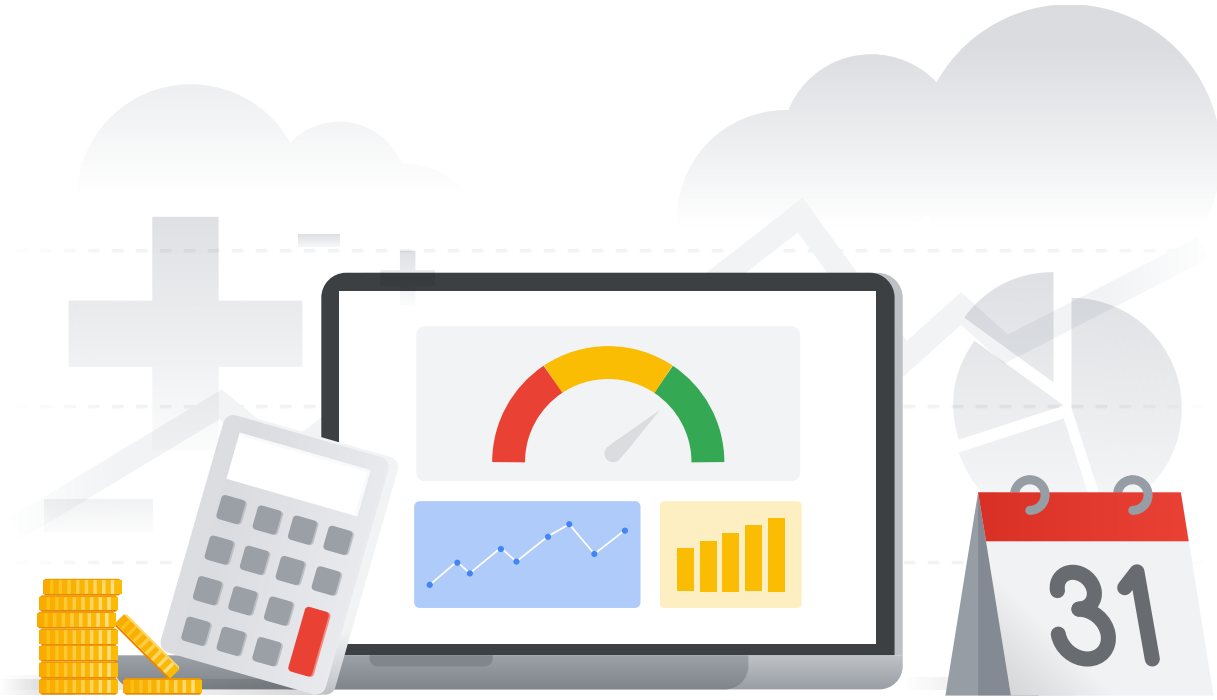
Informe de Google Cloud

Entender los principios de la optimización de costos



Google Cloud

Google Cloud

Índice

Google Cloud

Índice (continued)

Google Cloud

Colaboradores

Gracias a los miembros del equipo de Servicios profesionales de Google Cloud por sus aportes:

- Justin Lerma
- Pathik Sharma
- Amber Yadron
- Andrew Sallaway
- Akshay Kumbhar

Introducción

Siempre hay un cierto grado de incertidumbre cuando haces planes y estableces metas para tu empresa. Nunca sabes realmente cómo responderán los clientes a un producto o si las proyecciones de ventas serán correctas. Sin embargo, puedes controlar el nivel de eficiencia con el que ejecutas tus sistemas tecnológicos y hacer un seguimiento del nivel de satisfacción de tus usuarios empresariales. En la actualidad, con una incertidumbre global aún mayor sumada a la incertidumbre habitual de los negocios, es todavía más importante que uses tus recursos tecnológicos existentes de forma inteligente, y saques el máximo provecho a lo que ya tienes.

A fin de cuentas, obtener más de tus recursos de nube puede traducirse en más clientes satisfechos, más problemas resueltos y mayor adaptabilidad para la empresa en general. Usar tus recursos en la nube de manera más eficiente puede ayudar a tu equipo y a tu empresa a ajustarse a esta nueva realidad, y a ser lo más eficaces posible.

Nuestros propios equipos han trabajado durante años con los equipos de operaciones y TI de diferentes industrias en todo el mundo para saber cuáles son tus desafíos, tus logros y tus planes para el futuro. Aunque mucho ha cambiado, la capacidad de los especialistas en tecnología de adaptarse y tener éxito no lo ha hecho. Tu habilidad para adaptarte a medida que cambian los negocios es más importante que nunca, y nuestra tecnología de nube está diseñada para respaldar ese tipo de agilidad y resiliencia. Hemos recabado los consejos y las buenas prácticas que recomendamos para que puedas sacarle mayor provecho a tus recursos actuales (más VM, más almacenamiento, más consultas) y ayudar a tu empresa a alcanzar sus objetivos.

Continúa leyendo para aprender a usar atributos de optimización de costos integrados en Google Cloud y obtener muchas más recomendaciones para asegurarte de que estás aprovechando tus recursos al máximo. Y, algo importante: encontrarás consejos sobre cómo establecer procesos y trabajar con equipos interdisciplinarios para implementar estándares que conviertan a la eficiencia de la nube y la optimización de costos en parte de la cultura de tu empresa.

Tu capacidad para adaptarte a medida que cambian los negocios es más importante que nunca.

Capítulo 1
Principios y procesos para optimizar tus costos de nube

La nube es mucho más que un centro de costos. Migrar a la nube te permite lograr la innovación a escala global, incrementar la velocidad de los atributos para lograr un tiempo más rápido de salida al mercado y crear una ventaja competitiva al responder rápidamente a las necesidades de los clientes. Por lo tanto, no es ninguna sorpresa que muchas empresas estén buscando transformar la estrategia digital de sus organizaciones lo antes posible. Sin embargo, aunque tiene sentido adoptar la nube rápidamente, también es importante tomarse el tiempo necesario para revisar los conceptos clave antes de migrar o desplegar tus aplicaciones en la nube. Además, si ya tienes aplicaciones en la nube, es recomendable que audites tu entorno para asegurarte de estar implementando las mejores prácticas. La meta es maximizar el valor de la empresa a la vez que se optimizan los costos, teniendo en cuenta el uso más eficaz y eficiente de los recursos de nube.

A la hora de optimizar costos, es posible usar muchas herramientas y técnicas

Hemos estado trabajando a la par de algunos clientes complejos a medida que introducen la nueva generación de aplicaciones y servicios en Google Cloud. A la hora de optimizar costos, las organizaciones tienen muchas herramientas y técnicas a su disposición. Pero las herramientas son útiles hasta cierto punto. En nuestra experiencia, hay varios principios de primer nivel que las organizaciones, independientemente de su tamaño, pueden seguir para asegurarse de que están sacando el máximo provecho de la nube.

Optimización de costos con personas y procesos

Como sucede con la mayoría de las cosas en la tecnología, los estándares más altos sólo sirven en la medida en que sean respetados. En la mayoría de los casos, el factor limitante no es la capacidad de la tecnología, sino las personas y los procesos involucrados. La confluencia de los equipos ejecutivos, líderes de

proyectos, finanzas e ingenieros de fiabilidad de sitios (SRE) es la que entra en juego a la hora de optimizar los costos. El primer paso es que estos equipos se reúnan para diseñar un conjunto de estándares para la empresa que definan la rentabilidad, la fiabilidad y el rendimiento deseados a nivel de servicios. Recomendamos enérgicamente formar un equipo de especialistas con metas específicas para impulsar esta iniciativa. Más adelante en este libro abordaremos recomendaciones específicas a nivel de servicio, pero ninguna de ellas será escalable y eficaz hasta tanto tu marco de trabajo se aplique programáticamente.

Usar la mayor visibilidad de costos que ofrece la nube

Uno de los beneficios principales que proporciona un entorno de nube es la mayor visibilidad de tus datos de uso, ya que es posible realizar un seguimiento de cada servicio de nube y medirlo de forma independiente. Esto puede convertirse en una espada de doble filo: ahora tienes decenas de miles de SKU y, si no sabes quién está adquiriendo los servicios, qué servicios son y por qué, se hace difícil entender el costo total de propiedad (TCO) para la aplicación o servicio desplegado en la nube.

Este problema es común cuando los clientes realizan el cambio inicial de un modelo de gastos de capital on-premise (CapEx) a uno de gastos operativos basado en la nube (OpEx). Anteriormente, un equipo central de finanzas definía un presupuesto estático y luego adquiría los recursos necesarios. La proyección se basaba en una métrica, como el crecimiento histórico, para determinar las necesidades del siguiente mes, trimestre, año o incluso para varios años después. No se realizaba ninguna compra hasta que todos en la empresa tuvieran la posibilidad de reunirse y opinar si el gasto era necesario o no.

Para muchos clientes de servicios de nube, el escenario a menudo se parece bastante al lejano oeste.

