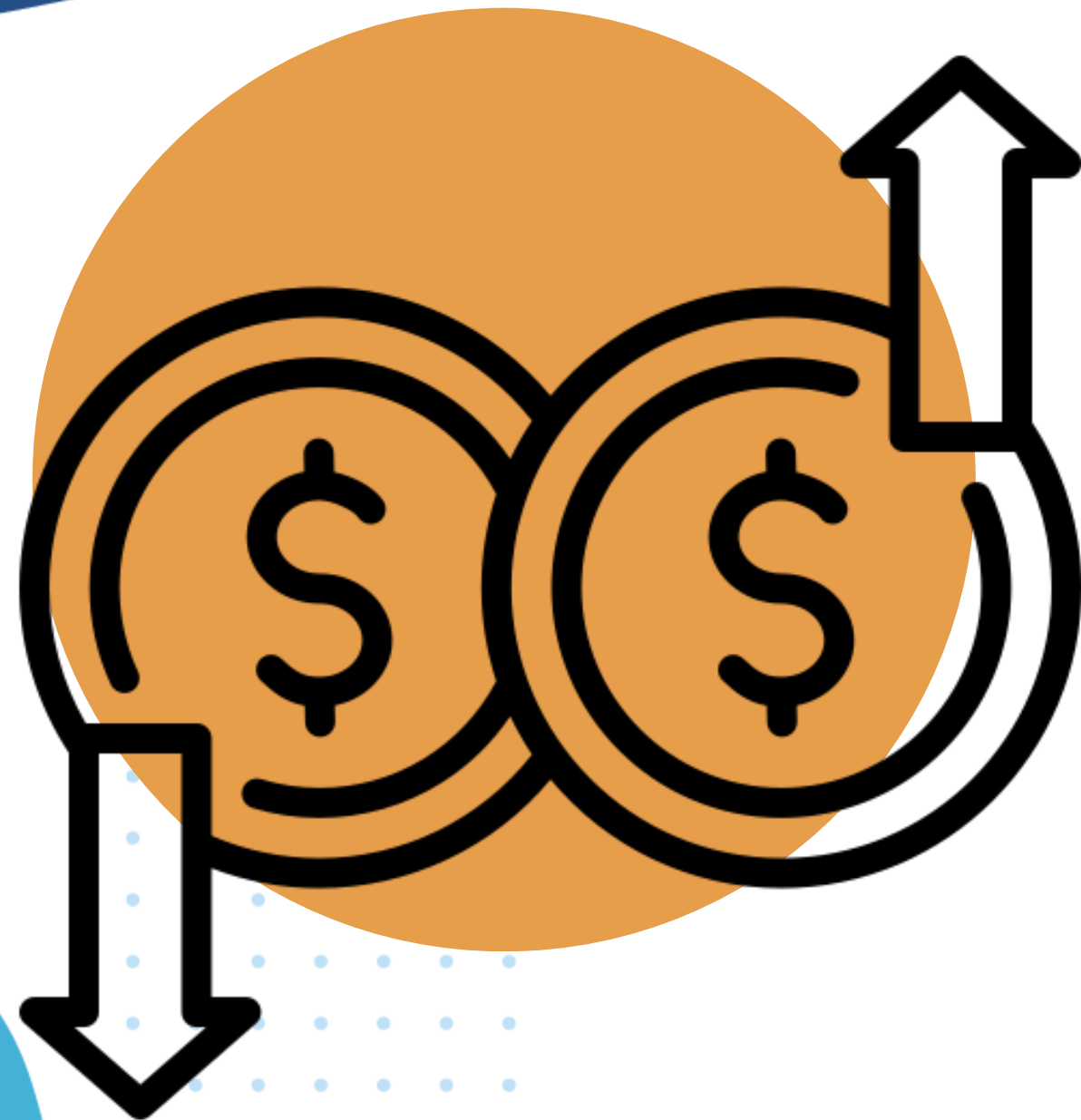




Lección 3

Optimización de costos de Amazon EC2





Instancias bajo demanda

- Pago por hora.
- No asume compromisos a largo plazo.
- Apto para el nivel gratuito de AWS.

Servidores dedicados

- Un servidor físico con capacidad de instancias de EC2 totalmente dedicado a su uso.

Instancias dedicadas

- Instancias que se ejecutan en una VPC en hardware dedicado a un solo cliente.

Instancias reservadas

- Pago total, parcial o sin pago inicial para la instancia que reserva.
- Descuento en el cargo horario por uso de dicha instancia.
- Plazo de 1 a 3 años.

Instancias reservadas programadas

- Adquiera una reserva de capacidad que esté disponible en un horario recurrente que especifique.
- Plazo de 1 año.

La facturación por segundo solo está disponible para las instancias bajo demanda, las instancias reservadas y las instancias de spot que ejecutan Amazon, Linux o Ubuntu.

Instancias de spot

- Las instancias se ejecutan siempre que estén disponibles y su puja sea superior al precio de las instancias de spot.
- AWS puede interrumpirlas con una notificación de 2 minutos.
- Las opciones de interrupción incluyen terminación, parada o hibernación.
- Los precios pueden ser de forma significativa menos costosos en comparación con las instancias bajo demanda.
- Una buena elección cuando tiene flexibilidad en cuanto a cuándo se pueden ejecutar sus aplicaciones.

Amazon ofrece diferentes modelos de precios entre los cuales elegir cuando desee ejecutar instancias EC2.

- La facturación por segundo solo está disponible para las instancias bajo demanda, las instancias reservadas y las instancias de spot que ejecutan Amazon Linux o Ubuntu.

