Pilar de la optimización de costos

Marco de Buena Arquitectura de AWS

Julio de 2020

This paper has been archived.

The latest version is now available at:

https://docs.aws.amazon.com/es_es/wellarchitected/latest/cost-optimization-pillar/welcome.html

Avisos

Los clientes son responsables de llevar a cabo su propia evaluación independiente de la información en este documento. Este documento: (a) solo tiene fines informativos, (b) representa las prácticas y las ofertas de productos de AWS actuales, las cuales están sujetas a cambios sin aviso previo, y (c) no crea compromisos ni promesas de parte de AWS y sus empresas afiliadas, proveedores o licenciantes. Los servicios o los productos de AWS se ofrecen “como son”, sin garantías, declaraciones ni condiciones de ningún tipo, ya sean expresas o implícitas. Las responsabilidades y las obligaciones de AWS frente a sus clientes se rigen por los acuerdos celebrados con AWS, y este documento no forma parte de ningún acuerdo entre AWS y sus clientes, ni lo modifica.

Índice

Resumen

Este documento técnico se centra en el pilar de la optimización de costos del Marco de Buena Arquitectura de Amazon Web Services (AWS). Ofrece asesoramiento para ayudar a los clientes a aplicar las prácticas recomendadas en el diseño, la entrega y el mantenimiento de los entornos de AWS.

Una carga de trabajo con costos optimizados utiliza todos los recursos, logra un resultado al menor precio posible y cumple los requisitos funcionales. Este documento técnico ofrece asesoramiento en profundidad para crear capacidades dentro de su organización, diseñar la carga de trabajo, seleccionar los servicios, configurar y operar los servicios, y aplicar técnicas de optimización de costos.

Introducción

El Marco de Buena Arquitectura de AWS lo ayuda a comprender las decisiones que toma a la hora de crear cargas de trabajo en AWS. El marco ofrece las prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo en la nube fiables, seguras, eficientes y rentables. Muestra una forma consistente de medir sus arquitecturas en función de las prácticas recomendadas y de identificar las áreas que admiten mejora. Creemos que tener cargas de trabajo con buena arquitectura aumenta considerablemente la probabilidad de éxito empresarial.

El marco se basa en cinco pilares:

- Excelencia operativa
- Seguridad
- Fiabilidad
- Eficiencia de rendimiento
- Optimización de costos

Este documento se centra en el pilar de la optimización de costos y en la manera de diseñar cargas de trabajo con el uso más eficaz posible de los servicios y los recursos, a fin de lograr los resultados empresariales al precio más bajo.

Aprenderá a aplicar las prácticas recomendadas del pilar de la optimización de costos dentro de su organización. La optimización de costos puede ser un desafío para las soluciones en las instalaciones tradicionales, ya que debe predecir la capacidad y las necesidades empresariales futuras mientras lidia con los complejos procesos de adquisición. La adopción de las prácticas que se describen en este documento ayudará a que su organización alcance los siguientes objetivos:

- Práctica de administración financiera en la nube
- Concientización sobre los gastos y el uso
- Recursos rentables
- Administración de los recursos de oferta y demanda
- Optimización con el paso del tiempo

Este documento está diseñado para quienes ocupan roles en tecnología y finanzas, como directores de tecnología (CTO), directores financieros (CFO), arquitectos, desarrolladores, consultores financieros, auditores financieros, analistas de negocio y miembros de equipos operativos. En este documento, no se ofrecen detalles sobre la implementación o los patrones de la arquitectura, pero se incluyen referencias para acceder a los recursos adecuados.

Optimización de costos

La optimización de costos en un proceso continuo de perfeccionamiento y mejora durante todo el ciclo de vida de una carga de trabajo. Las prácticas que se especifican en este documento lo ayudan a crear y operar cargas de trabajo para las que se tienen en cuenta los costos, las cuales logran resultados empresariales y, al mismo tiempo, minimizan los costos y permiten que su organización maximice el rendimiento de la inversión.

Principios de diseño

Tenga en cuenta los siguientes principios de diseño para la optimización de costos:

Implementar la administración financiera en la nube: para lograr el éxito financiero y acelerar la materialización del valor de negocio en la nube, debe invertir en la administración financiera en la nube. Su organización debe destinar el tiempo y los recursos necesarios para desarrollar capacidades en este nuevo ámbito de administración del uso y la tecnología. De manera similar a la capacidad de seguridad o de operaciones, necesita desarrollar capacidades mediante la incorporación de conocimientos, los programas, los recursos y los procesos que ayuden a su organización a ser rentable.

Adoptar un modelo de consumo: pague solo por los recursos informáticos que consuma y aumente o disminuya el nivel de uso en función de los requisitos empresariales. Por ejemplo, los entornos de desarrollo y prueba suelen utilizarse solo ocho horas al día durante la semana laboral. Puede detener estos recursos cuando no se estén utilizando para obtener un posible ahorro de costos del 75 % (40 horas en contraste con 168 horas).

Medir la eficiencia general: mida el resultado empresarial de la carga de trabajo y los costos asociados a la entrega. Utilice estos datos para comprender las ganancias que resultan del aumento en los resultados y la funcionalidad, y de la reducción de los costos.

Dejar de gastar dinero en tareas complicadas y no diferenciadas: AWS se encarga de las tareas complicadas que corresponden a las operaciones del centro de datos, como montar servidores en bastidores, apilarlos y proporcionarles electricidad. También elimina la carga operativa de administrar los sistemas operativos y las aplicaciones con servicios administrados. Esto le permite centrarse en los clientes y los proyectos empresariales, en lugar de en la infraestructura de TI.

Analizar y atribuir gastos: la nube permite identificar con precisión y más facilidad el costo y el uso de las cargas de trabajo, lo que a su vez permite atribuir con transparencia los costos de TI a las fuentes de ingresos y a los propietarios de las cargas de trabajo individuales. Esto ayuda a medir el rendimiento de la inversión (ROI) y ofrece a los propietarios de las cargas de trabajo la oportunidad de optimizar sus recursos y reducir los costos.

Definición

Existen cinco áreas de interés para la optimización de costos en la nube:

- Práctica de administración financiera en la nube
- Concientización sobre los gastos y el uso
- Recursos rentables
- Administración de los recursos de oferta y demanda
- Optimización con el paso del tiempo

De manera similar a los otros pilares del Marco de Buena Arquitectura, existen compensaciones que deben tenerse en cuenta para la optimización de costos. Por ejemplo, tiene que considerar si se debe optimizar la velocidad de la comercialización o los costos. En algunos casos, lo mejor es optimizar la velocidad (introducirse en el mercado rápidamente, lanzar características nuevas o cumplir un plazo) en lugar de invertir en la optimización de costos iniciales.

A veces, las decisiones de diseño se rigen por el apuro en lugar de los datos, y siempre existe la tentación de compensar en exceso en lugar de dedicar más tiempo a establecer puntos de referencia para que la implementación sea lo más rentable posible. La compensación en exceso puede generar implementaciones con demasiado aprovisionamiento y poca optimización. Sin embargo, puede ser una opción razonable si debe migrar mediante “lift and shift” los recursos de su entorno en las instalaciones hacia la nube y, luego, optimizarlos.

Invertir la cantidad adecuada de esfuerzo en una estrategia de optimización de costos desde el principio le permite obtener los beneficios económicos de la nube con mayor facilidad, ya que se asegura el cumplimiento consistente de las prácticas recomendadas y se evita el exceso de aprovisionamiento innecesario. En las siguientes secciones, se ofrecen técnicas y prácticas recomendadas para la implementación inicial y continua de la administración financiera en la nube y la optimización de costos correspondiente a sus cargas de trabajo.

Práctica de administración financiera en la nube

La administración financiera en la nube (CFM) permite a las organizaciones materializar el valor de negocio y el éxito financiero a medida que optimizan el costo y el uso, y crecen en AWS.

Estas son las prácticas recomendadas para la administración financiera en la nube:

- Propiedad funcional
- Asociación de las finanzas y la tecnología
- Presupuestos y previsiones para la nube

- Procesos con concientización sobre costos
- Cultura de concientización sobre costos
- Cuantificación del valor de negocio entregado a través de la optimización de costos

Propiedad funcional

Establecer un cargo de optimización de costos: este cargo se encarga de establecer y mantener una cultura de concientización sobre costos. Puede ser un individuo existente, un equipo dentro de su organización o un equipo nuevo de partes interesadas clave de finanzas, tecnología y organización de toda la empresa.

El cargo (individual o grupal) da prioridad al porcentaje requerido de su tiempo y lo dedica a actividades de administración y optimización de costos. En una organización pequeña, el cargo podría dedicar un porcentaje más bajo de tiempo en comparación con un cargo de tiempo completo en una empresa más grande.

El cargo requiere un enfoque multidisciplinario, con capacidades de administración de proyectos, ciencia de datos, análisis financiero y desarrollo de software o infraestructura. Puede mejorar la eficiencia de las cargas de trabajo mediante la ejecución de optimizaciones de costos (enfoque centralizado), el ejercicio de influencia en los equipos tecnológicos para aplicar optimizaciones (descentralizado) o la implementación de ambos enfoques combinados (híbrido). El cargo se puede evaluar en función de su capacidad para ejecutar y lograr los objetivos de optimización de costos (por ejemplo, métricas de la eficiencia de la carga de trabajo).

Debe conseguir el patrocinio ejecutivo para este cargo. El patrocinador se considera un defensor del consumo rentable en la nube y respalda el escalamiento del cargo para garantizar que las actividades de optimización de costos se aborden con el nivel de prioridad establecido por la organización. Juntos, el patrocinador y el cargo se aseguran de que el consumo de su organización en la nube sea eficiente y que continúe brindando valor de negocio.

Asociación de las finanzas y la tecnología

Establecer una asociación entre las finanzas y la tecnología: los equipos tecnológicos innovan más rápido en la nube porque se acortan los ciclos de aprobación, adquisición e implementación de la infraestructura. Esto puede requerir adaptación por parte de las organizaciones financieras que estaban acostumbradas a asignar costos solo después de la aprobación del proyecto y a ejecutar procesos que requieren mucho tiempo y hacen un uso intensivo de recursos para adquirir e implementar capital en centros de datos y entornos en las instalaciones.

Establezca una asociación entre las partes interesadas clave de finanzas y tecnología con el fin de poder comprender de la misma manera los objetivos organizativos y desarrollar mecanismos para tener éxito en cuando a las finanzas en el modelo de gasto variable de la informática en la nube.

Los equipos relevantes dentro de su organización deben participar en los debates sobre los costos y el uso en todas las etapas del proceso de traspaso a la nube, incluidos los siguientes:

- **Líderes en finanzas:** directores financieros; consultores financieros; auditores financieros; analistas de negocio; y encargados de la adquisición, el abastecimiento y las cuentas por pagar deben comprender el modelo de consumo en la nube, las opciones de compra y el proceso de facturación mensual. Debido a las diferencias esenciales entre la nube (como la tasa de cambio en uso, los precios de pago por uso, los precios por niveles, los modelos de precios y la información detallada de facturación y uso) y las operaciones en las instalaciones, es fundamental que el organismo de finanzas comprenda de qué manera el uso de la nube puede afectar aspectos comerciales, como los procesos de adquisición, el seguimiento de los incentivos, la asignación de costos y los estados financieros.
- **Líderes en tecnología:** los líderes en tecnología (incluidos los propietarios de productos y aplicaciones) deben conocer los requisitos financieros (por ejemplo, las restricciones presupuestarias) y los requisitos comerciales (por ejemplo, los acuerdos de nivel de servicio). Esto permite que la carga de trabajo se implemente para alcanzar los objetivos planteados por la organización.