En la actualidad, en un entorno de OpEx, un equipo de ingeniería puede crear recursos según lo desee para ejecutar sus servicios de forma óptima. Para muchos clientes de servicios de nube, el escenario a menudo se parece bastante al lejano oeste, ya que el equipo de ingeniería crea recursos sin protecciones estandarizadas, como crear presupuestos y alertas, sin una clasificación correcta de los recursos y sin abordar los costos desde una perspectiva de

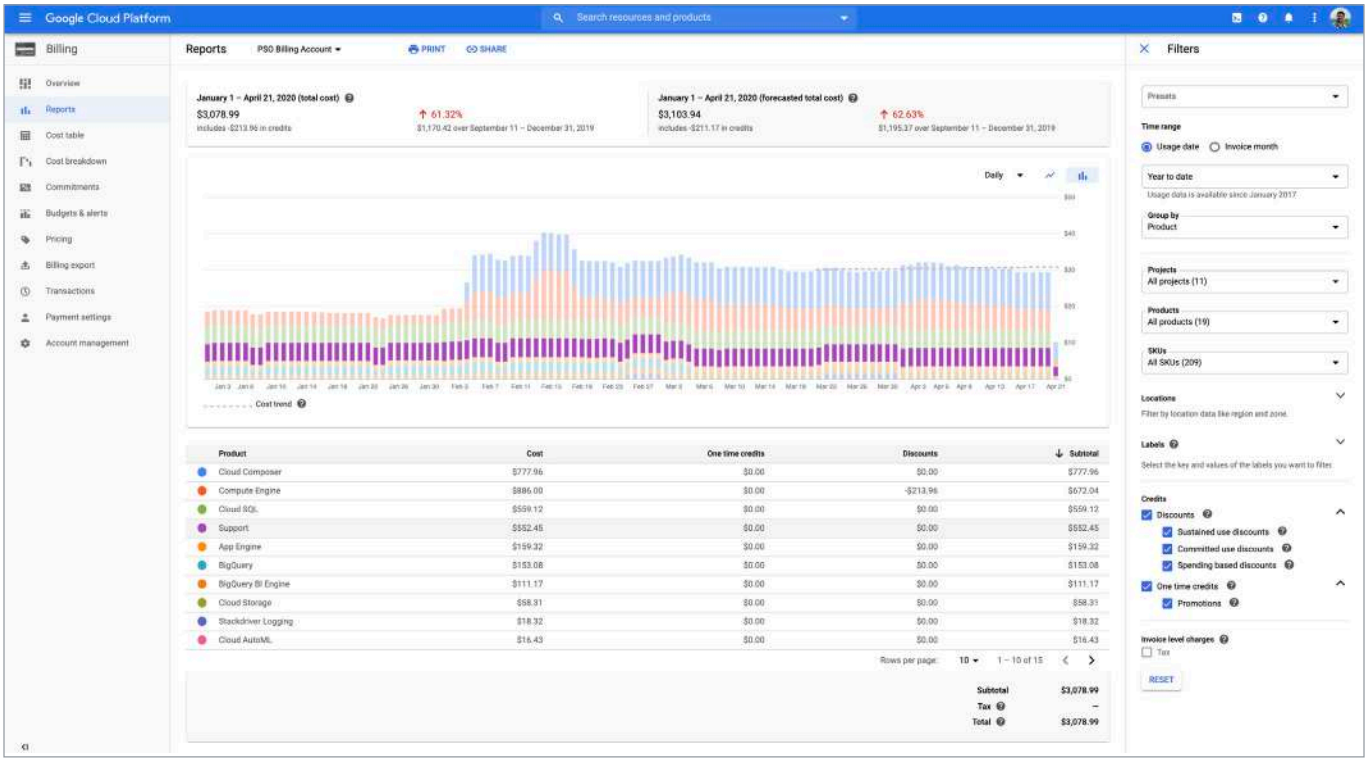
ingeniería y finanzas. Aunque esto aumenta la velocidad, realmente no es una buena posición de partida para diseñar eficazmente una ecuación de costo de generación de valor para un servicio (el valor generado por el servicio), mucho menos para optimizar los gastos. Vemos que los clientes tienen dificultades a la hora de identificar el costo de los proyectos de desarrollo en comparación con los proyectos de producción en sus entornos debido a la falta de una clasificación estandarizada. En otros casos, observamos que los ingenieros aprovisionan instancias en exceso para evitar problemas de rendimiento, lo que solo genera gastos generales considerables durante los periodos de poco uso. Esto genera un derroche de recursos en el largo plazo. Crear estándares aplicables a toda la compañía sobre qué tipos de recursos están disponibles y cuándo desplegarlos es fundamental para optimizar tus costos de nube.

Hemos observado esta dinámica muchas veces, y es lamentable que en ocasiones se considere a una de las características más deseables de la nube (la elasticidad) como un problema. Cuando hay un aumento imprevisto en una factura, algunos clientes podrían considerar preocupante ese aumento en los costos. A menos que atribuyas el costo a las métricas de la empresa, como cantidad de transacciones procesadas o de usuarios a los que se han prestado servicios, realmente no tienes suficiente contexto para interpretar tu factura de servicios de nube. Para muchos clientes, es más fácil ver que los costos están subiendo y atribuir ese aumento a un responsable o grupo determinado dentro de la empresa, pero no tienen suficiente contexto para brindar una recomendación específica al propietario del proyecto. El equipo podría estar gastando más dinero porque está proporcionando servicios a más clientes, lo cual es bueno. Por otro lado, los costos podrían estar subiendo porque alguien olvidó apagar una VM innecesaria con un alto consumo de CPU durante el fin de semana (y que está dirigiendo un tráfico innecesario a Australia).

Una forma de corregir este problema es organizar y estructurar tus costos en función de tus necesidades empresariales. Luego, puedes examinar a fondo los servicios con la ayuda de los informes de facturación de Cloud para obtener un panorama general de tus costos. También puedes visualizar los costos de tu entorno de forma más detallada atribuyendo los costos a los departamentos o equipos

A veces se considera a uno de los atributos más deseables de la nube (la elasticidad) como un problema.

a través de etiquetas y desarrollando tus propios paneles personalizados. Este enfoque te permite etiquetar un recurso con base en una métrica predefinida de la empresa y, posteriormente, hacer un seguimiento del gasto que genera en el tiempo. A largo plazo, la meta no es entender que gastaste "$X en Compute Engine el mes pasado", sino que "cuesta $X brindarles servicios a los clientes que nos generan $Y en ingresos". Este es el tipo de análisis en el que debes trabajar.



Los informes de facturación en la consola de Google Cloud te permiten explorar los costos en detalle.

Una de las principales características de la nube es que te permite incrementar la velocidad de los atributos para lograr un tiempo de salida al mercado más rápido, y es esta elasticidad la que te permite desplegar cargas de trabajo en cuestión de minutos, en lugar de tener que esperar meses en el entorno tradicional on-premise. Es probable que no sepas cuán rápido crecerá tu empresa, por lo que establecer un modelo de visibilidad de costos

desde el principio es esencial. Y una vez que superes las métricas simples de costo por servicio, puedes comenzar a medir nuevas métricas de la empresa; por ejemplo, la rentabilidad como métrica del rendimiento por proyecto..

Entender el valor versus el costo

El objetivo de crear un sistema de nube complejo no es simplemente reducir los costos. A modo de ejemplo, piensa en tus objetivos de entrenamiento. Cuando quieren ponerse en forma, muchas personas se concentran solo en perder peso. Sin embargo, la pérdida de peso no siempre es un indicador clave por sí sola. Puedes perder peso por padecer una enfermedad o tener deshidratación. Cuando apuntamos a un indicador como la pérdida de peso, lo que realmente nos importa es nuestro estado físico general, nuestro aspecto o lo bien que nos sentimos al estar activos, como poder jugar con nuestros hijos, tener una larga vida, bailar, ese tipo de cosas. En el mundo de la optimización de costos sucede lo mismo: no se trata solamente de bajar los costos, sino de identificar los costos innecesarios y asegurarse de maximizar el valor de cada dólar que se gasta.

Del mismo modo, nuestros clientes más sofisticados no están obsesionados con una cifra específica de reducción de costos, si no que se hacen una gran cantidad de preguntas para alcanzar su mejor estado operativo general:

- ¿Qué les estamos proporcionando realmente a nuestros clientes (unidad)?
- ¿Cuánto me cuesta proporcionar eso y sólo eso?
- ¿Cómo puedo optimizar todos los gastos relacionados por unidad creada?

En pocas palabras, se han adelantado y han creado su propio modelo de gastos e ingresos por unidad (unit economics). Se formulan estas preguntas desde un primer momento y luego trabajan para desarrollar un sistema que les permita responder a estas preguntas clave, así como evaluar su comportamiento. Esto no es algo habitual

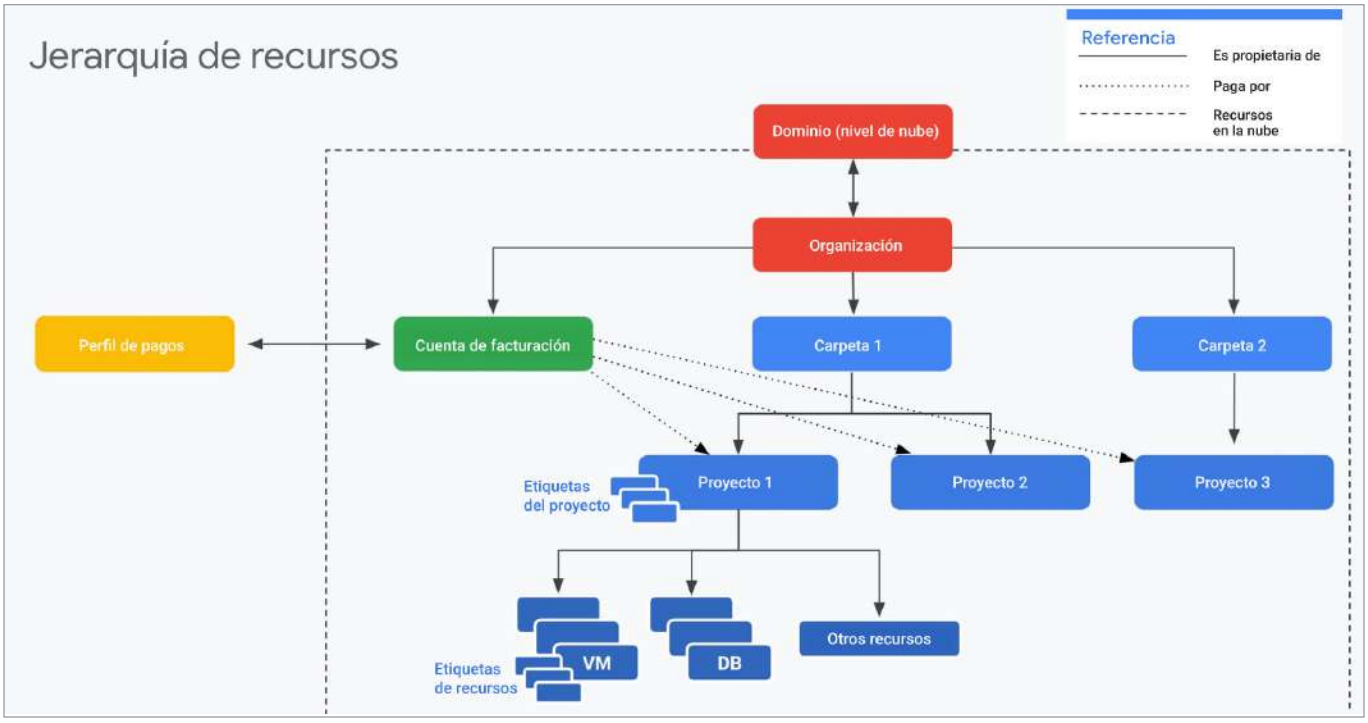
en los clientes que están dando sus primeros pasos, pero muchos de los clientes que ya han madurado están usando estos conceptos a medida que diseñan su sistema para el futuro.

