si se supera un límite durante un número determinado de periodos de evaluación.

La principal diferencia es que con el escalado por pasos la capacidad actual del grupo de Auto Scaling aumenta o disminuye en función de un conjunto de ajustes de escalado, llamados ajustes por pasos, que varían en función del tamaño de la interrupción de alarma.

Con las políticas de escalado de seguimiento de valores objetivo, se aumenta o se reduce la capacidad actual del grupo en función de un valor objetivo de una métrica determinada. Este tipo de escalado se parece a la manera en que los termostatos mantienen la temperatura del hogar: el usuario selecciona una temperatura y el termostato hace el resto del trabajo.

Las políticas de escalado de seguimiento de valores objetivo le permiten seleccionar una métrica de escalado y establecer un valor objetivo. Con Amazon EC2 Auto Scaling, se crean y se administran las alarmas de CloudWatch que activan la política de escalado y calculan el ajuste de escalado en función del valor objetivo y la métrica. Con la política de escalado, se aumenta o se reduce la capacidad en función de las necesidades para mantener la métrica en el valor objetivo determinado o en un valor próximo. Además de mantener la métrica próxima al valor objetivo, la política de escalado de seguimiento de valores objetivo también se ajusta a los cambios de la métrica que se produzcan por una carga de trabajo que cambia.

Para obtener más información acerca de las políticas de escalado de seguimiento de valores objetivo, consulte los siguientes recursos:

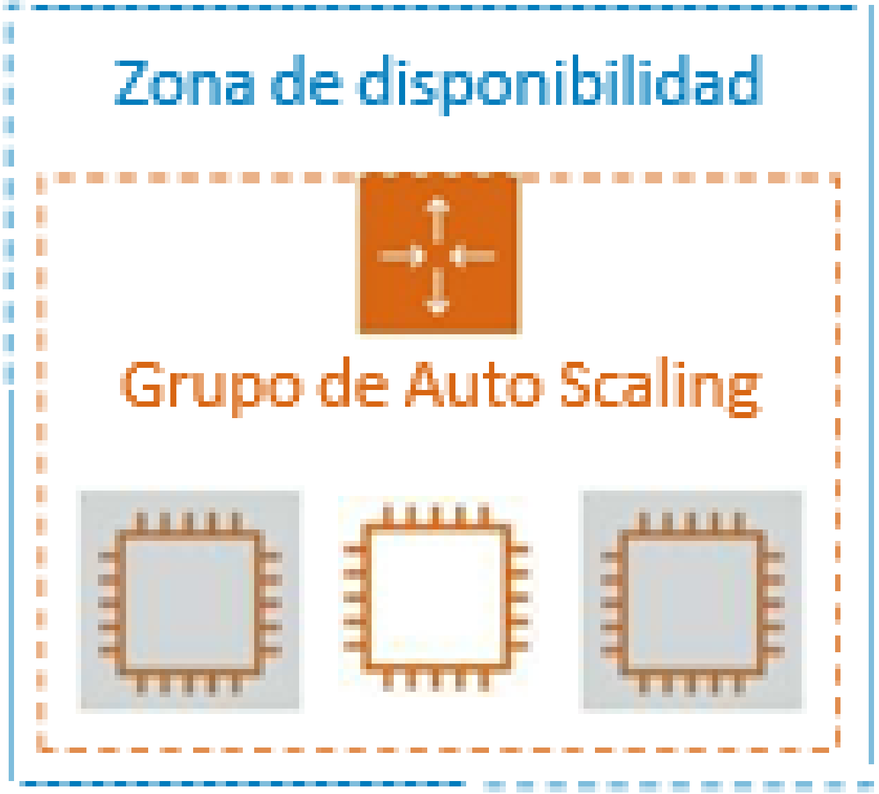
- TARGET TRACKING SCALING POLICIES FOR AMAZON EC2 AUTO SCALING
- SET IT AND FORGET IT AUTO SCALING TARGET TRACKING POLICIES
- AWS RE: INVENT 2017. AUTO SCALING PRIME TIME: TARGET TRACKING HITS THE BULLSEYE AT NETFLIX

GRUPOS DE AUTO SCALING

UN GRUPO DE AUTO SCALING DEFINE LO SIGUIENTE:

- LA CAPACIDAD MÍNIMA
- LA CAPACIDAD MÁXIMA
- LA CAPACIDAD DESEADA

La capacidad deseada refleja la cantidad de instancias que se estin ejecutando y puede fluctuar en respuesta a eventos.