La asociación de las finanzas y la tecnología ofrece los siguientes beneficios:

- Los equipos de finanzas y tecnología cuentan con visibilidad casi en tiempo real de los costos y el uso.
- Dichos equipos establecen un procedimiento operativo estándar para administrar la varianza del gasto en la nube.
- Las partes interesadas en las finanzas actúan como asesores estratégicos en cuanto al uso del capital para adquirir descuentos por compromiso de compra (por ejemplo, instancias reservadas o Savings Plans de AWS) y a la utilización de la nube para hacer crecer la organización.
- Las cuentas por pagar y los procesos de adquisición existentes se utilizan con la nube.
- Los equipos de finanzas y tecnología colaboran para realizar una previsión del costo y el uso futuros de AWS con el fin de alinear o crear presupuestos organizativos.
- Se logra una mejor comunicación entre las organizaciones gracias a un lenguaje compartido y a la comprensión unificada de los conceptos financieros.

Entre las partes interesantes adicionales dentro de su organización que deberían participar en los debates sobre costos y uso, se incluyen a las siguientes:

- **Propietarios de unidades de negocio:** estas personas deben comprender el modelo de negocio en la nube de manera que puedan ofrecer orientación tanto a las unidades de negocio como a toda la empresa. Este conocimiento sobre la nube es fundamental cuando se necesita realizar una previsión del crecimiento y el uso de la carga de trabajo, y cuando se evalúan las opciones de compra a largo plazo, como las instancias reservadas o los Savings Plans.
- **Terceros:** si en su organización participan terceros (por ejemplo, consultores o herramientas), asegúrese de que se encuentren en consonancia con sus objetivos financieros y de que puedan demostrar tanto la consonancia a través de los modelos de compromiso como el rendimiento de la inversión (ROI). Por lo general, los terceros contribuyen con la creación de informes y el análisis de cualquier carga de trabajo que administren, y proporcionan un análisis de costos de cualquier carga de trabajo que diseñen.

Presupuestos y previsiones para la nube

Establecer presupuestos y previsiones para la nube: los clientes utilizan la nube para conseguir eficiencia, velocidad y agilidad, lo que origina una cantidad muy variable de costo y uso. Los costos pueden disminuir cuando aumenta la eficiencia de la carga de trabajo o a medida que se implementan cargas de trabajo y características nuevas. O bien, las cargas de trabajo se ampliarán para atender a más clientes, lo que aumenta los niveles de uso y costos en la nube. Los procesos existentes de creación de presupuestos organizativos deben modificarse para incorporar esta variabilidad.

Ajustar los procesos existentes de creación de presupuestos y previsiones para que sean más dinámicos mediante un algoritmo basado en tendencias (con costos históricos como información de entrada), algoritmos basados en los factores de impulso del negocio (por ejemplo, lanzamiento de productos nuevos o expansión regional) o una combinación de tendencias y factores de impulso del negocio.

Puede utilizar AWS Cost Explorer para realizar previsiones sobre los costos en la nube a diario (hasta 3 meses) o mensualmente (hasta 12 meses) en función de los algoritmos de aprendizaje automático aplicados a los costos históricos (basados en tendencias).

Procesos con concientización sobre costos

Implementar la concientización sobre costos en sus procesos organizativos: la concientización sobre costos debe implementarse en los procesos organizativos nuevos y existentes. Se recomienda volver a utilizar y modificar los procesos existentes si es posible; esto minimiza el impacto sobre la agilidad y la velocidad. Las siguientes recomendaciones lo ayudarán a implementar la concientización sobre costos en la carga de trabajo:

- Asegúrese de que la administración de cambios incluya una medición de costos para cuantificar el impacto financiero de sus cambios. Esto ayuda a abordar de manera proactiva las preocupaciones relacionadas con los costos y a destacar el ahorro de costos.
- Asegúrese de que la optimización de costos sea un componente central de sus capacidades operativas. Por ejemplo, puede aprovechar los procesos existentes de administración de incidentes para investigar e identificar la causa raíz de las anomalías en el costo y el uso (excedentes de costos).
- Acelere la materialización del ahorro de costos y el valor de negocio a través de la automatización o la implementación de herramientas. A la hora de pensar en el costo de implementación, plantee la conversación de forma que se incluya un componente del ROI para justificar la inversión de tiempo o dinero.
- Amplíe los programas de formación técnica y desarrollo actuales para incluir formación técnica que tenga en cuenta los costos en toda su organización. Se recomienda que esto incluya formación técnica y certificación continuas. Con esto, se creará una organización capaz de administrar el costo y el uso por su cuenta.

Informar y notificar sobre la optimización de los costos y el uso: debe brindar información con regularidad sobre la optimización de los costos y el uso dentro de su organización. Puede implementar sesiones exclusivas para la optimización de costos o incluir la optimización de costos en sus ciclos regulares de creación de informes operativos sobre las cargas de trabajo. AWS Cost Explorer brinda paneles e informes. Puede realizar un seguimiento de su progreso respecto del costo y el uso en comparación con los presupuestos configurados con los informes de Presupuestos de AWS.

También puede usar Amazon QuickSight con los datos del Informe de uso y costo (CUR) para proporcionar informes con un alto nivel de personalización y con más detalles.

Implemente notificaciones sobre el costo y el uso para garantizar que se pueda reaccionar ante los cambios en estos factores con rapidez. Presupuestos de AWS le permite emitir notificaciones respecto de los objetivos. Recomendamos configurar las notificaciones tanto en los aumentos como en las disminuciones y tanto en el costo como en el uso de las cargas de trabajo.

Monitorear el costo y el uso de manera proactiva: se recomienda monitorear el costo y el uso de manera proactiva dentro de su organización y no solo cuando existen excepciones o anomalías. Los paneles con mucha visibilidad en toda su oficina o entorno de trabajo garantizan que las personas clave tengan acceso a la información que necesitan e indican el enfoque de la organización en la optimización de costos. Los paneles visibles le permiten promover de manera activa los resultados exitosos y también implementarlos en toda su organización.

Cultura de concientización sobre costos

Crear una cultura de concientización sobre costos: implemente cambios o programas en toda su organización para generar una cultura de concientización sobre costos. Se recomienda comenzar de a poco. Luego, a medida que aumenten sus capacidades y el uso de la nube en su organización, implemente programas grandes y de alcance amplio.

La cultura de concientización sobre costos le permite escalar la optimización de costos y la administración financiera en la nube a través de las prácticas recomendadas que se aplican de manera orgánica y descentralizada en toda su organización. Con esto, se crean niveles altos de capacidad en toda su organización con un esfuerzo mínimo, en comparación con lo que se logra con un enfoque estricto, descendente y centralizado.

Los pequeños cambios en la cultura pueden tener un gran impacto en la eficiencia de sus cargas de trabajo actuales y futuras. Algunos ejemplos de esto son los siguientes:

- Ludificación del costo y el uso en su organización. Se puede lograr a través de un panel con visibilidad pública o de un informe que compare los costos y el uso normalizados entre los equipos (por ejemplo, costo por carga de trabajo, costo por transacción).
- Reconocimiento de la rentabilidad. Recompense los logros voluntarios o no solicitados en materia de optimización de costos de forma pública o privada, y aprenda de los errores para evitar repetirlos en el futuro.
- Cree requisitos organizativos descendentes para que las cargas de trabajo se ejecuten de acuerdo con presupuestos definidos previamente.

Mantenerse actualizado con los nuevos lanzamientos de servicios: es posible que pueda implementar nuevos servicios y características de AWS para aumentar la rentabilidad en su carga de trabajo. Revise con regularidad el Blog de noticias de AWS, el Blog de administración de costos de AWS y las Novedades de AWS para obtener información sobre los nuevos lanzamientos de servicios y características.

Cuantificación del valor de negocio entregado a través de la optimización de costos

Cuantificar el valor de negocio correspondiente a la optimización de costos: además de brindar información sobre los ahorros de la optimización de costos, se recomienda que cuantifique el valor adicional entregado. Los beneficios de la optimización de costos generalmente se cuantifican en términos de costo más bajo por resultado empresarial. Por ejemplo, puede calcular los ahorros en costos correspondientes a las instancias bajo demanda de Amazon Elastic Compute Cloud (Amazon EC2) cuando adquiere Savings Plans, los cuales reducen los costos y mantienen los niveles de salida de la carga de trabajo. Puede cuantificar las reducciones de costos

en el gasto de AWS cuando se terminan las instancias inactivas de Amazon EC2 o se eliminan los volúmenes no vinculados de Amazon Elastic Block Store (Amazon EBS).

Cuantificar el valor de negocio correspondiente a la optimización de costos le permite comprender el conjunto completo de beneficios para su organización. Como la optimización de costos es una inversión necesaria, cuantificar el valor de negocio le permite explicar el rendimiento de la inversión a las partes interesadas. Cuantificar el valor de negocio puede ayudarlo a obtener más aceptación de las partes interesadas en futuras inversiones para la optimización de costos. También brinda un marco para medir los resultados de las actividades de optimización de costos de su organización.

Sin embargo, los beneficios de la optimización de costos superan la simple ventaja de reducir o evitar costos. Considere registrar datos adicionales para medir las mejoras en la eficiencia y el valor de negocio. Algunos ejemplos de mejoras son los siguientes:

- **Ejecución de las prácticas recomendadas para la optimización de costos:** por ejemplo, la administración del ciclo de vida de los recursos reduce los costos operativos y de infraestructura, y también libera tiempo y genera un presupuesto imprevisto para la experimentación. Como resultado, se mejoran los niveles de agilidad en la organización y se descubren nuevas oportunidades para la generación de ingresos.
- **Implementación de la automatización:** por ejemplo, tenga en cuenta Auto Scaling que garantiza la elasticidad con un esfuerzo mínimo y aumenta la productividad del personal, ya que elimina el trabajo de planificación de la capacidad manual. Consulte el documento técnico sobre el pilar de la fiabilidad del Marco de Buena Arquitectura para obtener más información acerca de la resiliencia operativa.
- **Previsión de los costos futuros de AWS:** la previsión permite a las partes interesadas de las finanzas establecer expectativas con otras partes interesadas internas y externas de la organización. También ayuda a mejorar la previsibilidad financiera de la organización. AWS Cost Explorer se puede utilizar para generar previsiones sobre el costo y el uso.

Recursos

Consulte los siguientes recursos a fin de obtener más información sobre las prácticas recomendadas de AWS para generar presupuestos y previsiones del gasto en la nube.