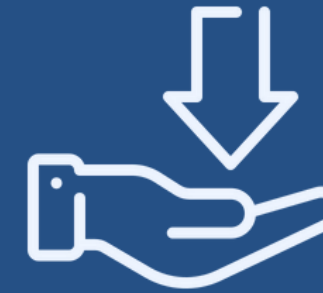
- Las instancias bajo demanda son aptas para la capa gratuita de AWS (<https://aws.amazon.com/free/>). Tienen el costo inicial más bajo y la mayor flexibilidad. No es necesario efectuar contratos por adelantado ni compromisos a largo plazo. Son una buena opción para aplicaciones con cargas de trabajo a corto plazo, con picos de demanda o impredecibles.
- Los alojamientos dedicados son servidores físicos con capacidad de instancia dedicada para su uso. Le permiten utilizar sus licencias de software existentes por socket, por núcleo o por máquina virtual, como Microsoft Windows o Microsoft SQL Server.
- Las instancias dedicadas son instancias de EC2 que se ejecutan en una nube virtual privada (VPC) en un hardware dedicado a un único cliente. Se aíslan físicamente de las instancias que pertenecen a otras cuentas de AWS a nivel del hardware que es host.



- Las instancias reservadas le permiten reservar capacidad de cómputo durante un año o tres años, con costos de funcionamiento menores por hora. El precio de uso con descuento se fija mientras sea propietario de la instancia reservada. Si espera un uso uniforme e intensivo, pueden proporcionar ahorros sustanciales en comparación con las instancias bajo demanda.
- Las instancias reservadas programadas le permiten comprar reservas de capacidad que se repiten diaria, semanal o mensualmente, con una duración especificada, por un periodo de 1 año. Paga por el tiempo en que se programan las instancias, incluso si no las utiliza.
- Las instancias de spot le permiten pujar por instancias EC2 no utilizadas, lo que puede reducir los costos. El precio por hora de una instancia de spot fluctúa en función de la oferta y la demanda. Su instancia de spot se ejecuta siempre que su puja supere el precio actual del mercado.



Modelos de precios de Amazon EC2: beneficios



Instancias bajo demanda
Ofrece bajo costo y flexibilidad



Instancias de spot
Carga de trabajo dinámica y a gran escala



Instancias reservadas
Ofrece previsibilidad, lo que garantiza que la capacidad de cómputo esté disponible cuando sea necesario



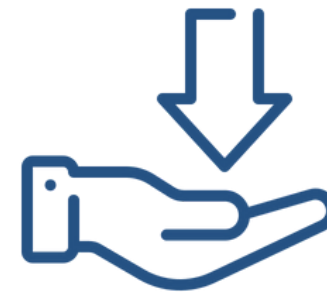
Servidores dedicados

- Permite ahorrar en costos de licencia
- Ayuda a cumplir los requisitos normativos y de conformidad

Cada modelo de precios de Amazon EC2 proporciona un conjunto diferente de beneficios.

- Las instancias bajo demanda ofrecen la mayor flexibilidad, sin contratos a largo plazo y con tarifas bajas.
- Las instancias de spot ofrecen gran escala a un precio con un descuento significativo.
- Las instancias reservadas son una buena opción si tiene necesidades de cómputo predecibles o de estado estable (por ejemplo, una instancia que sabe que desea seguir ejecutando casi todo el tiempo o durante meses o años).
- Los servidores dedicados son una buena opción cuando existen restricciones de licencia para el software que desea ejecutar en Amazon EC2 o cuando tiene requisitos normativos o de cumplimiento específicos que le impiden utilizar las demás opciones de implementación.

Modelos de precios de Amazon EC2: casos de uso



Instancias bajo demanda

- Cargas de trabajo a corto plazo, con picos de demanda o impredecibles
- Desarrollo o prueba de aplicaciones



Instancias de spot

- Aplicaciones con tiempos de inicio y finalización flexibles.
- Aplicaciones que solo son viables con precios de cómputo muy bajos
- Usuarios con necesidades de cómputo urgentes de grandes cantidades de capacidad adicional

Instancias reservadas

- Cargas de trabajo de uso predecible o estado estable
- Aplicaciones que requieren capacidad reservada, incluida la recuperación ante desastres (DR)
- Usuarios capaces de afrontar pagos iniciales para reducir aún más los costos de cómputo



Servidores dedicados

- Licencia "Bring your own license" (BYOL)
- Conformidad y restricciones normativas
- Seguimiento del uso y las licencias
- Control en una ubicación de instancia



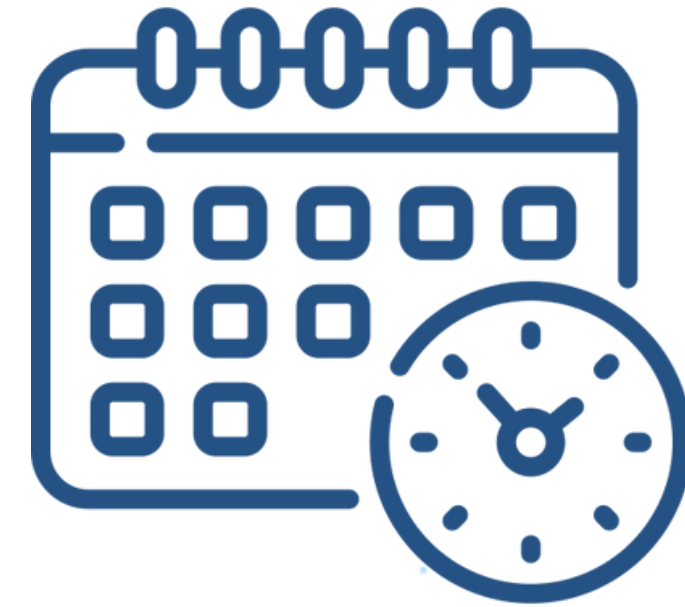
✗ Esta es una revisión de algunos casos de uso de las distintas opciones de precios.