Zona de disponibilidad
Grupo de Auto Scaling

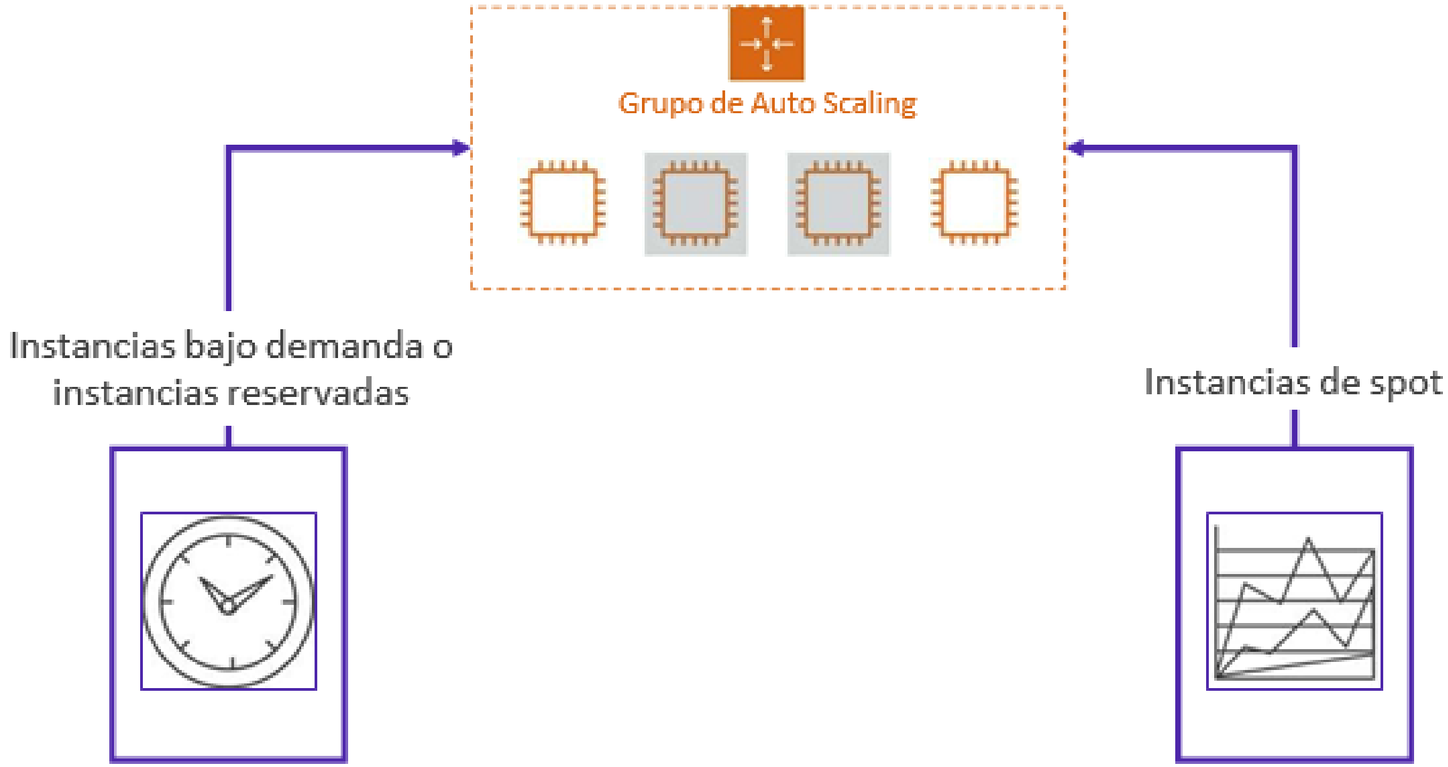
Amazon EC2 Auto Scaling lo ayuda a asegurarse de que cuenta con la cantidad correcta de instancias de Amazon EC2 para controlar la carga de la aplicación.

Crea colecciones de instancias EC2, denominadas grupos de Auto Scaling. Puede especificar la cantidad mínima de instancias en cada grupo de Auto Scaling, y Amazon EC2 Auto Scaling ayuda a que el grupo nunca tenga menos de esa cantidad de instancias. Puede especificar la cantidad máxima de instancias en cada grupo de Auto Scaling, y Amazon EC2 Auto Scaling ayuda a que el grupo nunca tenga menos de esa cantidad de instancias. Si especifica la capacidad deseada, ya sea cuando crea el grupo o después, Amazon EC2 Auto Scaling ayuda a que el grupo tenga esa cantidad de instancias.