- Generación de informes de métricas de presupuesto con informes de presupuesto
- Previsión con AWS Cost Explorer
- AWS Training
- AWS Certification
- AWS Cloud Management Tools Partners

Concientización sobre los gastos y el uso

Comprender los costos y los factores de impulso de su organización es fundamental para administrar el costo y el uso de manera eficaz. También es esencial para identificar oportunidades de reducción de costos. Las organizaciones generalmente operan varias cargas de trabajo, las cuales están bajo la dirección de múltiples equipos. Estos equipos pueden encontrarse en diferentes unidades organizativas, cada una con su propia fuente de ingresos. La capacidad de atribuir costos de recursos a las cargas de trabajo, la organización individual o los propietarios de productos fomenta un comportamiento de uso eficiente y ayuda a reducir el desperdicio. El monitoreo preciso de los costos y el uso le permite comprender cuán rentables son las unidades y los productos de la organización, además de permitirle tomar decisiones con más fundamentos sobre dónde asignar recursos dentro de su organización. La concientización sobre el uso en todos los niveles de la organización es clave para impulsar el cambio, ya que los cambios en el uso generan cambios en el costo.

Considere adoptar un enfoque multifacético para conocer el uso y los gastos. Su equipo debe recopilar datos, analizarlos y luego generar un informe sobre ellos. Los factores clave que se deben considerar son los siguientes:

- Gobernanza
- Monitoreo de los costos y el uso
- Retiro

Gobernanza

Para administrar sus costos en la nube, debe administrar el uso a través de las siguientes áreas de gobernanza:

Desarrollo de políticas organizativas: el primer paso para ejercer la gobernanza es utilizar los requisitos de su organización a fin de desarrollar políticas para el uso de la nube. Estas políticas definen la manera en que su organización utiliza la nube y la forma en que se administran los recursos. Las políticas deben abordar todos los aspectos relativos a los recursos y las cargas de trabajo que se relacionan con el costo o el uso, incluida la creación, la modificación y el retiro durante la vida útil del recurso.

Las políticas deben ser simples para que se comprendan fácilmente y puedan implementarse de manera efectiva en toda la organización. Comience con políticas amplias de alto nivel, como cuál es el uso permitido de la región geográfica o las horas del día en que los recursos deberían estar en ejecución. Ajuste gradualmente las políticas para las diversas unidades organizativas y cargas de trabajo. Las políticas comunes incluyen qué servicios y características se pueden usar (por ejemplo, almacenamiento de menor rendimiento en entornos de prueba o desarrollo), y qué

tipos de recursos pueden usar diferentes grupos (por ejemplo, el mayor tamaño de recurso en una cuenta de desarrollo es mediano).

Desarrollo de metas y objetivos: defina metas y objetivos sobre costo y uso para su organización. Las metas brindan orientación y dirección a la organización sobre los resultados esperados. Los objetivos especifican resultados que se pueden medir y se deben alcanzar. Un ejemplo de una meta es que el uso de la plataforma debería aumentar significativamente con solo un incremento reducido (no lineal) en el costo. Un ejemplo de objetivo es un aumento del 20 % en el uso de la plataforma, acompañado por un incremento de los costos de menos del 5 %. Otra meta común es que las cargas de trabajo deben ser más eficientes cada 6 meses. El objetivo que la complementa sería que el costo por salida de la carga de trabajo debe disminuir en un 5 % cada 6 meses.

Una meta común para las cargas de trabajo en la nube es aumentar la eficiencia de las cargas, lo que significa disminuir el costo por resultado empresarial de la carga de trabajo a lo largo del tiempo. Se recomienda implementar esta meta para todas las cargas de trabajo y también establecer un objetivo, como un aumento del 5 % en la eficiencia cada 6 a 12 meses. Esto se puede lograr en la nube mediante la creación de capacidad en optimización de costos y mediante el lanzamiento de nuevos servicios y características de servicios.

Estructura de la cuenta: AWS tiene una estructura de cuentas basada en la noción de un elemento primario para muchos elementos secundarios, los cuales se conocen en general como cuenta maestra (el elemento primario, antes pagador) y cuenta miembro (el elemento secundario, antes vinculado). Una práctica recomendada es tener siempre al menos una cuenta maestra con una cuenta miembro, independientemente del tamaño o del uso de su organización. Todos los recursos de la carga de trabajo deben encontrarse solo en las cuentas miembros.

No hay una respuesta universal para saber cuántas cuentas de AWS debería tener. Evalúe sus modelos operativos y de costos actuales y futuros para asegurarse de que la estructura de sus cuentas de AWS refleje las metas de la organización. Algunas empresas crean múltiples cuentas de AWS por razones comerciales, por ejemplo:

- Se requiere aislamiento administrativo, fiscal o de facturación entre las unidades organizativas, los centros de costos o las cargas de trabajo específicas.
- Los límites de servicio de AWS están configurados específicamente para determinadas cargas de trabajo.
- Existe un requisito de aislamiento y separación entre las cargas de trabajo y los recursos.

Dentro de AWS Organizations, la facturación unificada crea el vínculo entre una o más cuentas miembros y la cuenta maestra. Las cuentas miembros le permiten aislar y distinguir el costo y el uso por grupos. Una práctica común es tener cuentas miembros independientes para cada unidad organizativa (como finanzas, marketing y ventas), para cada ciclo de vida del entorno (como desarrollo, pruebas y producción) o para cada carga de trabajo (carga de trabajo a, b y c), y luego agrupar estas cuentas vinculadas con la facturación unificada.

La facturación unificada permite consolidar el pago de varias cuentas miembros de AWS en una sola cuenta maestra, al mismo tiempo que se proporciona visibilidad de la actividad a cada cuenta vinculada. Como los costos y el uso se agrupan en la cuenta maestra, esto le permite maximizar los descuentos por volumen de servicio y el uso de sus descuentos por compromiso (Savings Plans e instancias reservadas) para conseguir los descuentos más altos.

AWS Control Tower puede configurar rápidamente varias cuentas de AWS, lo que garantiza que la gobernanza esté alineada con los requisitos de su organización.

Grupos y roles organizativos: después de desarrollar las políticas, puede crear grupos y roles lógicos de usuarios dentro de la organización. Esto le permite asignar permisos y controlar el uso. Comience con grupos de personas de alto nivel; en general, esto se alinea con las unidades organizativas y los puestos de trabajo (por ejemplo, el administrador de sistemas en el Departamento de TI o el auditor financiero). Los grupos juntan a personas que realizan tareas similares y necesitan un acceso parecido. Los roles definen lo que debe hacer un grupo. Por ejemplo, un administrador de sistemas en TI requiere acceso para crear todos los recursos, pero un miembro del equipo de analítica solo necesita crear recursos analíticos.

Controles y notificaciones: un primer paso común en la implementación de controles de costos es configurar notificaciones cuando ocurren eventos relacionados con el costo o el uso fuera de las políticas. Esto le permite actuar rápidamente y verificar si se necesitan acciones correctivas sin restringir ni afectar de forma negativa las cargas de trabajo o la nueva actividad. Después de descubrir los límites del entorno y la carga de trabajo, puede hacer cumplir la gobernanza. En AWS, las notificaciones se envían con Presupuestos de AWS, que permite definir un presupuesto mensual para los costos, el uso y los descuentos por compromiso de AWS (Savings Plans e instancias reservadas). Puede crear presupuestos en un nivel de costo agrupado (por ejemplo, todos los costos) o en un nivel más pormenorizado donde incluya solo dimensiones específicas, como cuentas vinculadas, servicios, etiquetas o zonas de disponibilidad. También puede asociar notificaciones de email a sus presupuestos, las cuales se activarán cuando los costos o el uso actuales o previstos excedan un porcentaje límite ya establecido.

Controles y cumplimiento: como segundo paso, puede aplicar políticas de gobernanza en AWS a través de AWS Identity and Access Management (IAM) y políticas de control de servicios (SCP) de AWS Organizations. IAM le permite administrar de forma segura el acceso a los servicios y los recursos de AWS. Con IAM, puede controlar quién tiene permiso para crear y administrar los recursos de AWS, el tipo de recursos que se pueden crear y la ubicación en la que se crean. En consecuencia, se minimiza la creación de recursos que no son necesarios. Use los roles y los grupos creados previamente, y asígneles políticas de IAM para exigir el uso correcto. Las SCP ofrecen un control central sobre el número máximo de permisos disponibles para todas las cuentas en su organización, lo que garantiza que las cuentas respeten sus pautas de control de acceso. Las SCP están disponibles solo en una organización que tiene todas las características habilitadas, y usted puede configurar las SCP para denegar o permitir acciones a las cuentas miembros de forma predeterminada. Consulte el documento técnico sobre el pilar de la seguridad del Marco de Buena Arquitectura para obtener más información sobre cómo implementar la administración del acceso.

Controles y Service Quotas: la gobernanza puede implementarse también a través de la administración de Service Quotas. Si se asegura de que Service Quotas se fije con los costos generales mínimos y se mantenga de forma precisa, puede minimizar la creación de recursos fuera de lo establecido por los requisitos de la organización. Para lograr esto, debe entender la velocidad con la que pueden cambiar los requisitos y la velocidad con la que pueden implementarse los cambios de las cuotas, además de comprender los proyectos en proceso (la creación y el retiro de recursos). Service Quotas puede utilizarse para aumentar sus cuotas cuando sea necesario.

Los servicios de administración de costos de AWS se integran con el servicio AWS Identity and Access Management (IAM). Utilice el servicio IAM junto con los servicios de administración de costos para controlar el acceso a sus datos financieros y a las herramientas de AWS en la consola de facturación.

Seguimiento del ciclo de vida de las cargas de trabajo: asegúrese de realizar un seguimiento del ciclo de vida completo de la carga de trabajo. Esto garantiza que cuando las cargas de trabajo o sus componentes ya no sean necesarios se podrán modificar o retirar. Esto es sumamente útil cuando lanza nuevos servicios o características. Es posible que las cargas de trabajo y sus componentes figuren como en uso. Sin embargo, estos deben retirarse para redirigir a los clientes hacia nuevos servicios. Tenga en cuenta las etapas anteriores de las cargas de trabajo. Cuando una carga de trabajo se encuentra en la etapa de producción, los entornos anteriores pueden retirarse o reducirse considerablemente en cuanto a la capacidad hasta que se consideren necesarios otra vez.

AWS ofrece una serie de servicios de administración y gobernanza que puede utilizar para realizar el seguimiento del ciclo de vida de las entidades. Puede utilizar AWS Config o AWS Systems Manager para proporcionar un inventario detallado de los recursos y las configuraciones de AWS. Se recomienda la integración con los sistemas de administración de proyectos o recursos existentes para realizar un seguimiento de los proyectos y los productos activos dentro de su organización. La combinación de su sistema actual con el conjunto abundante de eventos y métricas que AWS brinda le permite generar una visualización de los eventos significantes del ciclo de vida y administrar los recursos de manera proactiva a fin de reducir la cantidad de costos innecesarios.

Consulte el documento técnico sobre el pilar de la excelencia operativa del Marco de Buena Arquitectura para obtener más información sobre cómo implementar el seguimiento del ciclo de vida de las entidades.