Los precios de las instancias bajo demanda funcionan bien para cargas de trabajo con picos de demanda o si solo necesita probar o ejecutar una aplicación durante un breve periodo (por ejemplo, durante el desarrollo o las pruebas de aplicaciones). A veces, las cargas de trabajo son impredecibles y las instancias bajo demanda son una buena opción para estos casos.



Las instancias de spot son una buena opción si sus aplicaciones pueden tolerar interrupciones con una notificación de advertencia de 2 minutos. De forma predeterminada, las instancias se terminan, pero puede configurarlas para que se detengan o hibernen en su lugar. Los casos prácticos habituales incluyen aplicaciones tolerantes a errores, como servidores web, backends de API y procesamiento de big data. Las cargas de trabajo que guardan datos de forma constante en almacenamiento persistente (como Amazon S3) también son buenas candidatas.

Las instancias reservadas son una buena opción cuando tiene cargas de trabajo a largo plazo con patrones de uso predecibles, como servidores que sabe que deseará ejecutar de manera uniforme durante muchos meses.



Los alojamientos dedicados son una buena opción cuando tiene licencias de software por socket, por núcleo o por máquina virtual, o cuando debe cumplir requisitos normativos y de conformidad corporativa específicos.

Los cuatro pilares de la optimización de costos



Presionar cada tema para ver si contenido



Pilar 1 Dimensionamiento adecuado



Pilar 2 Aumento de la elasticidad



Pilar 3 Modelo de precios óptimo



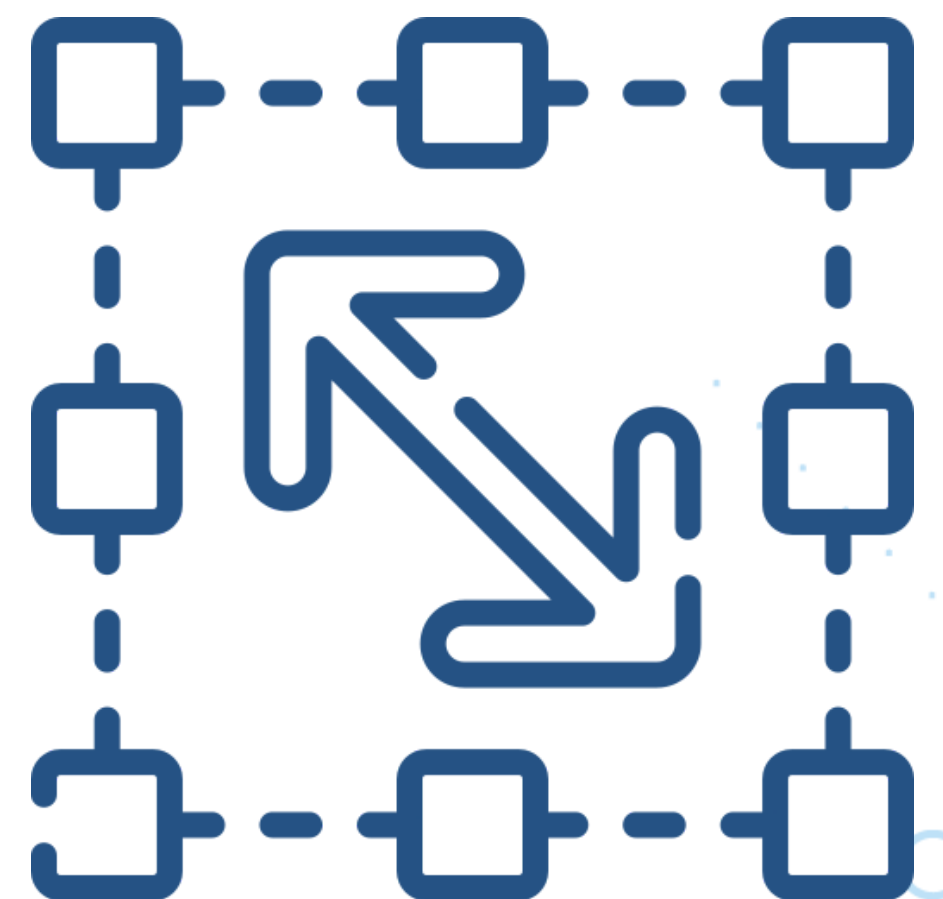
Pilar 4 Optimización de las opciones de almacenamiento



Optimización de costos

Para optimizar los costos, debe considerar cuatro factores potentes y consistentes:

- **Tamaño adecuado:** elija el equilibrio adecuado de tipos de instancias. Observe cuándo los servidores pueden reducirse o apagarse y aun así cumplir con los requisitos de rendimiento.
- **Aumento de la elasticidad:** diseña implementaciones para reducir la cantidad de capacidad del servidor que está inactiva mediante implementaciones que sean elásticas, tales como implementaciones que utilizan el escalado automático para manejar cargas máximas.