Implementar procesos estandarizados desde el principio

Asegurarte de que estás implementando estas recomendaciones de manera uniforme es algo que debe diseñarse y aplicarse de forma sistemática. Herramientas de automatización como Terraform y Cloud Deployment Manager pueden ayudarte a crear protecciones antes de desplegar un recurso de nube. Implementar un estándar de forma retroactiva es mucho más difícil. Hemos visto de todo, desde departamentos de operaciones de TI que desactivaban o amenazaban con desactivar recursos no etiquetados hasta «muros de la vergüenza» para quienes no respetaban los estándares. (Apoyamos el refuerzo positivo, como regalar una pizza o dar un trofeo, o incluso dar un trofeo hecho de pizza).

¿Cuál sería un ejemplo de un proceso de optimización que sería recomendable estandarizar desde el principio? En primer lugar, desplegar recursos. ¿Cualquier ingeniero debería ser capaz de desplegar cualquier cantidad de cualquier recurso? Probablemente no. Esta es un área en la que crear un estándar desde el principio puede hacer una gran diferencia.

Estructurar tus recursos para una administración eficaz de los costos también es importante. Lo más recomendable es adoptar la estructura más simple que satisfaga tus requisitos iniciales y, luego, ajustar tu jerarquía de recursos a medida que evolucionen tus necesidades. Puedes usar el asistente de configuración que te guiará a través de las recomendaciones y los pasos necesarios para crear tu entorno óptimo. Dentro de esta jerarquía de recursos, puedes usar proyectos, carpetas y etiquetas para crear agrupamientos lógicos de recursos que cumplan con tus requisitos de administración y atribución de costos.

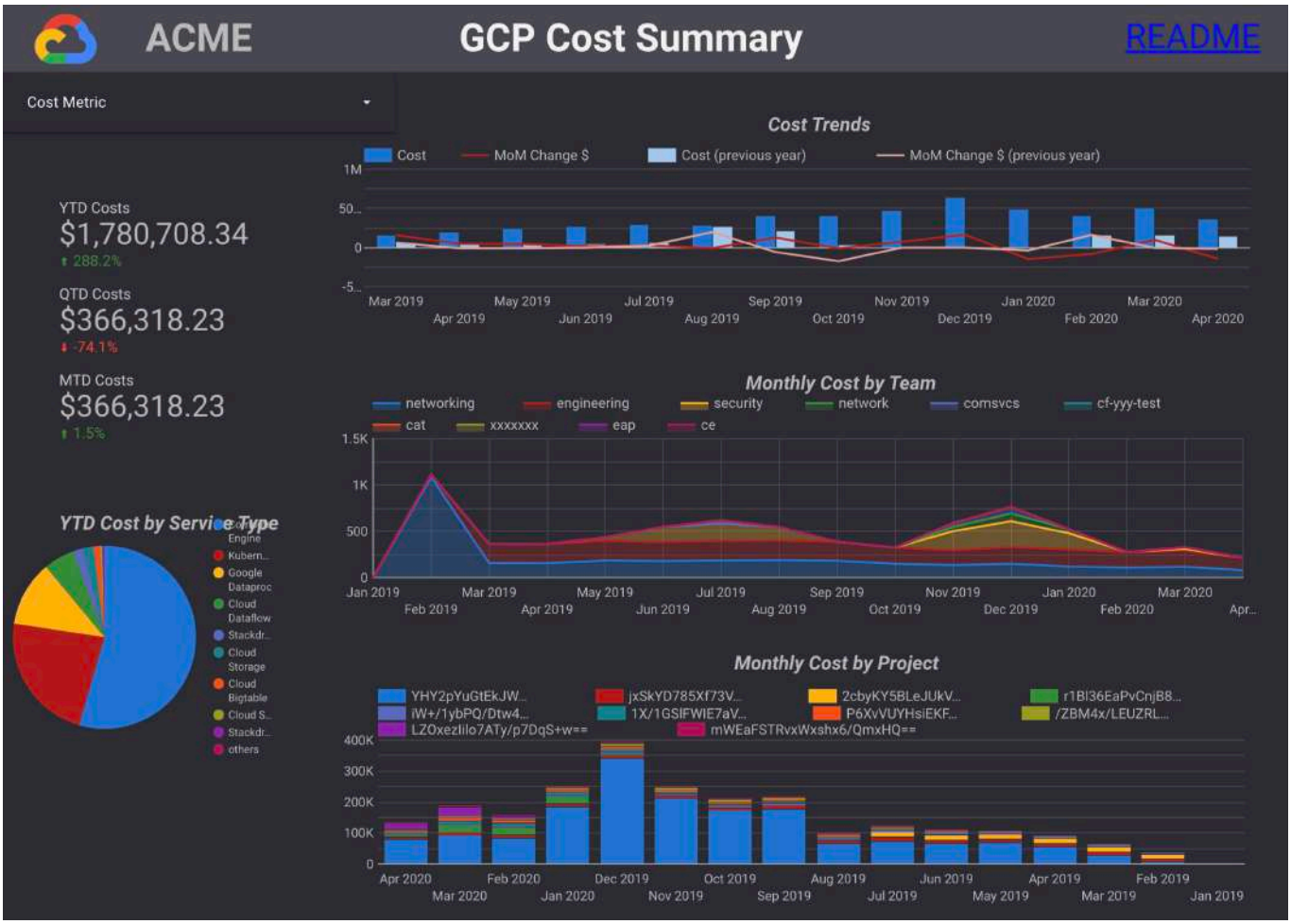


Ejemplo de una jerarquía de recursos para la nube

En tu jerarquía de recursos, etiquetar los recursos es una de las principales prioridades para las organizaciones interesadas en administrar los costos. Esta es en esencia tu capacidad de atribuir costos a una empresa, un servicio, una unidad, un líder, etc. específicos. Sin etiquetar los recursos, es increíblemente difícil descifrar cuáles son los costos de hacer cualquier cosa específica. En lugar de decir que gastaste USD36,000 en Compute Engine, es preferible poder decir que gastaste USD36,000 en brindarle memes a 400,000 usuarios el mes pasado. La segunda afirmación es mucho más detallada que la primera. Recomendamos crear etiquetas estandarizadas en conjunto con los equipos de ingeniería y finanzas, y usar etiquetas para tantos recursos como puedas

Revisar y repetir para tener los mejores resultados

Como práctica general, deberías reunirte regularmente con los equipos correspondientes para revisar las tendencias de uso y ajustar también la proyección en la medida en que sea necesario. La consola de facturación de Cloud hace que sea muy fácil ver y auditar tus gastos de nube de forma regular, mientras que los paneles personalizables proporcionan una visualización más detallada de los costos. Sin revisiones regulares y un modelo de gastos e ingresos por unidad apropiado, y sin tener visibilidad de tus gastos, es difícil actuar de otra forma que no sea reactiva cuando observas un aumento en tu factura.



Visualiza el gasto a lo largo del tiempo con Google Data Studio

Si eres un cliente estable, puedes revisar tus gastos con menor frecuencia, dado que las oportunidades para ajustar tus estrategias dependerán de elementos como nuevos atributos de Google Cloud, en lugar de un cambio empresarial en la hoja de ruta de tus productos. Pero si vas a desplegar muchas aplicaciones nuevas y gastar millones de dólares por mes, hacer una pequeña inversión en revisiones de costos más frecuentes puede generar un gran ahorro en un corto período de tiempo. En algunos casos, nuestros clientes más avanzados alcanzan y ajustan las proyecciones incluso a diario. Cuando estás gastando millones de dólares por mes, tan solo un pequeño porcentaje mayor en tu factura total puede desfinanciar proyectos como experimentación con nuevas tecnologías o contratación de ingenieros adicionales.

Para funcionar verdaderamente de forma eficiente y maximizar el valor de la nube, se necesitan varios equipos con diversas especializaciones que trabajen juntos para diseñar un sistema que satisfaga tus necesidades empresariales específicas. Las buenas prácticas incluyen establecer una cadencia de revisiones basada en la velocidad a la que estás desarrollando y gastando en la nube. El triángulo de la

gestión de proyectos es un marco de trabajo que se usa con frecuencia y que mide el costo versus la velocidad versus la calidad. Puedes trabajar con tus equipos para acordar un marco de trabajo que funcione para tu empresa. A partir de ese marco, puedes apretar el cinturón o invertir más.

Las herramientas del oficio de optimizar costos

Una vez que tengas una noción firme sobre cómo abordar la optimización de costos en la nube, es hora de pensar en las diversas herramientas que tienes a disposición. En un nivel alto, la administración de costos en Google Cloud depende de tres tipos amplios de herramientas.

1. **Visibilidad de los costos**: esto incluye saber en detalle cuánto gastas, cómo se facturan los servicios específicos y la capacidad de mostrar cómo (o por qué) gastaste una cantidad determinada para alcanzar un resultado empresarial. En este punto, ten en cuenta las capacidades clave, como la posibilidad de crear una responsabilidad compartida, realizar revisiones frecuentes de costos, analizar tendencias y visualizar el impacto de tus acciones con una frecuencia semejante al tiempo real. Con una estrategia estandarizada para organizar tus recursos, puedes mapear correctamente tus costos respecto de la estructura operativa de tu organización para crear un modelo de análisis de costos (showback)/facturación de costos a la unidad empresarial que usa los recursos (chargeback). También puedes usar controles de costos como alertas de presupuesto y cuotas para monitorear los costos en el tiempo.

2. **Optimización del uso de los recursos**:reducir los costos innecesarios en tu entorno mediante la optimización del uso. El objetivo es implementar un conjunto específico de estándares que genere una intersección adecuada entre costo y rendimiento dentro de un entorno. Es desde esta perspectiva que hay que revisar si existen recursos ociosos, mejores servicios en los cuales desplegar una aplicación, o incluso decidir si lanzar una forma personalizada de VM sería más apropiado. La mayoría de las compañías que tienen

éxito a la hora de evitar costos innecesarios están optimizando el uso de recursos de forma descentralizada, ya que los responsables de las aplicaciones individuales suelen ser quienes están mejor equipados para desactivar o cambiar el tamaño de los recursos por estar íntimamente familiarizados con las cargas de trabajo. Además, puedes usar un Recomendador que te ayude a detectar problemas como instancias de VM subaprovisionadas o sobreaprovisionadas, o recursos ociosos. Permitir que tu equipo haga estas recomendaciones de forma automática es el objetivo de cualquier esfuerzo importante de optimización.