Monitoreo de los costos y el uso

Permita a los equipos tomar las medidas necesarias respecto de los costos y el uso a través de la visualización detallada de la carga de trabajo. La optimización de costos comienza con el conocimiento detallado del desglose del uso y los costos; la capacidad de modelar y predecir gastos, usos y características futuras; además de la implementación de los mecanismos necesarios para alinear los costos y el uso con los objetivos de la organización. A continuación, se detallan las áreas necesarias para monitorear los costos y el uso:

Configurar orígenes de datos detallados: habilite el nivel de detalle por hora en Cost Explorer y cree un Informe de uso y costo (CUR). Estos orígenes de datos ofrecen la visualización más precisa del uso y los costos de toda su organización. El CUR proporciona detalles sobre el uso con nivel de detalle por día u hora, tasas, costos y atributos de uso sobre todos los servicios de AWS que se pueden cobrar. Este informe incluye todas las dimensiones posibles, como el etiquetado, la ubicación, los atributos de los recursos y los ID de las cuentas.

Configure su CUR con las siguientes opciones de personalización:

- Incluir los ID de los recursos
- Actualizar de manera automática el CUR
- Brindar información con nivel de detalle por hora
- Controlar las versiones: sobrescribir el informe existente
- Integrar los datos: Athena (formato Parquet y compresión)

Utilice AWS Glue a fin de preparar los datos para su análisis correspondiente. Por otro lado, utilice Amazon Athena con el objetivo de ejecutar los análisis de datos con SQL para consultar los datos. Además, puede utilizar Amazon QuickSight para crear visualizaciones complejas y personalizadas, y luego distribuirlas a toda la organización.

Identificar las categorías de atribución de los costos: trabaje junto con su equipo de finanzas y otras partes interesadas relevantes para comprender los requisitos de atribución de los costos dentro de su organización. Los costos que implican las cargas de trabajo deben distribuirse entre todas las etapas el ciclo de vida, incluido el desarrollo, la prueba, la producción y el retiro. Comprenda cómo los costos generados para el aprendizaje, el desarrollo del personal y la creación de ideas se atribuyen en la organización. Esto puede resultarle útil para asignar de forma correcta las cuentas que se utilizan para este fin a los presupuestos de formación técnica y desarrollo, en lugar de a los presupuestos genéricos de costos de TI.

Establecer métricas para las cargas de trabajo: comprenda cómo se miden los resultados de su carga de trabajo según el éxito de la empresa. Generalmente, cada carga de trabajo cuenta con un pequeño conjunto de resultados principales que indican el rendimiento. Si tiene una carga de trabajo compleja con muchos componentes, puede priorizar la lista o definir las métricas para cada componente y realizar un seguimiento de estas. Trabaje junto con sus equipos para identificar las métricas que debe utilizar. Esta unidad se utilizará para comprender la eficiencia de la carga de trabajo o los costos de cada resultado empresarial.

Atribuir significado al uso y los costos en el marco de la organización: implemente etiquetas en AWS para agregar la información de la organización a los recursos, la cual se sumará posteriormente a la información sobre uso y costos. Las etiquetas son pares de clave-valor. La clave está definida y debe ser única para toda la organización. Por otro lado, el valor es único para un grupo de recursos. La clave “Entorno” con el valor “Producción” es un claro ejemplo de un par de clave-

valor. Todos los recursos en el entorno de producción tendrán este par de clave-valor. El etiquetado le permite categorizar y realizar un seguimiento de sus costos con información relevante y significativa de la organización. Puede asignar etiquetas que representen categorías de organización (por ejemplo, centros de costos, nombres de aplicaciones, proyectos o propietarios) e identificar las cargas de trabajo y las características de estas (por ejemplo, prueba o producción) para distribuir los costos y el uso entre toda la organización.

Si asigna etiquetas a los recursos de AWS (por ejemplo, las instancias EC2 o los buckets de Amazon S3) y las activa, AWS agrega esta información a sus Informes de uso y costo. Puede ejecutar informes y análisis sobre los recursos con y sin etiquetas para permitir que se cumplan las políticas de administración de costos internas más plenamente, además de garantizar que la atribución de costos sea precisa.

Crear e implementar un estándar de etiquetado de AWS en todas las cuentas de su organización le permite administrar y controlar los entornos de AWS de forma coherente y uniforme. Utilice las políticas de etiquetado en AWS Organizations para definir las reglas de uso de las etiquetas sobre los recursos de AWS en sus cuentas de AWS Organizations. Las políticas de etiquetado le permiten adoptar con facilidad una estrategia estandarizada para el etiquetado de los recursos de AWS.

AWS Tag Editor le permite agregar, eliminar y administrar las etiquetas de varios recursos.

AWS Cost Categories le permite atribuir significado a los costos en el marco de la organización sin la necesidad de etiquetar los recursos. Puede asignar su información de costos y uso a las estructuras internas y únicas de la organización. Además, puede definir las reglas de las categorías para asignar y categorizar los costos a través de dimensiones de facturación, como cuentas y etiquetas. Esto brinda otro nivel de capacidad de administración, además del etiquetado. Puede también asignar cuentas y etiquetas específicas a varios proyectos.

Configurar las herramientas de facturación y optimización de costos: para modificar el uso y ajustar los costos, cada miembro de su organización debe tener acceso a su información de costos y uso. Se recomienda que todos los equipos y las cargas de trabajo cuenten con las siguientes herramientas configuradas cuando utilicen la nube:

- **Informes:** resuma toda la información relacionada con los costos y el uso.
- **Notificaciones:** emita notificaciones cuando los costos o el uso superen los límites definidos.
- **Estado actual:** configure un panel que muestre los niveles actuales de costos y uso. Los paneles deben estar disponibles en lugares sumamente visibles dentro del entorno de trabajo (similar a un panel de operaciones).
- **Tendencias:** brinde la capacidad de mostrar la variabilidad en los costos y el uso durante el periodo y con el nivel de detalle necesarios.
- **Previsiones:** proporcione la capacidad de mostrar los costos futuros estimados.

- **Seguimiento:** muestre el uso y los costos actuales en comparación con los objetivos o las metas que se han configurado.
- **Análisis:** brinde a los miembros del equipo la capacidad de ejecutar análisis profundos y personalizados con el nivel de detalle por hora inclusive y todas las dimensiones posibles.

Puede utilizar las herramientas nativas de AWS, como AWS Cost Explorer, Presupuestos de AWS y Amazon Athena con QuickSight, para proporcionar esta capacidad. Además, puede utilizar herramientas de terceros. Sin embargo, debe asegurarse de que los costos de estas herramientas aporten valor a su organización.

Asignar costos en función de las métricas de las cargas de trabajo: la optimización de costos implica entregar resultados empresariales al precio más bajo, el cual solo puede alcanzarse a través de la asignación de los costos de las cargas de trabajo de acuerdo con las métricas correspondientes (se miden según la eficiencia de las cargas de trabajo). Monitoree las métricas definidas de las cargas de trabajo a través de los archivos de registro o cualquier otro forma de monitoreo de aplicaciones. Combine estos datos con los costos de las cargas de trabajo, los cuales se obtienen observando los costos que contengan valores de etiquetas o ID de cuentas específicos. Se recomienda ejecutar este análisis a cada hora. En general, su eficiencia cambiará si dispone de algunos componentes con costo estático (por ejemplo, una base de datos de backend que se ejecuta las 24 horas del día, los 7 días de la semana) con una tasa de solicitudes que varía (por ejemplo, el uso alcanza su máximo nivel entre las 9:00 h y las 17:00 h, con pocas solicitudes durante la noche). Comprender cómo funciona la relación entre los costos variables y estáticos lo ayudará a enfocarse en las actividades de optimización.

Retiro de recursos

Después de administrar una lista de proyectos, empleados y recursos tecnológicos a lo largo del tiempo, podrá identificar qué recursos ya no se utilizan y qué proyectos ya no tienen un propietario.

Realizar un seguimiento de los recursos durante su vida útil: retire los recursos de las cargas de trabajo que ya no sean necesarios. Los recursos que se utilizan para ejecutar pruebas son un claro ejemplo de componentes que se pueden sacar de circulación. Una vez que se completan las pruebas, los recursos pueden descartarse. Realizar seguimientos de los recursos a través de etiquetas (y ejecutar informes sobre esas etiquetas) lo ayudará a identificar los recursos que deben retirarse. El uso de etiquetas es una forma efectiva de realizar un seguimiento de los recursos a través de la identificación del recurso con su función o una fecha conocida para su retiro. Los informes pueden ejecutarse sobre estas etiquetas. Los valores de ejemplo para el etiquetado de características son “featureX testing”. De esta forma, se podrá identificar el propósito del recurso en relación con el ciclo de vida de la carga de trabajo.

Implementar un proceso de retiro: implemente un proceso estandarizado en toda la organización para identificar y eliminar los recursos que no se utilizan. Este proceso debe

definir la frecuencia con la que se realizan las búsquedas y los procesos necesarios para eliminar el recurso y garantizar que se cumplan todos los requisitos de la organización.

Retirar recursos: la frecuencia y el esfuerzo de búsqueda de recursos que no se utilizan debe reflejarse en los ahorros potenciales. De esta forma, una cuenta con costos bajos debe analizarse con menor frecuencia que una cuenta con costos más altos. Los eventos de búsqueda y retiro pueden activarse debido a los cambios en el estado de la carga de trabajo, por ejemplo, un producto cuya vida útil está próxima a finalizar o un producto que está a punto de ser reemplazado. Además, es posible que estos eventos de búsqueda y retiro se activen por eventos externos, como los cambios en las condiciones del mercado o la discontinuación de un producto.

Retirar recursos de manera automática: utilice la automatización para reducir o eliminar los costos asociados al proceso de retiro. Diseñar una carga de trabajo para que ejecute un proceso de retiro automatizado reducirá los costos generales que implica dicha carga durante su vida útil. Puede utilizar AWS Auto Scaling para ejecutar el proceso de retiro. También puede implementar código personalizado a través de las API o los SDK para retirar automáticamente los recursos de las cargas de trabajo.

Recursos

Para obtener más información acerca de las prácticas recomendadas de AWS para la concientización sobre los gastos, consulte los siguientes recursos.

- Estrategias de etiquetado de AWS
- Activación de etiquetas de asignación de costos definidas por el usuario
- Administración de costos y facturación de AWS
- Blog de administración de costos
- Estrategia de facturación para varias cuentas
- Herramientas y SDK de AWS
- Prácticas recomendadas para el etiquetado
- Well-Architected Labs - Cost Fundamentals
- Well-Architected Labs – Expenditure Awareness

Recursos rentables

El uso de las configuraciones, los servicios y los recursos adecuados para sus cargas de trabajo es fundamental para el ahorro de costos. A la hora de crear recursos rentables, tenga en cuenta lo siguiente:

- Evalúe los costos a la hora de seleccionar los servicios
- Seleccione la cantidad, el tipo y el tamaño correctos de los recursos
- Seleccione el mejor modelo de precios
- Planifique la transferencia de datos

Para conseguir ayuda en la elección de una arquitectura de acuerdo con los conocimientos adquiridos, puede recurrir a los arquitectos de soluciones de AWS, las soluciones de AWS, las arquitecturas de referencia de AWS y los socios de APN.