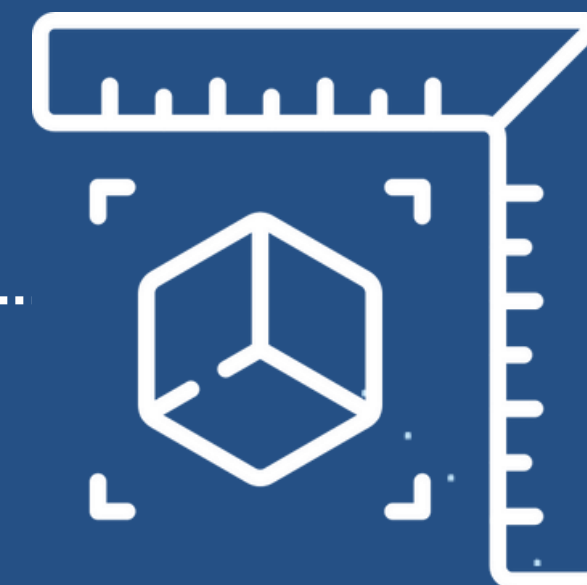
- **Modelo de precios óptimo:** reconoce las opciones de precios disponibles. Analice sus patrones de uso de manera que pueda ejecutar las instancias EC2 con la combinación adecuada de opciones de precios.
- **Optimización de las opciones de almacenamiento:** analiza los requisitos de almacenamiento de las implementaciones. Reduzca la sobrecarga de almacenamiento sin utilizar cuando sea posible y elegir opciones de almacenamiento más económicas si siguen cumpliendo con los requisitos de rendimiento de almacenamiento.



Pilar 1 Dimensionamiento adecuado



- Aprovisionar instancias para satisfacer la necesidad
- CPU, memoria, almacenamiento y rendimiento de red
 - Seleccionar los tipos de instancia apropiados para su uso
- Uso de métricas de Amazon CloudWatch
- ¿Qué tan inactivas están las instancias? ¿Cuándo?
 - Reducir el tamaño de las instancias
- Práctica recomendada: tamaño correcto, luego reserve





Primero, considere el tamaño correcto. AWS ofrece aproximadamente 60 tipos y tamaños de instancias. La amplia variedad de opciones permite a los clientes seleccionar la instancia que mejor se adapte a su carga de trabajo. Puede resultar difícil saber por dónde empezar y qué instancia será la mejor, tanto desde el punto de vista técnico como de costos. El ajuste del tamaño es el proceso de revisar los recursos implementados y buscar oportunidades para reducir el tamaño de las implementaciones cuando sea posible.



Dimensión correcta

- Seleccione la instancia más barata disponible que aún cumpla con sus requisitos de rendimiento.
- Verifica el uso de CPU, RAM, almacenamiento y red para identificar las instancias cuyo tamaño podría reducirse. Es posible que desee aprovisionar una variedad de tipos y tamaños de instancias en un entorno de prueba y luego probar su aplicación en esas diferentes implementaciones de prueba para identificar qué instancias ofrecen la mejor relación rendimiento-costos. Para lograr el tamaño correcto, utilice técnicas como las pruebas de carga a su favor.
- Utilice métricas de Amazon CloudWatch y configure métricas personalizadas. Una métrica representa un conjunto de valores ordenados en el tiempo que se publican en CloudWatch (por ejemplo, el uso de CPU de una instancia de EC2 particular). Los propios puntos de datos pueden proceder de cualquier aplicación o actividad empresarial desde la que se recopilen datos.

Pilar 2 Aumento de la elasticidad

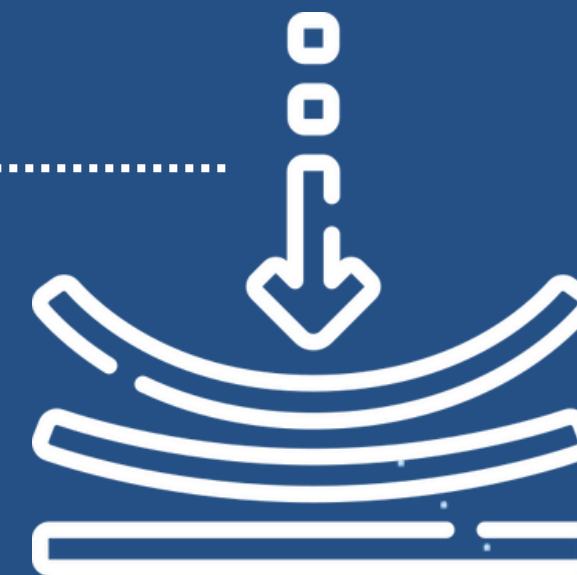


Detener o hibernar instancias respaldadas por Amazon EBS que estén en uso inactivo