3. **Eficiencia de precios**: esto incluye capacidades como descuentos por uso continuo, descuentos por compromiso de uso, precios fijos, facturación por segundo u otros atributos de descuento por volumen que te permiten optimizar las tarifas correspondientes a un servicio específico. Los equipos más centralizados de tu empresa, como un Centro de excelencia para la nube (CCoE) o un equipo de FinOps, son los que mejor aprovechan estas funcionalidades, ya que pueden reducir los posibles costos innecesarios y, a su vez, optimizar la cobertura de todas las unidades de negocios. Esto es algo que se debe seguir analizando regularmente, tanto antes de la migración a la nube como después.

Tener en cuenta a las personas y los procesos será de gran ayuda para asegurarte de que tus estándares son útiles y están alineados con lo que tu empresa necesita. Además, entender los atributos de visibilidad de costos, optimización de uso de recursos y eficiencia de precios de Google Cloud te brindará las herramientas que necesitas para optimizar los costos en todas tus tecnologías y equipos

Cómo priorizar las recomendaciones

Con muchas iniciativas que compiten entre sí, puede resultar difícil priorizar las recomendaciones de optimización de costos y asegurarte de que tu organización está dedicando el tiempo

suficiente al análisis de estos esfuerzos de forma constante. Tener visibilidad sobre la cantidad de esfuerzo de ingeniería, así como el posible ahorro de costos, puede ayudar a tu equipo a establecer sus prioridades. Algunos clientes se concentran únicamente en la innovación y velocidad de migración durante años y, con el transcurso del tiempo, sus malos hábitos de optimización se agravan, lo que conduce a costos innecesarios aún mayores. Esos fondos podrían haberse usado para desarrollar nuevos atributos, comprar infraestructura adicional o contratar a más ingenieros para incrementar la velocidad de desarrollo de atributos. Es importante hallar un equilibrio entre costo y velocidad, y comprender cuáles son las posibles ramificaciones de ir demasiado en una dirección en desmedro de otra.

Para ayudarte a priorizar una recomendación de optimización de costos sobre otra, etiquetamos cada recomendación que realizamos con nuestros clientes con dos características:

- **Esfuerzo**: nivel estimado de trabajo (en semanas) que necesita el cliente para coordinar los recursos e implementar una recomendación de optimización de costos.
- **Ahorro**: monto del ahorro potencial previsto (en porcentaje por servicio) que obtendrán los clientes si implementan una recomendación de optimización de costos.

Esfuerzo	Duración	Ahorro	% por servicio
Bajo	Hasta 2 semanas	Bajo	0-10%
Medio	2-6 semanas	Medio	10-20%
Alto	Más de 6 semanas	Alto	20% o más

Aunque no siempre es posible calcular con una precisión milimétrica cuánto te ahorrará en costos una medida antes de probarla, es importante intentar realizar una estimación informada para cada esfuerzo. Por ejemplo, saber que un determinado cambio podría potencialmente ahorrarte 60% en Cloud Storage para el proyecto X debería ser suficiente para ayudar con la matriz de priorización y establecer prioridades de ingeniería en tu equipo. A veces, puedes calcular el ahorro real. Especialmente con las opciones de compra, un equipo de FinOps puede calcular el potencial ahorro aprovechando atributos como descuentos por compromiso de uso para una cantidad específica de su infraestructura. Al realizar este ejercicio, es recomendable que el equipo sea capaz de tomar decisiones informadas sobre el rumbo del equipo de ingeniería, para que pueda concentrar su energía desde el punto de vista de la cultura.

Capítulo 2
Optimización de los costos de procesamiento

Los tres principios de optimización de costos que expusimos en la introducción (visibilidad de costos, optimización de uso y eficiencia en la fijación de precios) pueden aplicarse a la hora de evaluar y establecer estándares para diversos aspectos de tu infraestructura de nube. Comenzaremos con el procesamiento, la espina dorsal de la infraestructura de nube.

Cuando los clientes migran a Google Cloud, su primer paso suele ser adoptar Compute Engine, ya que facilita la adquisición y configuración de máquinas virtuales (VM) en la nube que proporcionan una cantidad enorme de capacidad de procesamiento. Compute Engine, lanzado en 2012, ofrece varios tipos de máquinas, muchos atributos innovadores y, al momento de esta publicación, está disponible en 23 regiones y 70 zonas.

Los tipos de máquinas predefinidas y personalizadas de Compute Engine hacen que sea fácil escoger las VM que estén más cerca de tu infraestructura local, lo cual acelera el proceso de migración de cargas de trabajo de forma rentable. Cloud te da la ventaja de "pagar a medida que usas" y también proporciona un ahorro significativo cuanto más procesamiento uses con descuentos por uso continuo. Aquí encontrarás algunas recomendaciones basadas en nuestra experiencia de trabajar con clientes empresariales para analizar su gasto mensual y detectar oportunidades de optimización

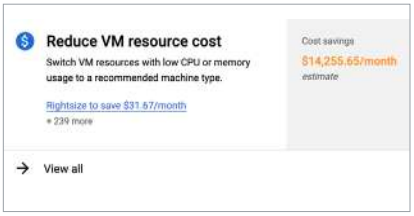
Prepararse para ahorrar

Si buscas optimizar, deberías familiarizarte con la página de precios de instancias de VM, que es de lectura obligatoria para quien necesite entender en detalle el modelo de facturación y los precios basados en

recursos de Compute Engine. Esto será de gran ayuda en el camino hacia una mejora en la visibilidad de tus costos.

El siguiente paso importante para conocer mejor tus costos de Compute Engine es usar informes de Facturación de Cloud en Google Cloud Console, y personalizar tus vistas con filtros y agrupamientos por proyectos, etiquetas, y mucho más. Aquí encontrarás información sobre los diversos tipos de máquinas de Compute Engine, los descuentos por compromiso de uso y cómo ver tu uso, entre otras cosas. Si necesitas una vista avanzada, puedes incluso exportar los detalles de uso de Compute Engine a BigQuery para realizar un análisis más detallado. Esto te permite consultar el almacén de datos para comprender las tendencias de uso de vCPU de tu proyecto y cuántas vCPU se pueden aprovechar. Si tienes umbrales definidos para la cantidad de núcleos por proyecto, las tendencias de uso pueden ayudarte a detectar anomalías y actuar de forma proactiva. Actuar puede significar ajustar el tamaño de las instancias de VM o aprovechar las VM que no están en uso.

Con una mejor visibilidad de costos disponible, repasemos las cinco formas en las que puedes optimizar tus recursos de Compute Engine que creemos ofrecerán el beneficio más inmediato.



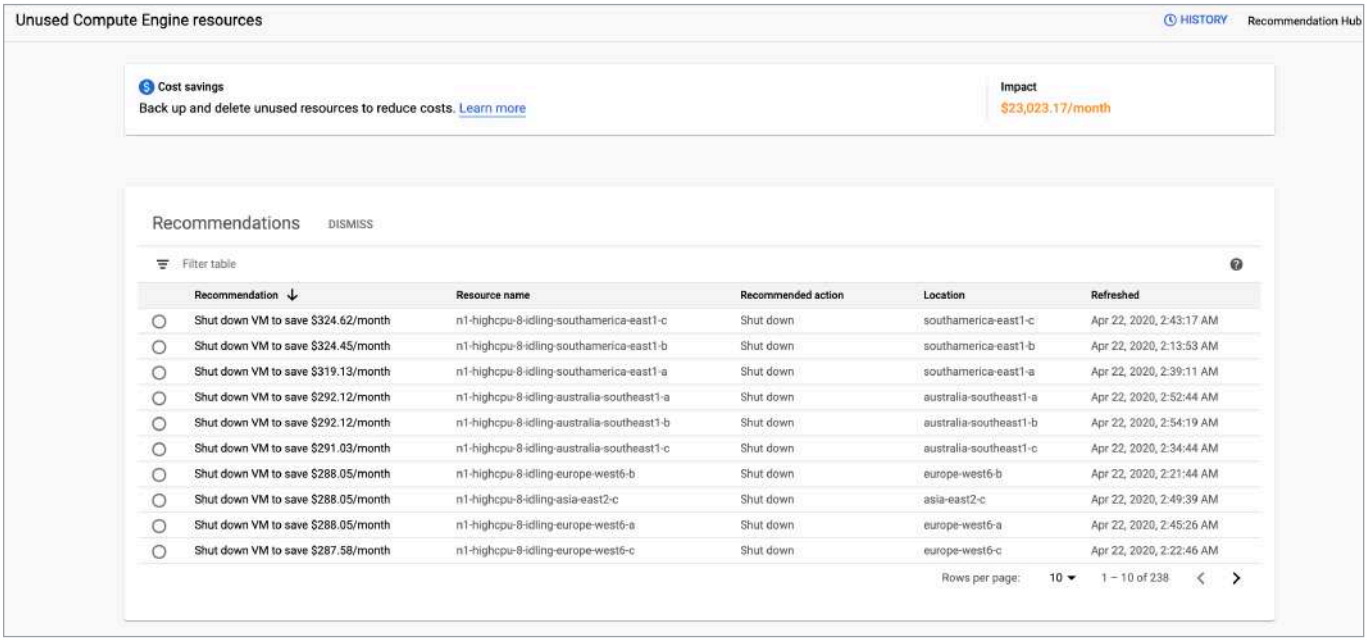
1. Paga solo por el procesamiento que necesitas