Evaluación de los costos a la hora de seleccionar servicios

Identificar los requisitos de la organización: cuando seleccione los servicios para su carga de trabajo, es fundamental que conozca las prioridades de la organización. Asegúrese de que existe equilibrio entre los costos y los demás pilares de Buena Arquitectura, como el rendimiento y la fiabilidad. Una carga de trabajo con costos completamente optimizados es la solución que más se alinea a los requisitos de la organización. Sin embargo, no se trata necesariamente de la solución con el costo más bajo. Reúnase con todos los equipos de su organización para recopilar información, como información sobre los productos, la empresa, el sector técnico y las finanzas.

Analizar todos los componentes de la carga de trabajo: ejecute un análisis exhaustivo de todos los componentes en su carga de trabajo. Asegúrese de que existe equilibrio entre el costo de los análisis y los ahorros potenciales en la carga de trabajo durante su ciclo de vida. Debe averiguar el impacto actual y el posible impacto futuro del componente. Por ejemplo, si el costo del recurso propuesto es de USD 10 por mes y las cargas con previsiones inferiores no superan los USD 15 por mes, destinar un día de esfuerzo para reducir los costos en un 50 % (USD 5 por mes) podría superar el posible beneficio durante la vida del sistema. Utilizar una estimación más eficiente y rápida que se basa en datos creará el mejor resultado general para este componente.

Las cargas de trabajo pueden cambiar a lo largo del tiempo, por lo que es posible que el conjunto adecuado de servicios no sea óptimo si la arquitectura o el uso de la carga de trabajo cambian. El análisis para la selección de los servicios debe incorporar los estados y los niveles de uso actuales y futuros de la carga de trabajo. La implementación de un servicio para el estado o uso futuros de la carga de trabajo puede reducir los costos generales al reducir o eliminar el esfuerzo que se requiere para efectuar cambios en el futuro.

AWS Cost Explorer y el CUR pueden analizar el costo de una prueba de concepto (PoC) o de un entorno en ejecución. También puede utilizar la calculadora de costo mensual de AWS o la calculadora de precios de AWS para estimar los costos de la carga de trabajo.

Considere el ahorro de tiempo que le permitirá a su equipo enfocarse en redimir deudas técnicas, la innovación y las características que agreguen valor. Por ejemplo, es posible que necesite migrar directamente su entorno en las instalaciones a la nube tan rápido como sea

posible y optimizarlo luego. Vale la pena analizar los ahorros que podría obtener si usara servicios administrados que eliminen o reduzcan los costos de licencia.

Generalmente, los servicios administrados poseen atributos que se pueden configurar para garantizar una capacidad suficiente. Debe configurar y monitorear estos atributos de manera que el exceso de capacidad se mantenga en el mínimo y se maximice el rendimiento. Puede modificar los atributos de AWS Managed Services mediante el uso de la consola de administración de AWS o las API y los SDK de AWS a fin de alinear las necesidades de los recursos con la demanda cambiante. Por ejemplo, puede aumentar o disminuir la cantidad de nodos en un clúster de Amazon EMR (o un clúster de Amazon Redshift) para que aumente o disminuya la escala.

También puede incluir varias instancias en un recurso de AWS para permitir un uso de mayor densidad. Por ejemplo, puede aprovisionar varias bases de datos pequeñas en una única instancia de base de datos de Amazon Relational Database Service (Amazon RDS). A medida que el uso crezca, puede migrar una de las bases de datos a una instancia de base de datos de RDS dedicada mediante el uso de una instantánea y un proceso de recuperación.

Cuando se aprovisionan las cargas de trabajo en los servicios administrados, debe comprender los requisitos para ajustar la capacidad del servicio. Por lo general, estos requisitos son tiempo, esfuerzo y cualquier impacto para lograr un normal funcionamiento de la carga de trabajo. El recurso aprovisionado debe dar tiempo para que se puedan efectuar cambios y aprovisionar los gastos generales necesarios para permitirlos. Se puede reducir el esfuerzo continuo necesario para modificar los servicios prácticamente a cero usando las API y los SDK que vienen integrados a las herramientas de sistemas y de monitoreo, tal como Amazon CloudWatch.

Amazon Relational Database Service (RDS), Amazon Redshift y Amazon ElastiCache proporcionan un servicio de base de datos administrado. Amazon Athena, Amazon Elastic Map Reduce (EMR) y Amazon Elasticsearch proporcionan un servicio de análisis administrado.

AWS Managed Services (AMS) es un servicio que opera la infraestructura de AWS en nombre de los clientes y socios empresariales. Proporciona un entorno seguro y conforme a los requisitos establecidos, en el cual se pueden implementar las cargas de trabajo. AMS utiliza modelos empresariales de operación en la nube que ofrecen automatización que le permite cumplir con los requisitos de su organización, trasladarse a la nube con mayor rapidez y reducir sus costos de administración actuales.

Servicios sin servidor o de nivel de aplicación: puede usar servicios sin servidor o de nivel de aplicación como AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS) y Amazon Simple Email Service (Amazon SES). Estos servicios eliminan la necesidad de tener que administrar un recurso y proporcionan las funciones de ejecución de código, servicios de cola y entrega de mensajes. El otro beneficio que ofrecen estos servicios es que adaptan su rendimiento y costos en función del uso, lo que permite la distribución y asignación eficientes de los costos.

Para obtener más información acerca de los recursos sin servidor, consulte el documento técnico sobre el enfoque de aplicaciones sin servidor del Marco de buena arquitectura de AWS.

Análisis de la carga de trabajo para diferentes usos con el tiempo: como AWS lanza nuevos servicios y características con el tiempo, es posible que los servicios para su carga de trabajo cambien. El esfuerzo necesario debe reflejar los potenciales beneficios. La frecuencia de las revisiones de la carga de trabajo depende de los requisitos de su organización. Si se trata de una carga de trabajo de un costo considerable, la temprana implementación de nuevos servicios maximizará los ahorros en los costos, por lo que realizar revisiones con mayor frecuencia puede tener ventajas. Otro desencadenador para las revisiones es el cambio en los patrones de uso. Los grandes cambios en el uso pueden indicar que los servicios alternativos serían más adecuados. Por ejemplo, para índices de transferencia de datos más altos, un servicio de conexión directa puede ser más económico que una VPN y proporcionar la conectividad necesaria. Estime el potencial impacto de los cambios de servicios, de manera que pueda monitorear estos desencadenadores de nivel de uso e implementar los servicios más rentables con mayor prontitud.

Costos de licencia: se puede eliminar el costo de las licencias de software a través del uso de software de código abierto. Esto puede tener un gran impacto en los costos de la carga de trabajo a medida que su tamaño aumenta. Evalúe los beneficios del software con licencia en comparación con el costo total para asegurarse de que tiene la carga de trabajo más optimizada. Modele cualquier cambio en las licencias y el modo en que afectarían los costos de su carga de trabajo. Si un proveedor cambia el costo de su licencia de base de datos, investigue el modo en que eso afectará la eficiencia general de su carga de trabajo. Considere los anuncios históricos de precios de sus proveedores para detectar tendencias en los cambios de licencias en todos sus productos. Los costos de las licencias también pueden escalar independientemente del rendimiento o el uso, como las licencias que escalan en función del hardware (licencias vinculadas a la CPU). Estas licencias se deberían evitar debido a que los costos pueden aumentar rápidamente sin obtener los resultados correspondientes.

Puede utilizar AWS License Manager para administrar las licencias de software de su carga de trabajo. Puede configurar las reglas de las licencias y aplicar las condiciones necesarias para evitar violaciones a las licencias y, además, para reducir los costos de los excedentes de licencias.

Selección del tipo, el tamaño y la cantidad correctos de recursos

Si selecciona la mejor opción de tipo, tamaño y cantidad de recursos, podrá cumplir los requisitos técnicos con el menor costo de recursos. Las actividades de redimensionamiento tienen en cuenta todos los recursos de una carga de trabajo, todos los atributos de cada recurso individual y el esfuerzo que implica la operación de redimensionamiento. Esta operación puede ser un proceso iterativo, desencadenado por los cambios en los patrones de uso y por factores externos, como las caídas en los precios de AWS o los nuevos tipos de recursos de AWS. El redimensionamiento también puede ser una excepción si el costo del esfuerzo para redimensionar los recursos supera los potenciales ahorros durante el tiempo de vida de la carga de trabajo.

En AWS, existe una serie de diferentes enfoques:

- Realizar el modelado de costos
- Seleccionar el tamaño en función de las métricas o los datos
- Seleccionar el tamaño de forma automática (en función de las métricas)

Modelado de costos: realice el modelado de costos para su carga de trabajo y cada uno de sus componentes a fin de comprender el equilibrio entre los recursos y de encontrar el tamaño adecuado para cada uno de los recursos de la carga de trabajo, según un nivel específico de rendimiento. Realice actividades de comparación para la carga de trabajo con diferentes cargas estimadas y compare los costos. El esfuerzo de modelado debería reflejar el beneficio potencial, por ejemplo, el tiempo dedicado es proporcional al costo de los componentes o a los ahorros estimados. Para conocer las prácticas recomendadas, consulte la sección Revisión del documento técnico Pilar de eficiencia de rendimiento del Marco de Buena Arquitectura de AWS.

AWS Compute Optimizer puede ayudarlo con el modelado de costos para las cargas de trabajo en ejecución. Proporciona recomendaciones para el redimensionamiento de los recursos informáticos en función del historial de uso. Este es el origen de datos ideal para los recursos informáticos debido a que se trata de un servicio gratuito y utiliza el aprendizaje automático para hacer varias recomendaciones en función de los niveles de riesgo. También puede utilizar Amazon CloudWatch y CloudWatch Logs con registros personalizados como orígenes de datos para las operaciones de redimensionamiento de otros servicios y componentes de la carga de trabajo.

A continuación, se presentan las recomendaciones para los datos y las métricas del modelado de costos:

- El monitoreo debe reflejar con exactitud la experiencia del usuario final. Seleccione el nivel de detalle correcto para el periodo de tiempo y elija a conciencia el valor máximo o el percentil 99 en lugar de un porcentaje.
- Seleccione el nivel de detalle correcto del periodo de análisis necesario para cubrir todos los ciclos de la carga de trabajo. Por ejemplo, si se lleva a cabo un análisis de dos semanas, es posible que pase por alto un ciclo mensual de alta utilización, que podría conducir a un aprovisionamiento insuficiente.

Selección en función de las métricas o los datos: seleccione el tamaño o el tipo de los recursos en función de la carga de trabajo y de las características de los recursos, como por ejemplo, la informática, la memoria, el rendimiento o el uso intensivo de la escritura. En general, esta selección se efectúa mediante el uso de modelado de costos, una versión previa de la carga de trabajo (como una versión en las instalaciones), de documentación o de otras fuentes de información sobre la carga de trabajo (como documentos técnicos o soluciones publicadas).

Selección automática basada en las métricas: cree un bucle de retroalimentación dentro de la carga de trabajo que utilice métricas activas de la carga de trabajo en ejecución con el fin de realizar cambios en dicha carga de trabajo. Puede utilizar un servicio administrado, como AWS Auto Scaling, que se puede configurar para que efectúe las operaciones de redimensionamiento por usted. AWS también ofrece API, SDK y características que permiten

modificar los recursos con el mínimo esfuerzo. Puede programar una carga de trabajo para que detenga y vuelva a iniciar una instancia EC2 a fin de permitir cambios en el tamaño de la instancia o su tipo. Esto proporciona los beneficios del redimensionamiento al mismo tiempo que elimina casi todos los costos operativos necesarios para efectuar el cambio.