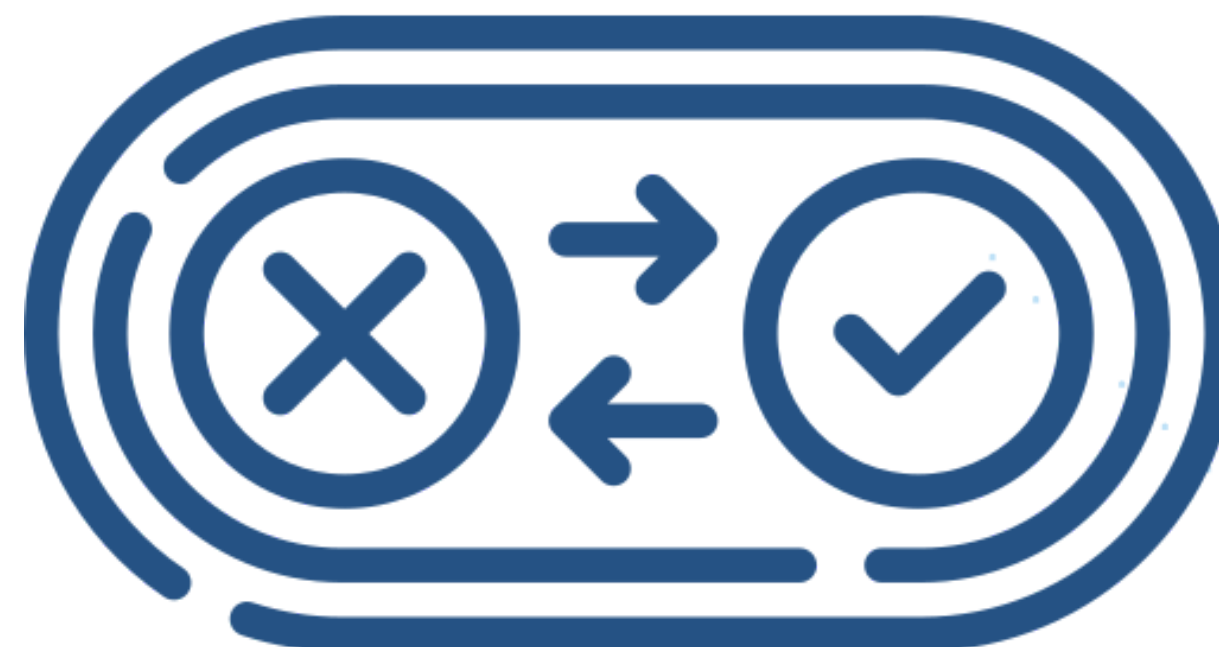
- Ejemplo: desarrollo no productivo o instancias de prueba

Utilice el escalado automático para satisfacer las necesidades de acuerdo con el uso

- Elasticidad automatizada y basada en el tiempo

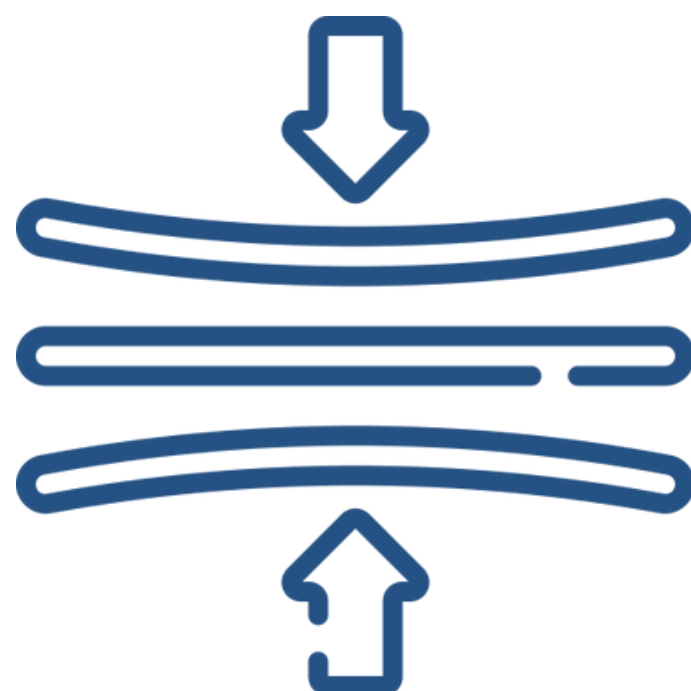


Una forma de elasticidad es crear, iniciar o usar instancias de EC2 cuando sean necesarias, pero luego desactivarlas cuando no estén en uso. La elasticidad es uno de los principios centrales de la nube, pero los clientes a menudo pasan por un proceso de aprendizaje para poner en práctica la elasticidad y generar ahorros de costos.