Tenga en cuenta que la capacidad deseada es una configuración basada en desencadenador y que puede fluctuar en respuesta a eventos como un límite que se ha superado. Refleja la cantidad de instancias que se están ejecutando en ese momento. Nunca puede ser menor que el valor mínimo o mayor que el valor máximo. El rol de la política de escalado es actuar como representante automatizado y tomar decisiones sobre cómo ajustar la capacidad deseada. A continuación, Amazon EC2 Auto Scaling responde al cambio con la configuración de capacidad deseada.

Inicialmente, usted establece la capacidad deseada para indicar al grupo de Auto Scaling cuántas instancias desea ejecutar en un momento determinado. La cantidad de instancias que se están ejecutando puede ser diferente del valor deseado hasta que Amazon EC2 Auto Scaling las inicie o las termine.

AMAZON EC2 AUTO SCALING: OPCIONES DE COMPRA



Amazon EC2 Auto Scaling le permite hacer escalados ascendentes y descendentes en la infraestructura en respuesta a cambios en las condiciones. Cuando configura un grupo de Auto Scaling, puede especificar los tipos de instancias EC2 que utiliza. También puede especificar qué porcentaje de la capacidad deseada se debe atender con instancias bajo demanda, instancias reservadas e instancias de spot. Con Amazon EC2 Auto Scaling, después se aprovisiona de la combinación de instancias de precio más bajo para atender la capacidad deseada en función de estas preferencias.

Solo puede utilizar un tipo de instancia. Sin embargo, una práctica recomendada es utilizar unos cuantos tipos de instancias para no intentar iniciar instancias desde grupos de instancias que tengan capacidad insuficiente. Si la solicitud del grupo de Auto Scaling respecto de instancias de spot no se puede atender en un grupo de instancias de spot, el grupo de Auto Scaling sigue intentándolo en otros grupos de instancias de spot en lugar de lanzar instancias bajo demanda.

Para obtener más información acerca de las opciones de compra, consulte Grupos de Auto Scaling con varios tipos de instancias y opciones de compra.

CONSIDERACIONES RELATIVAS AL ESCALADO AUTOMÁTICO

- Varios tipos de escalado automático
- Escalado sencillo, por pasos o de seguimiento de valores objetivo
- Varias métricas (no solo relativas a la CPU)
- Cuándo utilizar el escalado ascendente y cuándo utilizar el escalado descendente
- Uso de enlaces de ciclo de vida

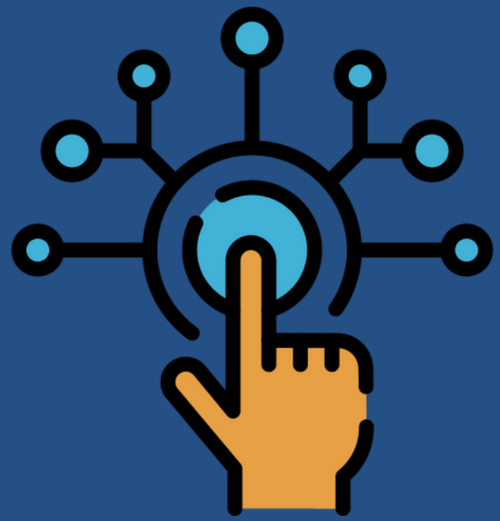
COLOMBIA POTENCIA DE LA VIDA

TIC

LAS COSAS QUE DEBE TENER EN CUENTA CUANDO UTILICE AMAZON EC2 AUTO SCALING PARA ESCALAR LA ARQUITECTURA SON LAS SIGUIENTES:

Varios tipos de escalado automático:

Es posible que necesite implementar una combinación de escalado programado, dinámico y predictivo.