Algunos servicios de AWS cuentan con selección automática integrada del tipo o tamaño de los recursos, como S3 Intelligent Tiering. S3 Intelligent Tiering traslada sus datos de forma automática entre dos capas de acceso, la de acceso frecuente y la de acceso poco frecuente, en función de sus patrones de uso.

Selección del mejor modelo de precios

Realizar modelado de costos de la carga de trabajo: tenga en cuenta los requisitos de los componentes de la carga de trabajo y comprenda los modelos de precios potenciales. Defina los requisitos de disponibilidad de los componentes. Determine si existen varios recursos independientes que lleven a cabo la función en la carga de trabajo y cuáles son sus requisitos con el paso del tiempo. Compare los costos de los recursos considerando el modelo predeterminado de precios bajo demanda y otros modelos aplicables. Tenga en cuenta cualquier cambio potencial en los recursos o en los componentes de la carga de trabajo.

Realizar análisis regulares del nivel de cuenta: llevar a cabo modelados de costos de forma regular garantiza que se puedan implementar las oportunidades de optimización en varias cargas de trabajo. Por ejemplo, si varias cargas de trabajo utilizan el modelo bajo demanda, a un nivel agregado, el riesgo de cambios es más bajo, y la implementación de un descuento en función del compromiso lograría una reducción del costo general. Se recomienda llevar a cabo análisis en ciclos regulares de dos semanas a un mes. Esto le permite efectuar pequeñas compras de ajuste, de manera que la cobertura de sus modelos de precios continúe evolucionando con los cambios en las cargas de trabajo y sus componentes.

Utilice la herramienta de recomendaciones de AWS Cost Explorer para encontrar oportunidades de descuentos por compromiso.

Con el fin de encontrar oportunidades de cargas de trabajo de spot, utilice una vista de su uso general por hora y busque periodos regulares de cambios en el uso o la elasticidad.

Modelos de precios: AWS cuenta con múltiples modelos de precios que le permiten pagar sus recursos de la manera más rentable que se ajuste a las necesidades de su organización. La siguiente sección describe cada modelo de compra:

- Bajo demanda
- De spot
- Descuentos por compromiso: Savings Plans
- Descuentos por compromiso: capacidad o instancias reservadas

- Selección geográfica
- Precios y acuerdos de terceros

Bajo demanda: este es el modelo de precios predeterminado, de precio por uso. Cuando utiliza recursos (por ejemplo, instancias EC2 o servicios como DynamoDB bajo demanda), paga una tarifa plana y no asume compromisos a largo plazo. Puede aumentar o disminuir la capacidad de sus recursos o servicios en función de las demandas de su aplicación. Con el modelo bajo demanda, se aplica una tarifa por hora pero, en función del servicio, se puede facturar por incrementos de 1 segundo (por ejemplo, AWS Lambda o instancias EC2 de Linux). Se recomienda el modelo bajo demanda para las aplicaciones con cargas de trabajo de corto plazo (por ejemplo, un proyecto de cuatro meses) que alcanzan el máximo de forma periódica o con cargas de trabajo no predecibles que no se pueden interrumpir. El modelo bajo demanda también se ajusta a las cargas de trabajo, como los entornos de preproducción, que requieren tiempos de ejecución ininterrumpida pero que no se ejecutan el tiempo suficiente como para obtener un descuento por compromiso (Savings Plans o instancias reservadas).

De spot: Una instancia de spot consiste en capacidad de cómputo de EC2 libre disponible con descuentos de hasta el 90 % en los precios bajo demanda, sin requerir compromisos a largo plazo. Con las instancias de spot, puede reducir de manera significativa el costo de ejecución de sus aplicaciones o escalar la capacidad de cómputo de sus aplicaciones por el mismo presupuesto. A diferencia del modelo bajo demanda, las instancias de spot se pueden interrumpir con una advertencia de 2 minutos si EC2 necesita la capacidad de regreso o si el precio de la instancia de spot excede su precio configurado. En promedio, las instancias de spot se interrumpen menos del 5 % del tiempo.

El modelo de spot es ideal cuando existe una cola o un búfer, o bien, cuando varios recursos trabajan de forma independiente para procesar las solicitudes (por ejemplo, el procesamiento de datos de Hadoop). Por lo general, estas cargas de trabajo son tolerantes a los errores, sin estado y flexibles, como el procesamiento por lotes, el big data y los análisis, los entornos en contenedores y la informática de alto rendimiento (HPC). Las cargas de trabajo no críticas, como los entornos de prueba o desarrollo, también son candidatas para el modelo de spot.

El modelo de spot también está integrado en varios servicios de AWS, como los grupos de EC2 Auto Scaling (ASG), Elastic MapReduce (EMR), Elastic Container Service (ECS) y AWS Batch.

Cuando se debe recuperar una instancia de spot, EC2 envía una advertencia de dos minutos mediante un aviso de interrupción de instancia de spot a través de CloudWatch Events, así como también en los metadatos de la instancia. Durante el periodo de dos minutos, la aplicación puede usar ese tiempo para guardar su estado, purgar contenedores que se están ejecutando, cargar archivos de registro finales o eliminarse de un balanceador de carga. Cuando se cumplen los dos minutos, tiene la opción de hibernar, detener o finalizar la instancia de spot.

Considere las siguientes prácticas recomendadas cuando adopte instancias de spot en sus cargas de trabajo:

- **Establecer su precio máximo como la tarifa bajo demanda:** esto garantiza que pagará la tarifa de spot actual (el precio más bajo disponible) y que nunca pagará más que la tarifa bajo demanda. Las tarifas actuales e históricas están disponibles vía consola y API.
- **Ser flexible en tantos tipos de instancias como sea posible:** sea flexible tanto en la familia como en el tamaño del tipo de instancia, con el fin de mejorar la probabilidad de satisfacer los requisitos de capacidad de destino, obtener el costo más bajo posible y minimizar el impacto de las interrupciones.
- **Ser flexible acerca del lugar donde se ejecutará su carga de trabajo:** la capacidad disponible puede variar según la zona de disponibilidad. Esto mejora la probabilidad de satisfacer la capacidad de su objetivo aprovechando varias reservas de capacidad libre y proporciona el menor costo posible.
- **Diseñar en busca de continuidad:** diseñe sus cargas de trabajo para que tengan tolerancia a errores y no tengan estado, de modo que si parte de su capacidad de EC2 se interrumpe, esto no impacte la disponibilidad y el rendimiento de la carga de trabajo.
- Recomendamos utilizar las instancias de spot junto con instancias bajo demanda y Savings Plans/instancias reservadas para maximizar la optimización de costos de la carga de trabajo con rendimiento.

Descuentos por compromiso (Savings Plans): AWS le ofrece varias maneras de reducir los costos. Puede reservar o comprometerse a utilizar cierta cantidad de recursos y recibir una tarifa con descuento para sus recursos. Un Savings Plan le permite hacer un compromiso de gasto por hora, durante uno o tres años, y recibir un precio con descuento para todos sus recursos. Los Savings Plans ofrecen descuentos para servicios informáticos de AWS como EC2, Fargate y Lambda. Cuando realiza el compromiso, paga el monto de ese compromiso por cada hora y se resta de su uso bajo demanda con la tarifa de descuento. Por ejemplo, se compromete a pagar USD 50 por hora y tiene un uso bajo demanda de USD 150 por hora. Si se tiene en cuenta el precio de Savings Plans, su uso específico tiene una tasa de descuento del 50 %. Entonces, su compromiso de USD 50 cubre USD 100 del uso bajo demanda. Pagará USD 50 (compromiso) y USD 50 del uso bajo demanda restante.

Compute Savings Plans son la opción más flexible y ofrecen un descuento de hasta el 66 %. Se aplican de manera automática en las zonas de disponibilidad, el tamaño de instancia, la familia de instancia, el sistema operativo, la tenencia, la región y el servicio informático.

Instance Savings Plans tienen menos flexibilidad, pero ofrecen una mayor tasa de descuento (hasta el 72 %). Se aplican de manera automática en las zonas de disponibilidad, el tamaño de instancia, la familia de instancia, el sistema operativo y la tenencia.

Existen tres opciones de pago:

- **Sin pago inicial:** no hay pago inicial. Paga una tarifa reducida por hora todos los meses por el total de horas en el mes.

- **Pago inicial parcial:** proporciona una tasa más alta de descuento que la opción sin pago inicial. Parte del uso se paga por anticipado. Luego, usted paga una tarifa reducida por hora de menor monto todos los meses, por el total de horas en el mes.
- **Pago inicial total:** el uso de todo el periodo se paga por adelantado y no se tiene otros gastos por el resto del plazo del uso que está cubierto por el compromiso.

Puede aplicar cualquier combinación de estas tres opciones de compra en todas sus cargas de trabajo.

Los Savings Plans se aplican primero al uso de la cuenta en la que se adquirieron, desde el porcentaje de descuento más alto al más bajo; luego se aplican al uso consolidado en todas las otras cuentas, desde el porcentaje de descuento más alto al más bajo.

Se recomienda comprar todos los Savings Plans en una cuenta sin uso ni recursos, como la cuenta maestra. Esto garantiza que los Savings Plans se apliquen a las tasas de descuento más altas en todo su uso y, de esta forma, se maximiza el monto de descuento.

Las cargas de trabajo y el uso normalmente cambian con el paso del tiempo. Se recomienda comprar continuamente cantidades pequeñas de compromiso de Savings Plans a lo largo del tiempo. Esto garantiza que mantendrá altos niveles de cobertura para maximizar sus descuentos y que sus planes se ajustarán a su carga de trabajo y a los requisitos de su organización en todo momento.

No establezca un objetivo de cobertura en sus cuentas, debido a la posible variabilidad de descuento. Una cobertura baja no necesariamente indica grandes ahorros potenciales. Es posible que tenga una baja cobertura en su cuenta, pero si su uso se compone de instancias pequeñas, con un sistema operativo con licencia, el ahorro potencial puede ser de solo algunos puntos porcentuales. En su lugar, realice un seguimiento y monitoreo de los ahorros potenciales disponibles en la herramienta de recomendación Savings Plans. Revise a menudo las recomendaciones de Savings Plans en Cost Explorer (realice análisis regulares) y continúe adquiriendo compromisos hasta que los ahorros estimados se encuentren por debajo del descuento requerido para la organización. Por ejemplo, realice un seguimiento y monitoreo de sus descuentos potenciales para que permanezcan por debajo del 20 %; si superan ese porcentaje, se debe realizar una compra.

Monitoree la utilización y la cobertura, pero solo para detectar cambios. No apunte a un porcentaje específico de utilización o un porcentaje de cobertura, ya que esto no necesariamente escala con los ahorros. Asegúrese de que una compra de Savings Plans tenga como resultado un aumento en la cobertura y, si existen disminuciones en la cobertura o utilización, asegúrese de que estén contabilizadas y se conozcan. Por ejemplo, usted migra un recurso de carga de trabajo a un nuevo tipo de instancia, que reduce la utilización de un plan existente, pero el beneficio de rendimiento supera la reducción de ahorro.