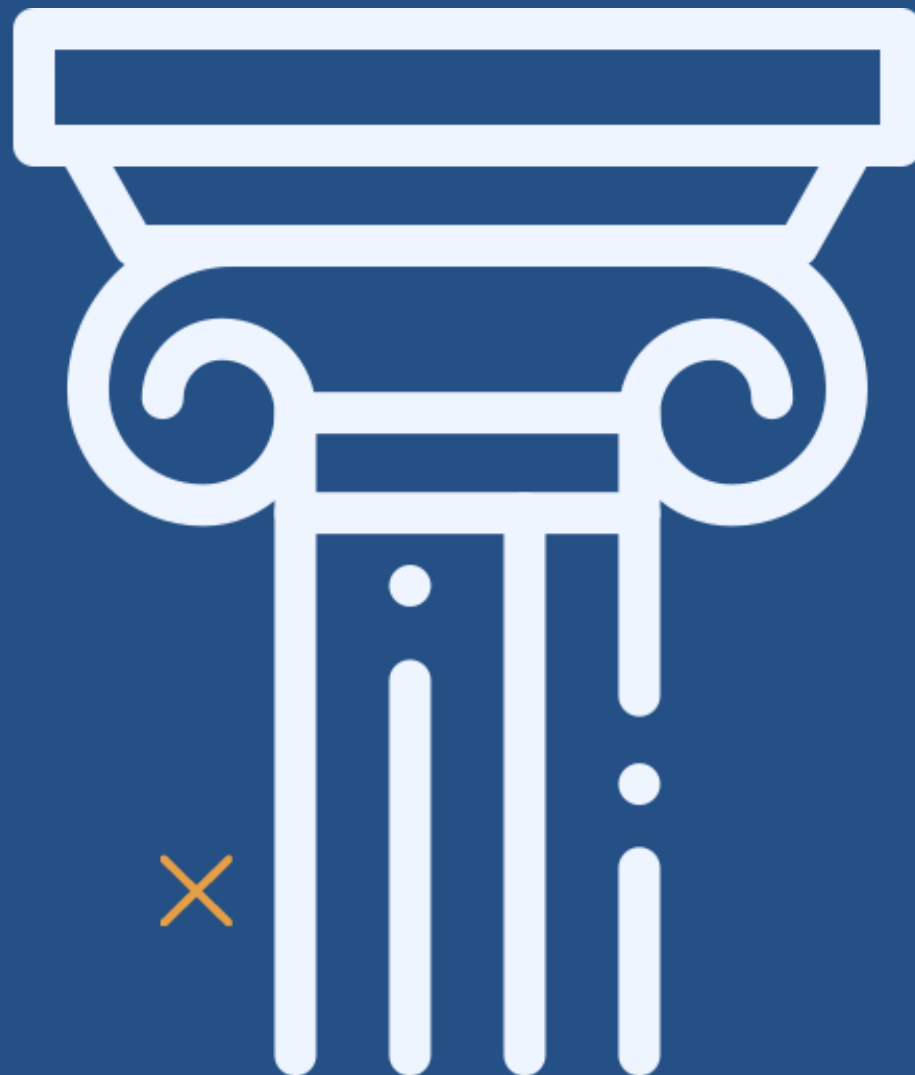
La forma más fácil para que los grandes clientes adopten la elasticidad es buscar recursos que parezcan buenos candidatos para detenerse o hibernar, como entornos que no son de producción, cargas de trabajo de desarrollo o cargas de trabajo de prueba. Por ejemplo, si ejecuta cargas de trabajo de desarrollo o prueba en una única zona horaria, puede desactivar fácilmente esas instancias fuera del horario comercial y así reducir los costos de tiempo de ejecución quizás en un 65 por ciento. El concepto es similar a por qué hay un interruptor de luz al lado de la puerta y por qué la mayoría de las oficinas alientan a los empleados a apagar las luces al salir de la oficina cada noche.

Para las cargas de trabajo de producción, configurar políticas de escalado automático más precisas y granulares puede ayudarle a aprovechar el escalado horizontal para satisfacer las necesidades de capacidad máxima y no pagar por la capacidad máxima todo el tiempo.



Como regla general, debe apuntar a que entre el 20 y el 30 por ciento de sus instancias Amazon EC2 se ejecuten como instancias bajo demanda o instancias de spot, y también debe buscar activamente formas de maximizar la elasticidad.

Pilar 3 Modelo de precios óptimo



Aproveche el modelo de precios adecuado para su caso práctico

- ¿Cuáles son los patrones de uso?

Optimizar y combinar tipos de compra

Ejemplos:

- Utilice instancias bajo demanda e instancias de spot para cargas de trabajo variables

Considere soluciones sin servidor (AWS Lambda)

AWS ofrece varios modelos de precios para Amazon EC2 para ayudar a los clientes a ahorrar dinero. Los modelos disponibles se analizaron en detalle antes en esta unidad. Los clientes pueden combinar múltiples tipos de compras para optimizar los precios en función de sus necesidades de capacidad actuales y previstas.

Además se anima a los clientes a considerar la arquitectura de su aplicación. Por ejemplo, ¿es necesario que la funcionalidad proporcionada por su aplicación se ejecute en una máquina virtual de Elastic Compute Cloud? Quizás al utilizar el servicio AWS Lambda, podría reducir significativamente sus costos.

Analizaremos AWS Lambda más adelante en esta unidad.



Pilar 4 Optimización de las opciones de almacenamiento



- Reduzca los costos manteniendo el rendimiento y la disponibilidad del almacenamiento
- Modificar el tamaño de los volúmenes de Amazon EBS
- Modificar los tipos de volúmenes de Amazon EBS
 - ¿Se pueden cumplir los requisitos de rendimiento con un almacenamiento menos costoso?
 - Ejemplo: el almacenamiento HDD (st1) con rendimiento optimizado de Amazon EBS por lo general cuesta la mitad que la opción de almacenamiento SSD de uso general (gp2) predeterminada.
- Elimine las instantáneas de EBS que ya no sean necesarias
- Identificar el destino más apropiado para tipos específicos de datos.
 - ¿La aplicación necesita que la instancia resida en Amazon EBS?
 - Las opciones de almacenamiento de Amazon S3 con políticas de ciclo de vida pueden reducir los costos

Los clientes también pueden reducir los costos de almacenamiento. Cuando inicia instancias de EC2, diferentes tipos de instancias ofrecen diferentes opciones de almacenamiento. Es una práctica recomendada el intentar reducir los costos y al mismo tiempo mantener el rendimiento y la disponibilidad del almacenamiento.