Esfuerzo ●●○ **Ahorro** ●●●

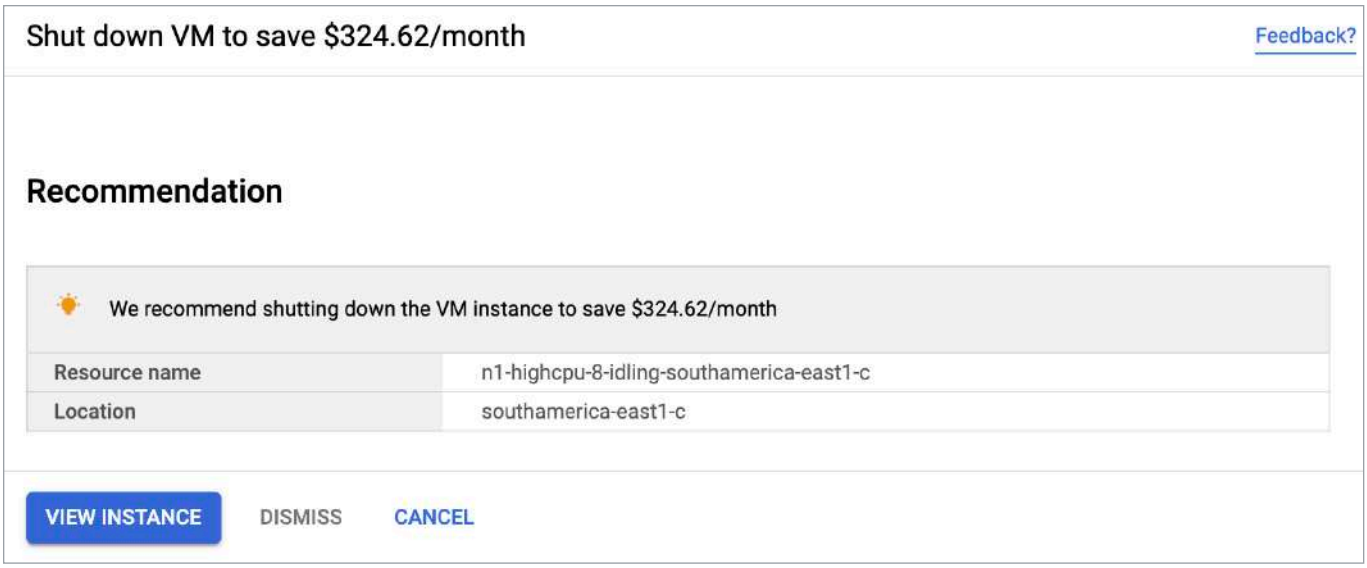
Identifica las VM (y los discos) que no estás usando: La forma más fácil de reducir tu factura de Google Cloud es deshacerte de los recursos que ya no estás usando. Piensa en esos proyectos de prueba de concepto que ya no son prioridad o los procesos zombis que nadie se ocupó de borrar. Google Cloud ofrece varios Recomendadores que pueden ayudarte a optimizar estos recursos, incluso un recomendador de VM inactivas que identifica los discos persistentes y las VM sin usar con base en las métricas de uso.

No obstante, ten siempre mucho cuidado a la hora de eliminar una VM. Antes de eliminar un recurso, pregúntate, "¿Qué impacto tendrá el

hecho de borrar este recurso? ¿Cómo puedo recrearlo, si fuera necesario?". Al borrar instancias te deshaces del disco o los discos subyacentes y todos sus datos. Una buena práctica es tomar una instantánea de la instancia antes de eliminarla. Como alternativa, puedes elegir simplemente detener la VM, lo que finaliza la instancia, pero conserva recursos como discos o direcciones IP hasta que los separes o elimines.



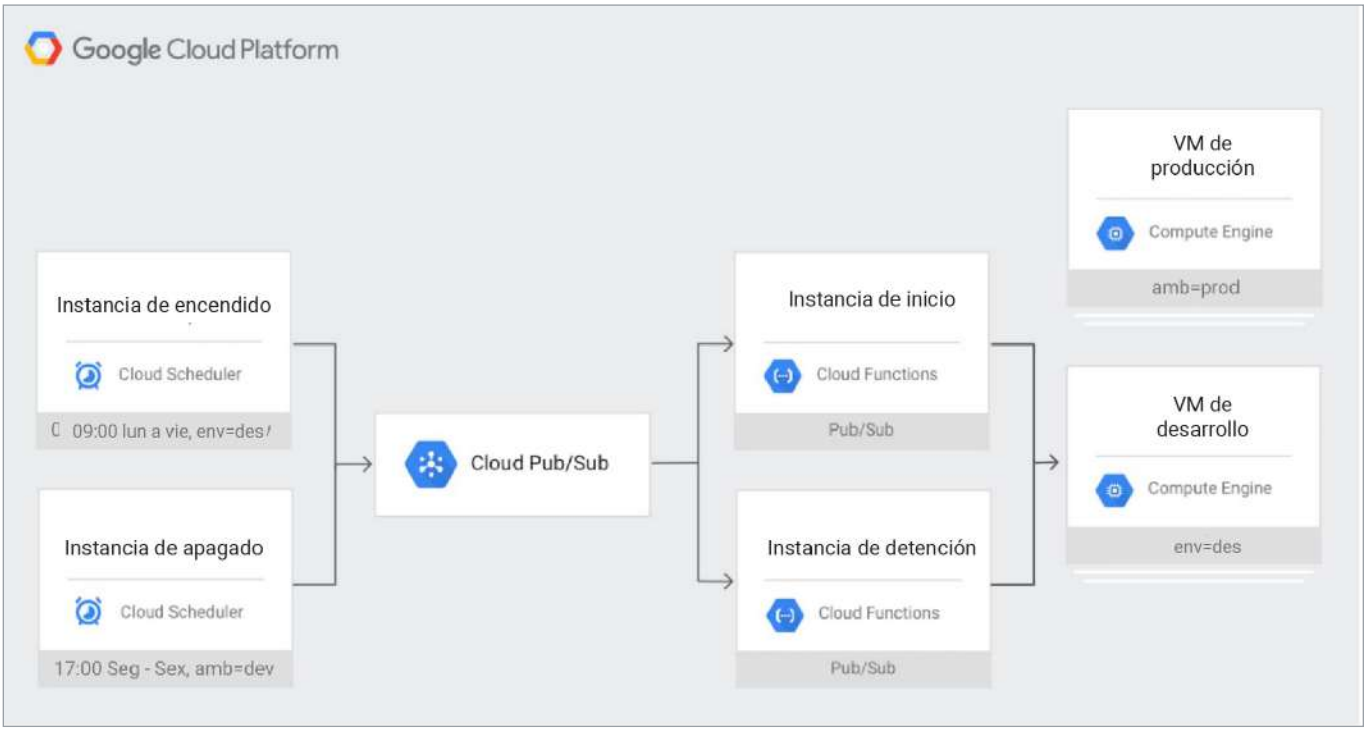
Los Recomendadores de Google Cloud ofrecen orientación específica sobre un potencial ahorro de costos.



Los Recomendadores pueden ofrecer consejos detallados sobre cómo apagar VM.

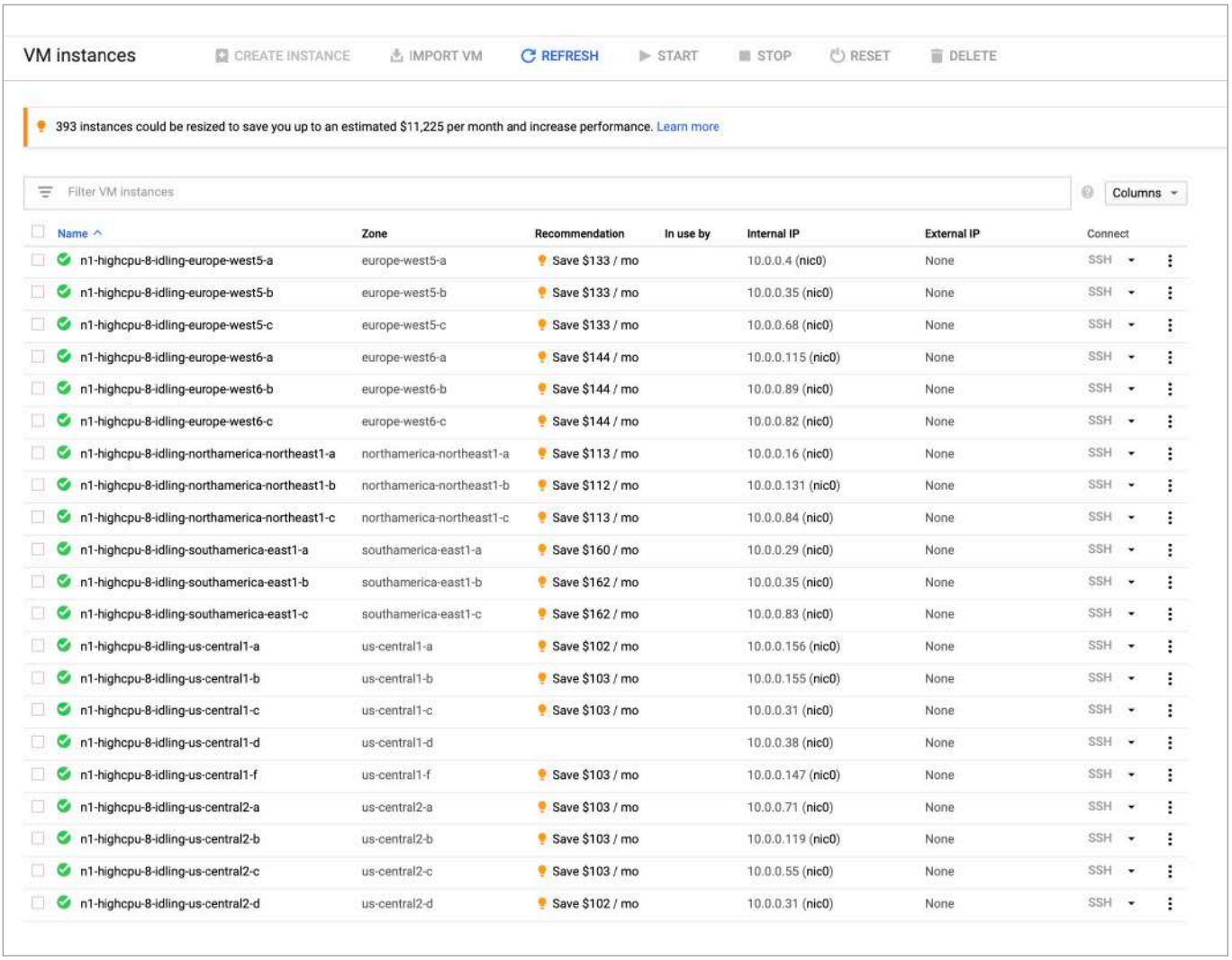
Para obtener más información, consulta la Documentación sobre Recomendadores. Estate atento, ya que agregamos recomendadores en función del uso.