Descuentos por compromiso (instancias reservadas/compromiso): de manera similar a Savings Plans, las instancias reservadas ofrecen descuentos de hasta 72 % por un compromiso

a ejecutar una cantidad mínima de recursos. Las instancias reservadas están disponibles para RDS, Elasticsearch, ElastiCache, Amazon Redshift y DynamoDB. Amazon CloudFront y AWS Elemental MediaConvert también ofrecen descuentos cuando realiza compromisos de uso mínimo. Las instancias reservadas están disponibles actualmente para EC2; sin embargo, Savings Plans ofrecen los mismos niveles de descuento con mayor flexibilidad y sin gastos generales de administración.

Las instancias reservadas ofrecen las mismas opciones de precios sin pago inicial, con pago inicial parcial y pago inicial total, y los mismos plazos de uno o tres años.

Las instancias reservadas se pueden adquirir en una región o una zona de disponibilidad específica. Proporcionan una reserva de capacidad cuando se adquieren una zona de disponibilidad.

EC2 cuenta con IR convertibles; sin embargo, los Savings Plans deberían utilizarse para todas las instancias de EC2 debido a su mayor flexibilidad y sus costos operativos reducidos.

Se deben usar el mismo proceso y las mismas métricas para controlar y realizar compras de instancias reservadas. Se recomienda no hacer un seguimiento de la cobertura de IR en todas las cuentas. También se recomienda no hacer un monitoreo ni un seguimiento del porcentaje de utilización, sino visualizar el informe de utilización en Cost Explorer y usar la columna de ahorro neto en la tabla. Si el ahorro neto es un monto negativo significativamente grande, debe tomar medidas para recuperar las IR sin utilizar.

EC2 Fleet: EC2 Fleet es una característica que le permite definir un objetivo de capacidad de cómputo y, luego, especificar los tipos de instancia y el equilibrio entre instancias bajo demanda y de spot para la flota. EC2 Fleet lanzará de manera automática la combinación más económica de recursos para satisfacer la capacidad definida.

Selección geográfica: cuando diseña sus soluciones, una práctica recomendada es buscar acercar los recursos informáticos a los usuarios para brindar menor latencia y una soberanía de datos sólida. En el caso de audiencias globales, debe utilizar varias ubicaciones para satisfacer estas necesidades. Debe seleccionar la ubicación geográfica que minimice sus costos.

La infraestructura de la nube de AWS se basa en las regiones y zonas de disponibilidad. Una región es una ubicación física en el mundo donde tenemos varias zonas de disponibilidad. Las zonas de disponibilidad consisten en uno o más centros de datos discretos, cada uno con potencia, redes y conectividad redundantes, alojados en instalaciones separadas.

Cada región de AWS opera dentro de las condiciones de mercado locales y el precio de los recursos es distinto en cada región. Elija una región específica para operar un componente de su solución o la solución entera, de modo que pueda operar al menor precio posible de manera global. Puede utilizar la calculadora de costo mensual de AWS para estimar los costos de su carga de trabajo en varias regiones.

Precios y acuerdos de terceros: cuando utiliza soluciones o servicios de terceros en la nube, es importante que las estructuras de precios se ajusten a los resultados de la optimización de

costos. El precio debe escalar junto con los resultados y el valor que proporciona. Un ejemplo de esto es el software que toma un porcentaje de los ahorros que proporciona; mientras más ahorra (resultado), más cobra. Los acuerdos que escalan con su factura normalmente no se ajustan a la optimización de costos, a menos que proporcionen resultados para cada parte de su factura específica. Por ejemplo, una solución que proporciona recomendaciones para EC2 y cobra un porcentaje de su factura total aumentará si usted utiliza otros servicios para los cuales no proporciona ningún beneficio. Otro ejemplo es un servicio administrado que se cobra a un porcentaje del costo de los recursos que se administran. Es posible que un tamaño más grande de instancia no requiera necesariamente más esfuerzo de administración, pero se cobrará más. Asegúrese de que los arreglos de precios de estos servicios incluyan un programa o características de optimización de costos en su servicio para impulsar la eficiencia.

Planificación de la transferencia de datos

Una ventaja de la nube es que es un servicio de red administrado. Ya no existe la necesidad de administrar y operar una flota de interruptores, enrutadores y otros equipos asociados a redes. Los recursos de redes en la nube se consumen y pagan de la misma manera en la que paga por una CPU y almacenamiento: solo paga por lo que usa. Es necesario un uso eficiente de los recursos de red para la optimización de costos en la nube.

Realizar modelado de transferencia de datos: comprenda dónde ocurre la transferencia de datos en su carga de trabajo, el costo de la transferencia y su beneficio asociado. Esto le permite tomar una decisión informada al modificar o aceptar la decisión de diseño. Por ejemplo, es posible que tenga una configuración de Multi-Availability Zone donde replica los datos entre las zonas de disponibilidad. Usted modela el costo de estructura y decide que es un costo aceptable (de manera similar a los gastos informáticos y de almacenamiento en ambas zonas de disponibilidad) para lograr la fiabilidad y resiliencia necesarias.

Modele los costos en diferentes niveles de uso. El uso de la carga de trabajo puede cambiar con el tiempo y es posible que diferentes servicios resulten más rentables en diferentes niveles.

Utilice AWS Cost Explorer o el informe de uso y costo (CUR) para comprender y modelar los costos de transferencia de datos. Configure una prueba de concepto (PoC) o pruebe su carga de trabajo, y ejecute una prueba con una carga simulada realista. Puede modelar los costos en diferentes demandas de carga de trabajo.

Optimizar la transferencia de datos: diseñar la transferencia de datos garantiza que minimice los costos de transferencia de datos. Esto puede comprender el uso de redes de entrega de contenido para ubicar datos más cerca de los usuarios o el uso de enlaces de red dedicada de sus instalaciones a AWS. También puede utilizar optimización de WAN y optimización de aplicación para reducir la cantidad de datos que se transfieren entre componentes.

Seleccionar servicios para reducir los costos de transferencia de datos: Amazon CloudFront es una red de entrega de contenido global que entrega datos con latencia baja y velocidades altas

de transferencia. Almacena en caché datos en ubicaciones de borde de todo el mundo, lo que reduce la carga en sus recursos. Si utiliza CloudFront, puede reducir el trabajo administrativo de entregar contenido a un gran número de usuarios a nivel global, con mínima latencia.

AWS Direct Connect le permite establecer una conexión de red dedicada con AWS. Esto puede reducir los costos de red, aumentar el ancho de banda y proporcionar una experiencia de red más estable que las conexiones basadas en Internet.

AWS VPN le permite establecer una conexión segura y privada entre su red privada y la red global de AWS. Es ideal para oficinas pequeñas o socios comerciales pequeños ya que proporciona una conectividad rápida y fácil, y es un servicio completamente administrado y elástico.

Los puntos de enlace de la VPC permiten conectividad entre los servicios de AWS en redes privadas y pueden utilizarse para reducir la transferencia pública de datos y los costos de gateways NAT. Los puntos de enlace de la VPC de gateway no tienen cargos por hora y son compatibles con Amazon S3 y Amazon DynamoDB. Los puntos de enlace de la VPC de interfaz son proporcionados por AWS PrivateLink y tienen una tarifa por hora y por costo de uso de GB.

Recursos

Consulte los siguientes recursos para obtener más información acerca de las prácticas recomendadas de AWS para recursos rentables.

- AWS Managed Services: Enterprise Transformation Journey Video
- Análisis de los costos con Cost Explorer
- Acceso a recomendaciones de instancias reservadas
- Introducción a las recomendaciones de redimensionamiento
- Prácticas recomendadas para instancias de spot
- Flotas de spot
- Funcionamiento de las instancias reservadas
- Infraestructura global de AWS
- Asistente de instancias de spot
- Well-Architected Labs - Recursos rentables

Administración de los recursos de oferta y demanda

Una vez que migra a la nube, solo paga por lo que necesita. Puede proporcionar recursos para adaptarse a la demanda de la carga de trabajo en el momento en el que son necesarios, lo que elimina la necesidad de un aprovisionamiento excesivo de gran costo y que derrocha recursos. También puede modificar la demanda usando un límite, un búfer o una cola para mitigar la demanda y satisfacerla con menos recursos.

Los beneficios económicos de un suministro justo a tiempo deben equilibrarse con la necesidad de aprovisionamiento para adaptarse a las fallas de recursos, la alta disponibilidad y el tiempo de aprovisionamiento. Dependiendo de si su demanda es fija o variable, planifique la creación de métricas y automatización que garanticen que la administración de su entorno sea mínima, incluso a medida que escala. Cuando modifica la demanda, debe conocer la demora admisible y máxima que puede permitir la carga de trabajo.

En AWS, puede usar distintos enfoques para administrar los recursos de oferta y demanda. Las siguientes secciones describen cómo utilizar esos enfoques:

- Análisis de la carga de trabajo
- Administración de la demanda
- Oferta en función de la demanda
- Oferta en función del tiempo

Analizar la carga de trabajo: conozca los requisitos de la carga de trabajo. Los requisitos de organización deben indicar los tiempos de respuesta de la carga de trabajo para las solicitudes. El tiempo de respuesta puede usarse para determinar si la demanda está administrada o si la oferta de recursos cambiará para cumplir con la demanda.

El análisis debería incluir la predictibilidad y repetibilidad de la demanda, la tasa de cambio de la demanda y la medida del cambio de la demanda. Asegúrese de que el análisis se realice durante un periodo lo suficientemente largo como para incorporar cualquier tipo de variación estacional, como el procesamiento de fin de mes o los picos por las fiestas.

Asegúrese de que el trabajo de análisis refleje los beneficios potenciales de implementar el escalado. Observe el costo total esperado del componente y cualquier aumento o disminución del uso y del costo durante la vida útil de la carga de trabajo.

Puede usar AWS Cost Explorer o Amazon QuickSight con el CUR o sus registros de aplicaciones para realizar un análisis visual de la demanda de carga de trabajo.

Administración de la demanda

Administrar la demanda (limitación controlada): si el origen de la demanda tiene capacidad para reintentos, puede implementar una limitación controlada. La limitación controlada le indica al origen que, si no puede atender la solicitud en el momento actual, debe volver a intentarlo más tarde. El origen esperará durante un periodo y intentará realizar la solicitud de nuevo. Implementar la limitación controlada tiene como ventaja que se coloca un límite a la cantidad máxima de recursos y costos de la carga de trabajo. En AWS, puede usar Amazon API Gateway para implementar la limitación controlada. Consulte el documento técnico sobre el pilar de fiabilidad de Well-Architected para obtener más detalles acerca de la implementación de la limitación controlada.

Administrar la demanda (en función del búfer): de manera similar a la limitación controlada, un búfer posterga el procesamiento de solicitudes, lo que permite que aplicaciones que funcionan a velocidades diferentes se comuniquen con eficacia. Un enfoque en función del búfer usa una cola para aceptar mensajes (unidades de trabajo) de los productores. Los consumidores leen los mensajes y los procesan, lo que permite que los mensajes se ejecuten a la velocidad que cumpla los requisitos del negocio de los consumidores. No debe preocuparse por que los productores deban enfrentar los problemas de la limitación controlada, como la durabilidad de los datos y la contrapresión (donde los productores reducen su velocidad porque el consumidor funciona lentamente).