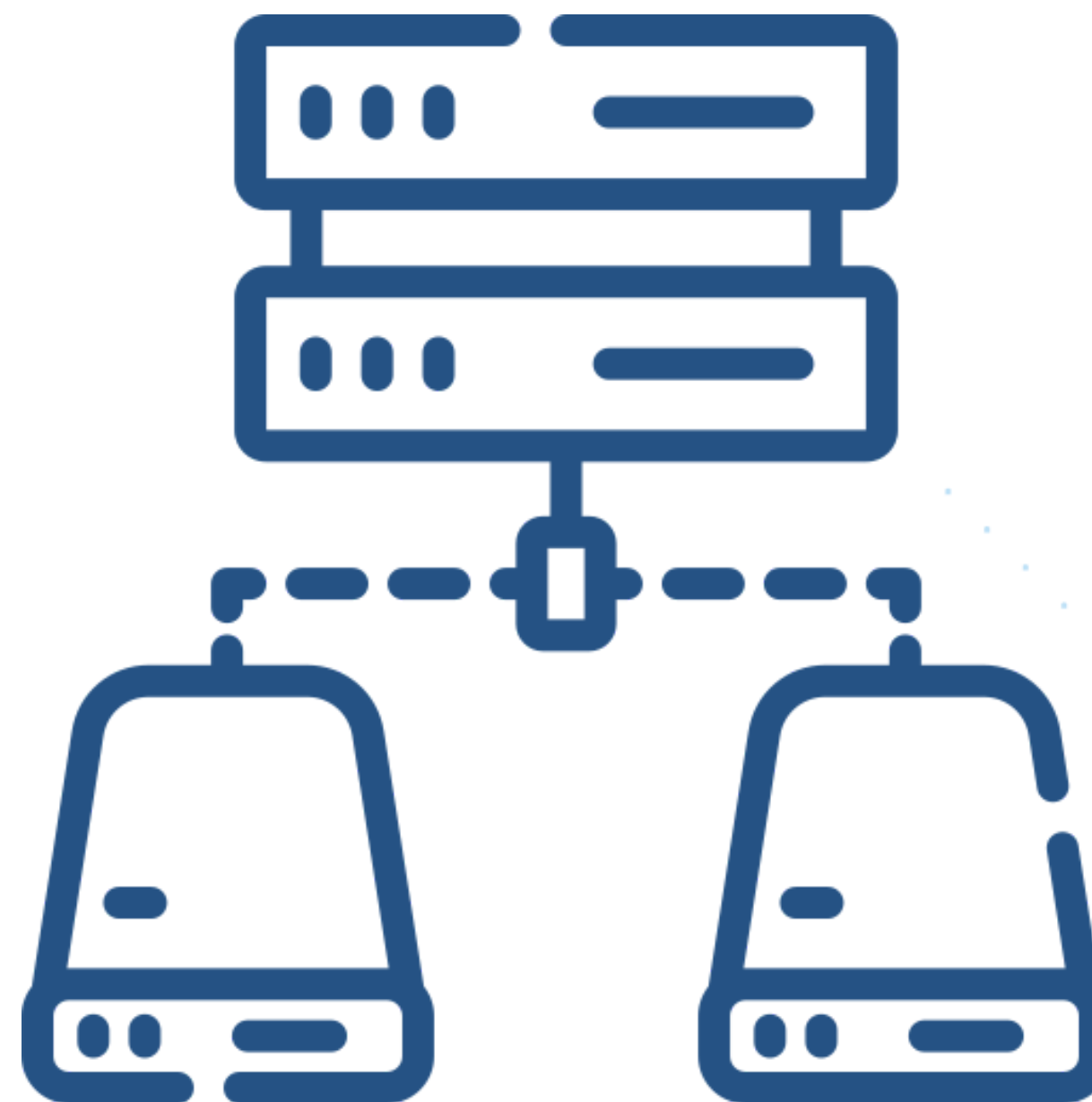


Una forma de lograrlo es mediante el redimensionamiento de los volúmenes de EBS. Por ejemplo, si en el inicio provisionó un volumen de 500 GB para una instancia de EC2 que solo necesitará un máximo de 20 GB de espacio de almacenamiento, puede reducir el tamaño del volumen y ahorrar costos.

También hay una variedad de tipos de volúmenes de EBS. Elija el tipo menos costoso que aún cumpla con sus requisitos de rendimiento. Por ejemplo, el almacenamiento HDD (st1) con rendimiento optimizado de Amazon EBS por lo general cuesta la mitad que la opción de almacenamiento SSD de uso general (gp2) predeterminada.

Si una unidad st1 satisface las necesidades de su carga de trabajo, aproveche el ahorro de costos.

Los clientes suelen utilizar instantáneas de EBS para crear respaldos de datos. Sin embargo, algunos clientes se olvidan de eliminar las instantáneas que son innecesarias. Elimine estas instantáneas innecesarias para ahorrar costos.





Finalmente, trate de identificar el destino más apropiado para tipos específicos de datos. ¿Su aplicación necesita que los datos que utiliza residan en Amazon EBS? ¿La aplicación se ejecutaría igual de bien si utilizara Amazon S3 para el almacenamiento? La configuración de políticas de ciclo de vida de datos también puede reducir los costos. Por ejemplo, puede automatizar la migración de datos antiguos a los que se accede a ubicaciones de almacenamiento más económicas con poca frecuencia, como Amazon Simple Storage Service Glacier.