Programa VM para que se inicien y detengan automáticamente: El beneficio de una plataforma como Compute Engine es que solo pagas por los recursos de procesamiento que usas. Los sistemas de producción tienden a funcionar las 24 horas, los siete días de la semana, mientras que las VM en entornos de desarrollo, prueba o personales tienden a usarse solo en horario comercial. ¡Apagarlas puede ahorrarte mucho dinero! Por ejemplo, ejecutar una VM que está encendida durante 10 horas por día, de lunes a viernes, cuesta 75% menos por mes en comparación con dejarla encendida todo el tiempo.



Vale la pena explorar cómo se configuran y ajustan los tamaños de las VM en toda tu infraestructura de nube para detectar un ahorro de costos

Ajustar el tamaño de las instancias de VM: En Google Cloud, puedes generar ahorros significativos con tan solo crear tipos de máquinas personalizadas con la cantidad de CPU y RAM necesaria para satisfacer tus necesidades. Sin embargo, las necesidades de carga de trabajo pueden cambiar a lo largo del tiempo. Las instancias que en algún momento estuvieron optimizadas ahora podrían tener menor cantidad de usuarios y tráfico. Nuestras recomendaciones de tamaño para instancias de VM pueden ayudarte mostrándote cómo reducir el tamaño de tu tipo de máquina de forma eficaz con base en cambios en el uso de vCPU y RAM. Estas recomendaciones de ajuste de tamaño para el tipo de máquina de tu instancia (o grupo administrado de instancias) se generan gracias a las métricas del sistema recopiladas por Cloud Monitoring durante los ocho días anteriores.



Explora las recomendaciones de ajuste de tamaño para VM en la nube.

Si tu organización usa la infraestructura como código para administrar tu entorno, consulta esta guía, que te mostrará cómo desplegar las recomendaciones de ajuste de tamaño de VM a escala.

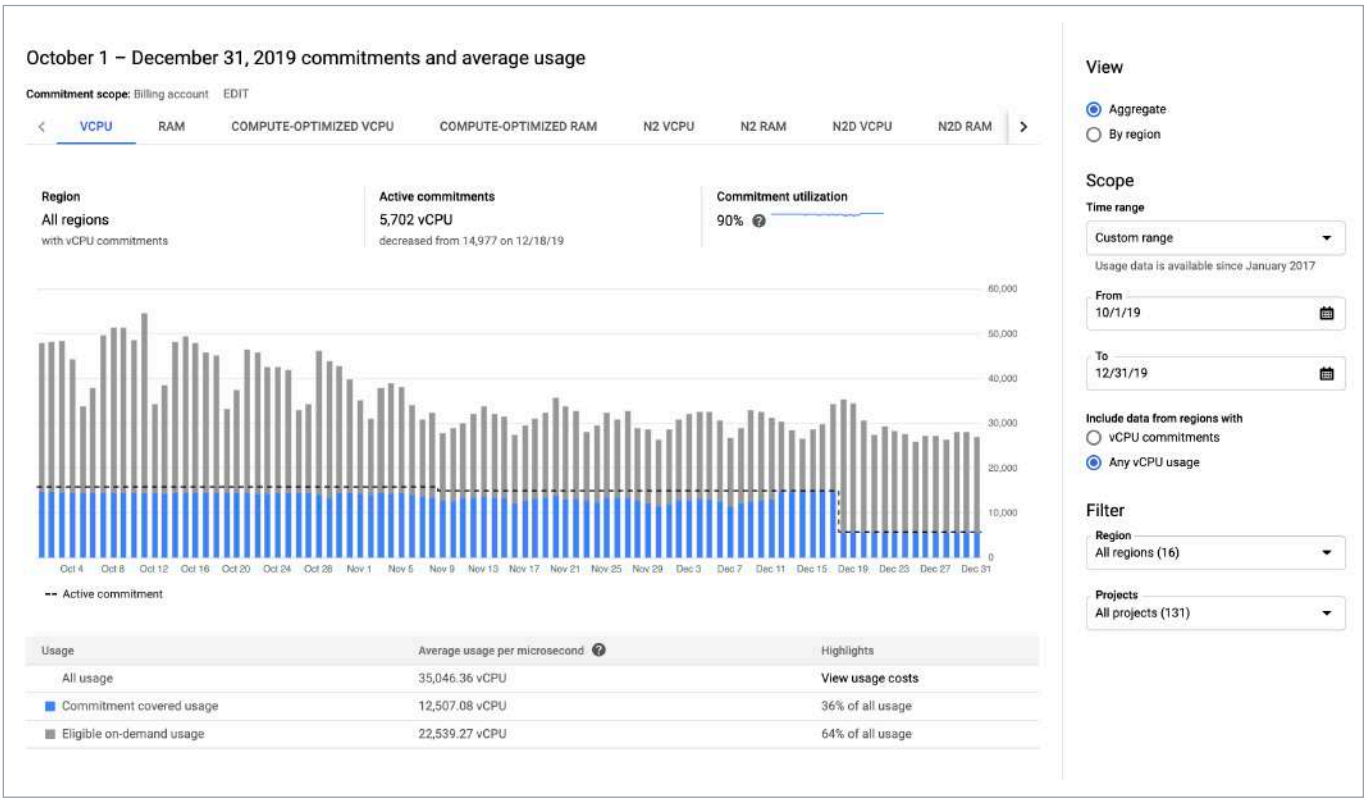
2. Compra compromisos

Esfuerzo ●○○ **Ahorro** ●●●

La eficiencia de precios es otro concepto clave para emplear como parte de un esfuerzo de optimización de la nube. La capacidad de comprar compromisos y recibir los descuentos relacionados es un paso enorme hacia el objetivo de garantizar que realmente estás optimizando tus costos de nube.

Nuestros clientes tienen cargas de trabajo diversas ejecutándose en Google Cloud, con requisitos de disponibilidad variados. Muchos clientes siguen una regla de 70/30 para administrar su flota de VM (tienen un uso constante durante todo el año de aproximadamente un 70% y un pico de temporada de aproximadamente un 30% durante los feriados y eventos especiales).

Si esto te suena familiar, probablemente estés aprovisionando recursos para la capacidad pico. Sin embargo, después de migrar a Google Cloud, puedes establecer los valores de referencia de tu uso y aprovechar descuentos mayores para las cargas de trabajo de procesamiento. Los descuentos por compromiso de uso son ideales si tienes una carga de trabajo de estado estable y predecible, ya que puedes adquirir un compromiso de un año o tres años a cambio de un descuento sustancial en tu uso de VM. Hace poco tiempo, publicamos un informe de análisis sobre los descuentos por compromiso de uso en Cloud Console que te ayudará a entender y analizar la eficacia de los compromisos que has adquirido e incluso calcular cuál sería tu piso de recursos según datos históricos.



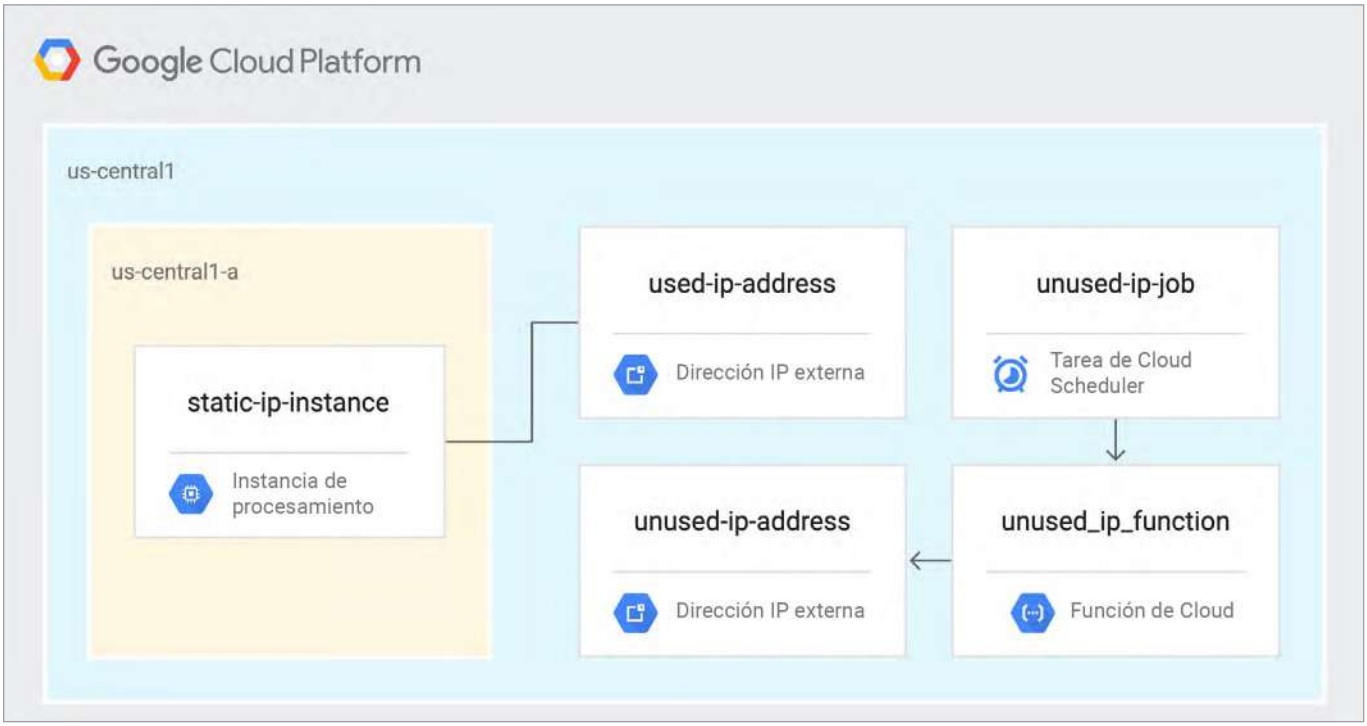
Los informes sobre los descuento por compromiso de uso pueden mostrar los posibles ahorros.

Además, conversa sobre el uso con tu equipo interno y evalúa si un descuento por compromiso de uso tiene sentido o no para tu carga de trabajo. Puedes trabajar de forma proactiva con tu equipo para incrementar la cobertura de tu descuento por compromiso de uso y maximizar tu ahorro.