En AWS, puede elegir entre múltiples servicios para implementar un enfoque de almacenamiento en búfer. Amazon SQS es un servicio administrado que proporciona colas para permitir que un único consumidor lea mensajes individuales. Amazon Kinesis proporciona un flujo que permite que muchos consumidores lean los mismos mensajes.

Cuando diseña con un enfoque en función del búfer, asegúrese de que diseña su carga de trabajo para atender la solicitud en el tiempo requerido y que puede administrar solicitudes de trabajo duplicadas.

Oferta dinámica

Oferta en función de la demanda: aproveche la elasticidad de la nube para proporcionar recursos y satisfacer la demanda cambiante. Utilice las API o las características de servicio para variar mediante programación la cantidad de recursos de la nube en su arquitectura de manera dinámica. Esto le permite escalar componentes de su arquitectura y aumentar automáticamente la cantidad de recursos durante picos de demanda para mantener el rendimiento, y disminuir la capacidad cuando la demanda se contrae con el fin de reducir costos.

Auto Scaling lo ayuda a ajustar su capacidad para mantener un rendimiento estable y predecible con el mínimo costo posible. Es un servicio gratuito completamente administrado que se integra con instancias de Amazon EC2 y flotas de spot, Amazon ECS, Amazon DynamoDB y Amazon Aurora.

Auto Scaling proporciona una función de detección automática de recursos para ayudar a encontrar recursos en su carga de trabajo que puedan configurarse. También cuenta con estrategias de escalado incorporadas para optimizar el rendimiento, los costos o un equilibrio entre ambos, y proporciona escalado predictivo para ayudar con los picos que se producen regularmente.

Auto Scaling puede implementar un escalado manual, programado o en función de la demanda. Además, puede usar métricas y alarmas de Amazon CloudWatch con el fin de desencadenar eventos de escalado para su carga de trabajo. Algunas métricas típicas pueden ser métricas estándar de Amazon EC2, como la utilización de CPU, el rendimiento de la red y la latencia de solicitud/respuesta observada de ELB. Siempre que sea posible, debe usar una métrica representativa de la experiencia de usuario. En general, esta es una métrica personalizada que puede surgir del código de la aplicación dentro de su carga de trabajo.

Cuando diseña con un enfoque en función de la demanda, tenga en cuenta dos consideraciones clave. En primer lugar, debe comprender la rapidez con la que debe aprovisionar nuevos recursos. En segundo lugar, debe comprender que el tamaño del margen entre la oferta y la demanda cambiará. Debe estar listo para lidiar con la tasa de cambio de demanda y, adicionalmente, con errores en los recursos.

Elastic Load Balancing (ELB) lo ayuda a escalar distribuyendo la demanda entre múltiples recursos. A medida que implementa más recursos, debe agregarlos en el balanceador de carga para hacerse cargo de la demanda. AWS ELB admite instancias de EC2, contenedores, direcciones IP y funciones Lambda.

Oferta en función del tiempo: un enfoque en función del tiempo ajusta la capacidad de recursos a la demanda predecible o bien definida por el tiempo. En general, este enfoque no depende de los niveles de utilización de los recursos. Un enfoque en función del tiempo garantiza que los recursos estén disponibles en el momento específico en el que se requieren y que puedan proporcionarse sin demoras causadas por procedimientos de arranque y verificaciones del sistema o de consistencia. Con un enfoque en función del tiempo, puede proporcionar recursos adicionales o aumentar la capacidad durante periodos con más carga.

Puede usar Auto Scaling de manera programada para implementar un enfoque en función del tiempo. Se puede programar que aumente la escala de cargas de trabajo en determinadas horas (por ejemplo, al comienzo del horario de trabajo), lo que asegura que los recursos estén disponibles cuando lleguen los usuarios o la demanda.

También puede aprovechar las API y los SDK de AWS y AWS CloudFormation para aprovisionar y retirar entornos enteros a medida que lo necesite. Este enfoque es adecuado para los entornos de desarrollo o prueba que funcionan solo en horarios de trabajo o periodos definidos.

Puede usar las API para escalar el tamaño de los recursos dentro de un entorno (escalado vertical). Por ejemplo, puede aumentar la escala de una carga de trabajo de producción cambiando la clase o el tamaño de instancia. Esto puede lograrse deteniendo e iniciando nuevamente la instancia, y seleccionando una clase o un tamaño de instancia diferentes. Esta técnica puede aplicarse a otros

recursos, como los volúmenes elásticos de EBS, que puede aumentar de tamaño, ajustar el rendimiento (IOPS) o cambiar de tipo de volumen mientras se encuentran en uso.

Cuando diseña con un enfoque en función del tiempo, tenga en cuenta dos consideraciones clave. En primer lugar, ¿qué tan consistente es el patrón de uso? En segundo lugar, ¿cuál es el impacto si cambia el patrón? Puede aumentar la precisión de las predicciones monitoreando sus cargas de trabajo y utilizando inteligencia empresarial. Si ve cambios significativos en el patrón de uso, puede ajustar las horas para garantizar que se proporcione cobertura.

Oferta dinámica: puede usar AWS Auto Scaling o incorporar el escalado en su código con las API o los SDK de AWS. Esto reduce sus costos generales de carga de trabajo eliminando el costo operativo por realizar cambios manualmente a su entorno, que, además, pueden realizarse mucho más rápido. Esto garantizará que los recursos para la carga de trabajo coincidan de la mejor manera con la demanda en cualquier momento.

Recursos

Consulte los siguientes recursos para obtener más información acerca de las prácticas recomendadas de AWS para administrar los recursos de oferta y demanda.

- Limitación controlada de API Gateway
- Introducción a Amazon SQS
- Introducción a Amazon EC2 Auto Scaling

Optimización con el paso del tiempo

En AWS, puede lograr la optimización con el paso del tiempo revisando nuevos servicios e implementándolos en su carga de trabajo.

Revisión e implementación de nuevos servicios

A medida que AWS lanza nuevos servicios y características, una práctica recomendada es revisar sus decisiones sobre arquitectura para garantizar que continúen siendo rentables. Cuando los requisitos cambian, debe ser enérgico a la hora de retirar recursos, componentes y cargas de trabajo que ya no requiera. Considere hacer lo siguiente para ayudarlo a lograr la optimización con el paso del tiempo:

- Desarrollar un proceso de revisión de carga de trabajo
- Analizar e implementar servicios

Desarrollar un proceso de revisión de carga de trabajo: para garantizar que siempre tenga la carga de trabajo más rentable, debe revisarla con regularidad para saber si existen oportunidades

para implementar servicios, características y componentes nuevos. Para garantizar que obtiene costos generales menores, el proceso debe ser proporcional a la cantidad de ahorros potencial. Por ejemplo, las cargas de trabajo que representen el 50 % de su gasto general deben revisarse con más regularidad y minuciosidad que las cargas de trabajo que representan el 5 % de su gasto general. Tenga en cuenta todas la volatilidad o los factores externos. Si la carga de trabajo brinda servicios a una región geográfica o una parte del mercado específica, y se predicen cambios en dicha área, hacer revisiones con mayor frecuencia podría generar ahorros de costos. Otro factor en la revisión es el esfuerzo para implementar los cambios. Si los costos de probar y validar cambios son significativos, las revisiones deben ser menos frecuentes.

Tenga en cuenta el costo a largo plazo de mantenimiento de componentes y recursos desactualizados y heredados, y la incapacidad de implementar nuevas características en ellos. El costo actual de realizar pruebas y validaciones puede superar el beneficio propuesto. Sin embargo, con el paso del tiempo, el costo de implementar el cambio puede aumentar significativamente a medida que aumenta la brecha entre la carga de trabajo y las tecnologías actuales, lo que trae como resultado costos aún mayores. Por ejemplo, es posible que el costo de cambiar a un nuevo lenguaje de programación no sea rentable en la actualidad. Sin embargo, en cinco años, el costo de las personas con habilidades en ese lenguaje puede aumentar y, debido al crecimiento de la carga de trabajo, debería trasladar un sistema aún mayor al nuevo lenguaje, lo que requeriría un esfuerzo todavía mayor que el anterior.

Divida su carga de trabajo en componentes, asigne el costo de cada componente (una estimación es suficiente) y enumere los factores (por ejemplo, esfuerzo y mercados externos) junto a cada componente. Utilice estos indicadores para determinar una frecuencia de revisión para cada carga de trabajo. Por ejemplo, puede considerar a los servidores web como una carga de trabajo con esfuerzo de alto costo y pocos cambios y con alta incidencia de factores externos, lo que da como resultado una frecuencia alta de revisión. Una base de datos centralizada podría ser una carga de trabajo con esfuerzo de costo medio y muchos cambios de cambio y con baja incidencia de factores externos, lo que da como resultado una frecuencia media de revisión.

Revisar la carga de trabajo e implementar servicios: para aprovechar los beneficios de las características y los servicios nuevos de AWS, debe ejecutar el proceso de revisión sobre sus cargas de trabajo e implementar nuevos servicios y características según sean necesarios. Por ejemplo, puede revisar sus cargas de trabajo y reemplazar el componente de mensajería con Amazon Simple Email Service (SES). Esto elimina el costo de operar y mantener una flota de instancias, al mismo tiempo que proporciona toda la funcionalidad a un costo reducido.

Conclusión

La optimización de costos y la administración financiera en la nube representan un esfuerzo continuo. Debe trabajar regularmente con sus equipos de finanzas y tecnología, revisar su enfoque respecto de la arquitectura y actualizar su selección de componentes.

AWS se esfuerza por ayudarlo a minimizar costos mientras usted crea implementaciones con gran resiliencia, respuesta y adaptabilidad. Para optimizar verdaderamente el costo de su implementación, aproveche las herramientas, técnicas y prácticas recomendadas que se analizaron en este documento.

Colaboradores

Las personas que colaboraron en este documento son las siguientes:

- Philip Fitzsimons, director sénior de Well-Architected, Amazon Web Services
- Nathan Besh, líder de costos de Well-Architected, Amazon Web Services
- Levon Stepanian, Amazon Web Services
- Keith Jarrett, líder de desarrollo de negocios, Optimización de costos
- PT Ng, arquitecto comercial, Amazon Web Services
- Arthur Basbaum, director de desarrollo de negocios, Amazon Web Services
- Jarman Hauser, arquitecto comercial, Amazon Web Services

Documentación adicional

Para obtener información adicional, consulte lo siguiente:

- Marco de Buena Arquitectura de AWS

Revisiones del documento

Fecha	Descripción
Abril de 2020	Se actualizó para incorporar CFM, nuevos servicios y la integración con Well-Architected.
Julio de 2018	Se actualizó para reflejar cambios en AWS e incorporar aprendizajes de las revisiones con los clientes.
Noviembre de 2017	Se actualizó para reflejar cambios en AWS e incorporar aprendizajes de las revisiones con los clientes.
Noviembre de 2016	Primera publicación