Medición, supervisión y mejora

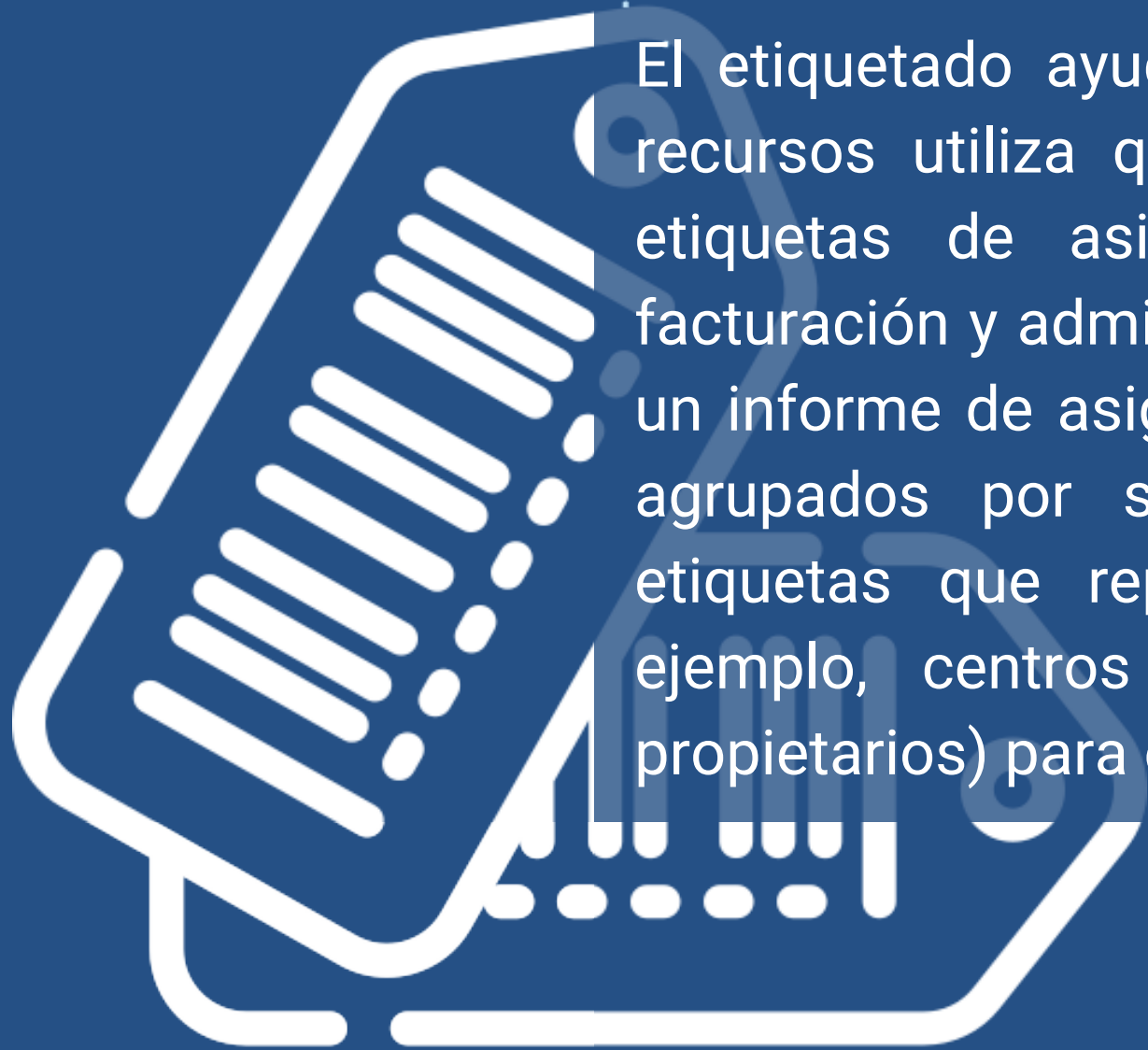


La optimización de costos es un proceso continuo.

Recomendaciones:

- Definir y aplicar el etiquetado de asignación de costos.
- Defina métricas, establezca objetivos y revíselos periódicamente.
- Aliente a los equipos a diseñar por costos.
- Asigne la responsabilidad de la optimización a un individuo o a un equipo.

Si se hace correctamente, la optimización de costos no es un proceso único que completa un cliente. En cambio, al medir y analizar sus sistemas de forma rutinaria, puede mejorar y ajustar sus costos de manera continua.



El etiquetado ayuda a proporcionar información sobre qué recursos utiliza quién y con qué propósito. Puede activar etiquetas de asignación de costos en la consola de facturación y administración de costos, y AWS puede generar un informe de asignación de costos con el uso y los costos agrupados por sus etiquetas activas. Pueden aplicarse etiquetas que representen categorías empresariales (por ejemplo, centros de costos, nombres de aplicación o propietarios) para organizar los costos entre varios servicios.

Aliente a los equipos a diseñar por costos

Explorador de costos de AWS es una herramienta que puede utilizar para ver y analizar sus costos y uso. Puede utilizar el explorador de costos para identificar las áreas que deben estudiarse más a fondo y analizar tendencias que puede utilizar para comprender los costos.



Utilice AWS como AWS Trusted Advisor, que proporciona orientación en tiempo real para ayudarlo a aprovisionar recursos que sigue las prácticas recomendadas de AWS.

Los esfuerzos de optimización de costos suelen tener más éxito cuando la responsabilidad de la optimización de costos se asigna a un individuo o a un equipo.



Algunos puntos clave de esta sección de la unidad son los siguientes:

- Los modelos de precios de Amazon EC2 incluyen instancias bajo demanda, instancias reservadas, instancias de spot, instancias dedicadas y servidores dedicados. la facturación por segundo solo está disponible para las instancias bajo demanda, las instancias reservadas y las instancias de spot que ejecutan Amazon Linux y Ubuntu.
- Las instancias de spot pueden ser interrumpidas con una notificación de 2 minutos. Sin embargo, pueden ofrecer importantes ahorros de costos en comparación con las instancias bajo demanda.

Los cuatro pilares de la optimización de costos son

Dimensionamiento
adecuado



Aumento de la
elasticidad



Modelo de precios
óptimo



Optimización de las opciones
de almacenamiento