3. Automatiza las optimizaciones de costos

Esfuerzo ●●● **Ahorro** ●●●

La mejor forma de asegurarte de que tu equipo siga siempre las buenas prácticas de optimización de costos es automatizarlas, lo que reduce la intervención manual. La automatización se simplifica en gran medida con el uso de una etiqueta (un par de clave-valor aplicado a varios servicios de Google Cloud). Por ejemplo, podrías etiquetar instancias que solo usan los desarrolladores durante el horario comercial con "env: desarrollo". Luego, podrías usar Cloud Scheduler para programar una función de Cloud sin servidores que las apague durante el fin de semana o después del horario comercial, y las vuelva a encender cuando sean necesarias. Usa este diagrama de arquitectura y ejemplos de código para hacerlo por tu cuenta. Este es un enorme salto hacia adelante en tu capacidad de optimizar el uso de tus recursos



El uso de etiquetas puede automatizar las optimizaciones de costos.

Usar Cloud Functions para automatizar la limpieza de otros recursos de Compute Engine también puede ahorrarle tiempo y dinero a tu equipo de ingeniería. Por ejemplo, los clientes a menudo olvidan los discos persistentes desconectados (huérfanos) o las direcciones IP sin uso. Esto genera gastos, incluso aunque no estén conectados a una instancia de máquina virtual. Las VM con la opción de regla de eliminación configurada en "mantener disco" retendrán los discos persistentes incluso después de que se haya eliminado la VM. Eso es muy bueno si necesitas guardar los datos en ese disco para otro momento, pero esos discos persistentes huérfanos pueden acumularse

con rapidez. Este artículo sobre las soluciones de Google Cloud describe la arquitectura y el código de ejemplo necesarios para que Cloud Functions, Cloud Scheduler y Cloud Monitoring rastreen automáticamente estos discos huérfanos, hagan una instantánea y los eliminen. Esta solución se puede usar como modelo para otras automatizaciones de costos, como la limpieza de direcciones IP sin usar o la detención de VM inactivas.

4. Usa VM interrumpibles

Esfuerzo ●●● **Ahorro** ●●●

Si tienes cargas de trabajo tolerantes a errores, como HPC, Big Data, transcodificación de medios, canalizaciones de IC/EC o aplicaciones web sin estado, usar VM interrumpibles para procesarlas por lotes puede generar un ahorro enorme de costos, mediante estrategias de optimización de uso de los recursos y eficiencia de fijación de precios. Por ejemplo, nuestro cliente Descartes Labs redujo sus costos de análisis en más del 70% al usar VM interrumpibles para procesar las imágenes satelitales y ayudar a las empresas y los gobiernos a predecir los suministros de alimentos a nivel global.

Las VM interrumpibles tienen un ciclo de vida corto. Funcionan durante un máximo de 24 horas y también podrían apagarse antes de que se hayan cumplido las 24 horas. Se envía un aviso de interrupción de 30 segundos a la instancia cuando se necesita reclamar la VM, y puedes usar un script de apagado para limpiar durante ese período de 30 segundos. Asegúrate de leer la lista completa de estipulaciones cuando consideres las VM interrumpibles para tu carga de trabajo. Todos los tipos de máquinas están disponibles como VM interrumpibles, y puedes lanzar una simplemente si agregas la línea de comando de gcloud "-preemptible" o seleccionas la opción desde Cloud Console.

Usar las VM interrumpibles en tu arquitectura es una excelente forma de escalar el procesamiento a una tarifa con descuento, pero deberás asegurarte de que la carga de trabajo pueda tolerar las posibles interrupciones si es necesario reclamarlas. Una forma de hacerlo es asegurarte de que tu aplicación está realizando puntos de control a medida que procesa los datos (es decir, que está escribiendo en un almacenamiento fuera de la VM en sí, como Google Cloud Storage o una base de datos). Para probarlo, usa este código de ejemplo para usar un script de apagado y escribir un archivo de punto de control en un depósito de Cloud Storage. Para las aplicaciones web detrás de un balanceador de cargas, considera usar un aviso de interrupción de 30 segundos para drenar las conexiones a esa VM de modo tal que el tráfico pueda redireccionarse a otra VM. Algunos clientes también eligen automatizar el apagado de VM interrumpibles de forma periódica antes de que acabe el plazo de 24 horas para evitar que varias VM se apaguen al mismo tiempo si se lanzaron en el mismo momento.

No tienes por qué limitar las VM interrumpibles a un entorno de Compute Engine. Las GPU, los clústeres de GKE y las instancias secundarias en Dataproc también pueden usar VM interrumpibles. Además, puedes reducir tus costos de análisis por lotes de Cloud Dataflow si usas una programación flexible de recursos para complementar las instancias regulares con VM interrumpibles.

5. Prueba el ajuste de escala automático

Esfuerzo ●○○ **Ahorro** ●●○

Otra buena forma de ahorrar costos es ejecutar solamente la capacidad que necesitas cuando la necesitas. Este es uno de los principios clave de la optimización de uso de los recursos. Como mencionamos anteriormente, por lo general, se necesita alrededor de un 70% de la capacidad para el uso de estado estable, pero cuando necesitas capacidad adicional, es fundamental tenerla a disposición.

En un entorno on-premise, debes adquirir esa capacidad extra con antelación, esperar a que te la envíen y luego configurarla correctamente. En la nube, puedes usar el ajuste de escala automático para pasar automáticamente a una capacidad superior solo cuando la necesites. Los grupos de instancias administradas de Compute Engine son los que te proporcionan esta capacidad de ajuste de escala automático en Google Cloud. Puedes subir la escala sin problemas para manejar un aumento en el tráfico, y luego reducirla cuando se necesiten menos instancias (downscaling). Puedes escalar con base en el uso de CPU, la capacidad de balanceo de cargas de HTTP o las métricas de Cloud Monitoring. Esto te brinda flexibilidad para escalar en función de lo que sea más importante para tu aplicación.

Los costos altos no tienen lógica

Como vimos anteriormente, hay muchas formas de optimizar tus costos de Compute Engine. Supervisar tu entorno y entender tus patrones de uso es clave para comprender las mejores opciones que debes considerar primero, tomándote el tiempo para simular tus costos de referencia con antelación. Luego, hay una amplia variedad de estrategias que puedes implementar según tu carga de trabajo y modelo operativo actual.

Capítulo 3
Optimización de los costos de almacenamiento

Tener un servicio de almacenamiento ilimitado y escalable significa que habrá algunos depósitos y objetos que fácilmente podrías guardar, aunque no deberías hacerlo.

Da lo mismo si eres parte de un conglomerado de varios miles de millones de dólares que intenta revisar los datos de ventas de la primera mitad del año o simplemente estás intentando subir un video de tu gato tocando el piano: en cualquiera de los dos casos, necesitas un lugar para almacenar esos datos.

Para los clientes de Google Cloud, ese lugar suele ser Cloud Storage, un almacén unificado de objetos. Cloud Storage cuenta con una API sólida que te permite integrarlo con una gran variedad de servicios. También incluye atributos que te ayudan a seguir los principios de la optimización de costos: optimización de uso, visibilidad de costos y eficiencia de precios. Aunque almacenar un objeto en la nube es una tarea sencilla, asegurarte de estar aplicando el enfoque más sensato a la situación en la que estás requiere un poco más de previsión.

Uno de los beneficios de tener un servicio de almacenamiento escalable e ilimitado es que, tal como sucedería si tuvieras un desván en tu casa que pudiera aumentar de tamaño infinitamente, habrá algunas cajas y elementos (o depósitos y objetos) que podrías guardar con facilidad, aunque no deberías hacerlo. Con el paso del tiempo, almacenar estos elementos genera un costo y, más allá de si los necesitas a los fines empresariales o si solo los estás guardando por la remota posibilidad de que algún día puedan ser útiles (como esa espada que te encanta), el primer paso es crear una práctica en torno a cómo identificar la utilidad de un objeto o depósito para tu empresa. ¡Así que toma la escoba, la pala, y a trabajar!

Limpiar tu almacenamiento antes de migrar a la nube

Hay varios factores que debes considerar a la hora de analizar la optimización de los costos de almacenamiento. El secreto es asegurarte de no generar un impacto negativo en el rendimiento ni

descartar nada que debas conservar para el futuro, ya sea por motivos de cumplimiento, razones legales o simplemente por su valor empresarial. Ahora que los datos emergen como una de las principales materias primas de las empresas, es recomendable que uses las clases correctas de almacenamiento en el corto plazo, y también para un análisis longitudinal. Cloud Storage ofrece una gran variedad de clases de almacenamiento para elegir, con diversos niveles de costos, durabilidad y solidez.

Cuando se trata de la arquitectura en la nube, rara vez hay un enfoque universal. No obstante, hay algunos temas recurrentes que hemos observado a medida que trabajamos a la par de nuestros clientes. Estas lecciones aprendidas pueden aplicarse a cualquier entorno, ya sea que estés almacenando imágenes o desarrollando modelos de aprendizaje automático avanzado.

Cuando se trata de la arquitectura en la nube, rara vez hay un enfoque universal.

El punto de partida natural es comprender primero "¿Qué me cuesta dinero?" al usar Cloud Storage. La página de precios es increíblemente útil, pero a la hora de analizar tu uso de Cloud Storage, también debes considerar lo siguiente:

1. Retención
2. Patrones de acceso
3. Rendimiento

Puede haber casos de uso adicionales que impliquen costos, pero nos concentraremos en las recomendaciones en torno a estos temas. A continuación, encontrarás más detalles sobre cada uno.