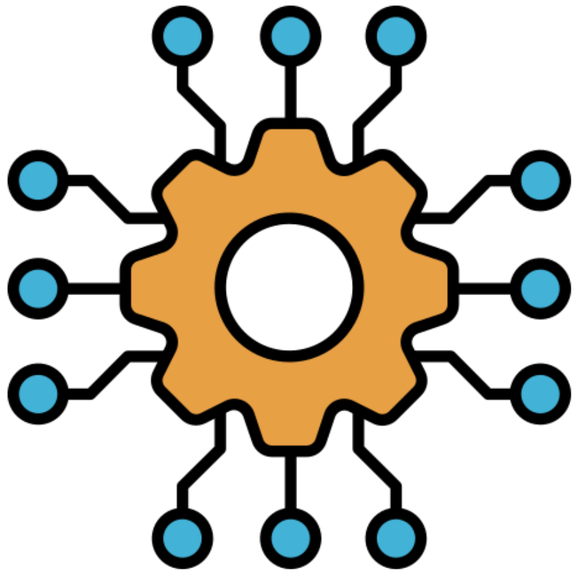


TALENTO TECH

AZ | PROYECTOS EDUCATIVOS

UTP Universidad Tecnológica de Pereira

Tipo de política de escalado dinámico:

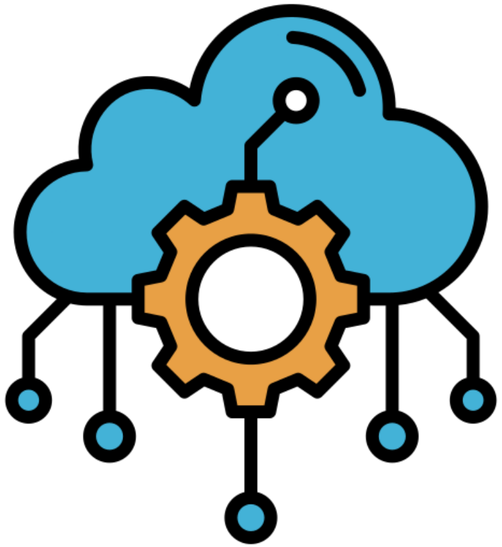


Con las políticas de escalado simple se aumenta o se disminuye la capacidad actual del grupo en función de un único ajuste de escalado. Con las políticas de escalado por pasos, se aumenta o se disminuye la capacidad actual del grupo en función de un conjunto de ajustes de escalado, llamados ajustes por pasos, que varían en función de la magnitud de la interrupción de alarma. Con las políticas de escalado de seguimiento de valores objetivo, se aumenta o se reduce la capacidad actual del grupo en función de un valor objetivo de una métrica determinada.

Varias métricas:

Algunas arquitecturas deben escalar en dos o más métricas (no solo relativas a la CPU). AWS recomienda que utilice una política de escalado de seguimiento de valores objetivo para escalar en una métrica, como la de la utilización media de la CPU o la métrica RequestCountPerTarget del balanceador de carga de aplicaciones. Las métricas que disminuyen si aumenta la capacidad y que aumentan si disminuye la capacidad se pueden utilizar para hacer un escalado proporcional ascendente o descendente de la cantidad de instancias que utilizan el seguimiento de valores objetivo. Con este tipo de métrica, se ayuda a velar por que Amazon EC2 Auto Scaling siga detenidamente la curva de demanda de las aplicaciones.

Cuándo hacer un escalado ascendente y cuándo hacer un escalado descendente:



Intente hacer los escalados ascendentes temprano y rápido, y haga los escalados descendentes de forma lenta y conforme avanza el tiempo.

Uso deenlaces de ciclo de vida :

Con los enlaces de ciclo de vida puede hacer acciones personalizadas pausando instancias si un grupo de Auto Scaling las lanza o las termina. Cuando se pausa, la instancia permanece en estado de espera hasta que lleva a cabo la acción de ciclo de vida mediante el comando complete-lifecycle-action o la operación CompleteLifecycleAction, o hasta que se cumpla el periodo del tiempo de espera (1 hora, de forma predeterminada).

Ahora el instructor puede elegir demostrar cómo crear políticas de escalado de seguimiento de valores objetivo y de escalado por pasos para Amazon EC2 Auto Scaling.

ESTOS SON ALGUNOS DE LOS APRENDIZAJES CLAVE DE ESTA LECCIÓN DE LA UNIDAD:

- Una infraestructura elástica puede ampliarse y contraerse a medida que cambian las necesidades de capacidad
- Amazon EC2 Auto Scaling agrega o elimina automáticamente instancias EC2 de acuerdo con las políticas definidas, las programaciones y las comprobaciones de estado
- Amazon EC2 Auto Scaling ofrece varias opciones de escalado para atender las necesidades de las aplicaciones de la mejor manera
- Cuando configura un grupo de Auto Scaling, puede especificar los tipos de instancias EC2 y la combinación de modelos de precios que utiliza