Consideraciones y recomendaciones con respecto a la retención

Esfuerzo ●○○ **Ahorro** ●●○

Lo primero que se debe considerar al ver un tipo de dato es su período de retención. Hacerte preguntas como “¿Por qué es valioso este objeto?” y “¿Durante cuánto tiempo esto será valioso?” es fundamental para ayudarte a determinar la política de ciclo de vida apropiada. Establecer una política de ciclo de vida te permite etiquetar objetos o depósitos específicos, y crea una regla automática que eliminará o incluso transformará las clases de

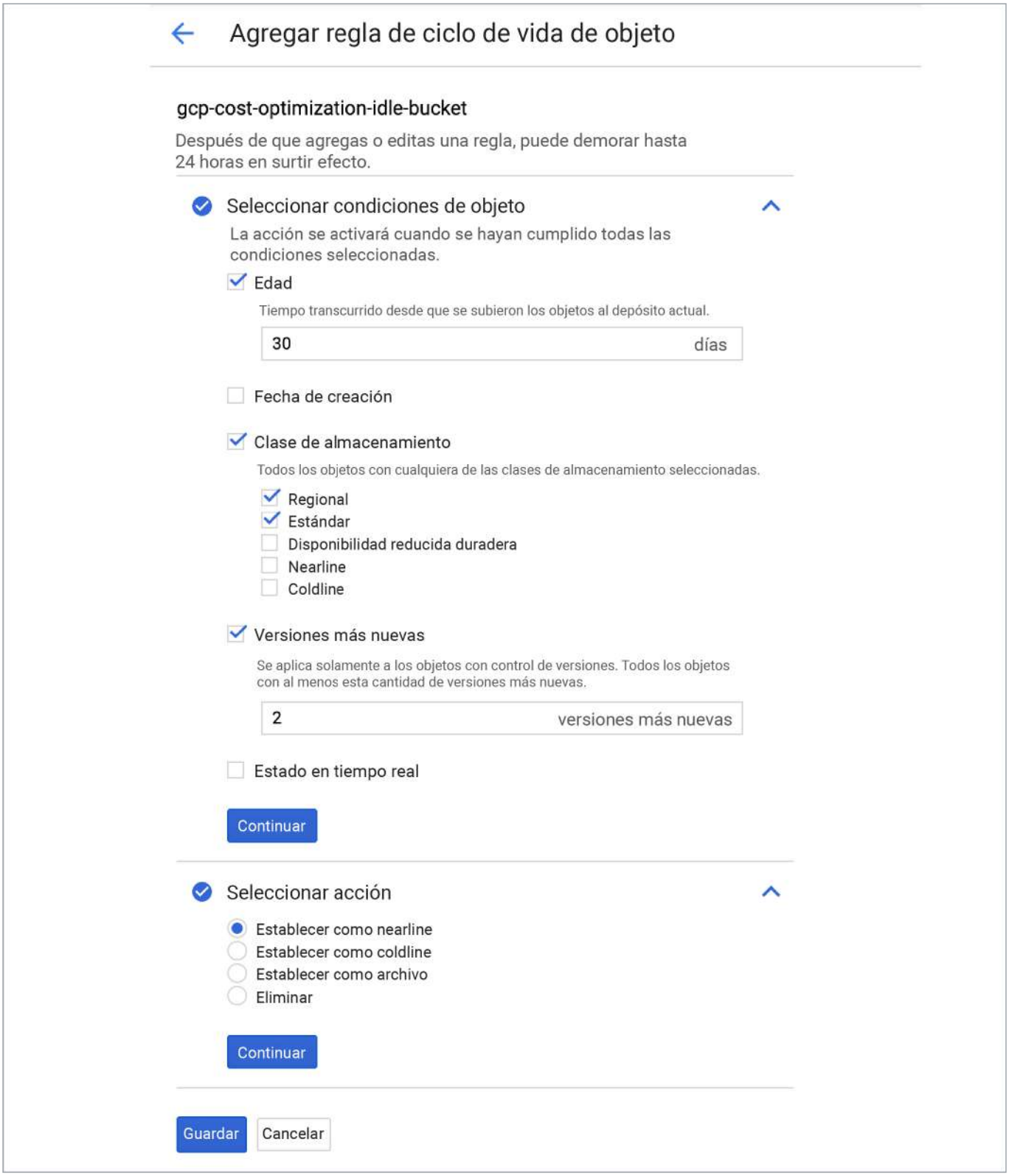
Lo primero que se debe considerar al ver un tipo de dato es su período de retención.

almacenamiento para ese tipo de depósito u objeto en particular con base en un conjunto de condiciones. Piensa en esto como tu propio mayordomo personal que se asegura sistemáticamente de que tu desván está limpio y ordenado (pero en lugar de gastar dinero, este mayordomo hace que ahorres dinero por estas operaciones).

Vemos cómo los clientes usan las políticas de ciclo de vida en una multiplicidad de formas con gran éxito. Un buen uso es el cumplimiento por motivos legales. Según tu industria y clase de datos, hay leyes que regulan los tipos de datos que deben retenerse y durante cuánto tiempo. Mediante una política de ciclo de vida de Cloud Storage, puedes etiquetar al instante un objeto para que sea eliminado cuando haya alcanzado el umbral mínimo para las necesidades de cumplimiento legal, con lo cual te aseguras de no pagar por su retención por más tiempo que el necesario y no tienes que recordar qué datos vencen cuándo.

Dentro de Cloud Storage, también puedes establecer políticas para transformar un tipo de almacenamiento en una clase diferente. Esto es particularmente útil para los datos a los que se accederá con relativa frecuencia durante un período corto de tiempo, pero a los que no se accederá demasiado en el largo plazo. Tal vez quieras retener estos objetos específicos durante más tiempo por motivos legales o de seguridad, o incluso por su valor comercial a largo plazo. Un buen lugar para ponerlo en práctica es un entorno de laboratorio. Una vez que completas un experimento, probablemente quieras analizar los resultados en profundidad en el corto plazo, pero que luego no accederás demasiado a esos datos en el largo plazo. Contar con una política para convertir este almacenamiento en clases de almacenamiento nearline y coldline luego de un mes es una excelente forma de ahorrar en los costos de datos que generan a largo plazo

Los clientes usan las políticas de ciclo de vida en una gran variedad de formas con mucho éxito.



Configura las políticas de Cloud Storage para mover datos automáticamente

Otra fuente común de gastos innecesarios en los entornos de almacenamiento son los datos duplicados. Por supuesto, hay ocasiones en que esos datos son necesarios. Por ejemplo, es recomendable que dupliques conjuntos de datos en varias regiones geográficas para que los equipos locales puedan acceder a ellos con rapidez. No obstante, en nuestra experiencia trabajando con los clientes, gran parte de los datos duplicados son el resultado de un control poco estricto de las versiones, y administrar los duplicados resultantes puede ser complicado y costoso.

Por suerte, hay muchas formas de prevenir los datos duplicados, así como herramientas para evitar que se eliminen datos por error. A continuación, presentamos algunas cuestiones que deberías considera:

- Si intentas mantener la resiliencia con una única fuente de verdad, es probable que tenga más sentido usar un depósito de multirregión en lugar de crear múltiples copias en varios depósitos. Con este atributo, obtienes redundancia geográfica para los objetos almacenados. Esto garantiza que tus datos se repliquen de modo asíncrono en dos o más ubicaciones, y brinda protección frente a fallas regionales en caso de un desastre natural.

- Una gran parte de los datos duplicados provienen de un uso incorrecto del atributo de control de versiones de objetos de Cloud Storage. El control de versiones de objetos evita que los datos se sobrescriban o eliminen accidentalmente, pero los duplicados que crea pueden acumularse. ¿Realmente necesitas cinco copias de tus datos? Una podría ser suficiente siempre y cuando esté protegida. ¿Te preocupa no poder volver a una versión anterior? Puedes configurar políticas de control de versiones de objetos para asegurarte de tener una cantidad adecuada de copias. ¿Aun así te preocupa perder algo accidentalmente? Considera usar el atributo de bloqueo de depósitos, que te asegura que no se eliminen los elementos antes de una fecha u hora específicas. Esto resulta realmente útil para demostrar el cumplimiento de varias regulaciones importantes. En

resumen, si implementas el control de versiones de objetos, hay varios atributos que puedes usar para mantener tus datos seguros sin desperdiciar espacio de forma innecesaria.

Consideraciones y recomendaciones con respecto a patrones de acceso

Esfuerzo ●●● **Ahorro** ●●●

Cloud Storage ofrece una gran variedad de clases de almacenamiento (standard, nearline, coldline y archive), todos con diversos costos y sus casos de uso más apropiados. Si solo usas la clase estándar, quizás sea hora de analizar tus cargas de trabajo y revaluar con qué frecuencia se está accediendo a tus datos. En nuestra experiencia, muchas compañías usan el almacenamiento de clase estándar como archivo, y podrían reducir sus gastos si aprovecharan el almacenamiento de las clases nearline o coldline. Y en algunos casos, si estás guardando objetos para casos de uso de almacenamiento en frío, como cuando se exige el almacenamiento para procesos judiciales, la clase de almacenamiento de archivo podría ofrecer incluso un ahorro mayor.

La capacidad para transformar objetos en clases de almacenamiento de bajo costo es una poderosa herramienta, pero debe usarse con cuidado. Aunque el almacenamiento a largo plazo es más económico de mantener para un objeto al cual se accede con una frecuencia menor, incurrirás en cargos adicionales si de pronto necesitas acceder con frecuencia a los datos o metadatos que se han migrado a una opción de almacenamiento más fría. También hay algunas implicancias con respecto al costo cuando necesitas eliminar los datos de una clase de almacenamiento particular. Por ejemplo, actualmente hay un plazo mínimo de 30 días para que un objeto permanezca en almacenamiento nearline. Si necesitas acceder a esos datos con una frecuencia mayor, puedes realizar una copia en una clase de almacenamiento regional para evitar cargos de acceso mayores.

Muchas compañías podrían reducir sus gastos si aprovecharan el almacenamiento nearline o coldline.

A la hora de considerar las oportunidades de ahorro de costos, también deberías pensar si se necesitará acceder a tus datos en el largo plazo, y con qué frecuencia se accederá a ellos si vuelven a ser

valiosos. Por ejemplo, si eres un CFO que analiza un informe trimestral sobre los gastos de nube y solo necesitas obtener esa información cada tres meses, quizás no debas preocuparte por el aumento en los cargos derivados de la recuperación de esos datos, ya que sigue siendo más económico que mantener el almacenamiento en un depósito regional durante todo el año. Pero algunos costos de recuperación en las clases de almacenamiento a largo plazo pueden ser sustanciales y deberían analizarse minuciosamente a la hora de tomar decisiones con respecto a la clase de almacenamiento.

Consideraciones y recomendaciones con respecto al rendimiento

Esfuerzo ●●○ **Ahorro** ●○○

"¿Desde qué lugar se accederá a estos datos?" es una de las principales preguntas que deben considerarse a la hora de evaluar el rendimiento e intentar establecer la mejor clase de almacenamiento para un caso de uso en particular. La localización puede influenciar directamente con cuánta rapidez se ingresa y extrae contenido de la ubicación de almacenamiento que has seleccionado. Por ejemplo, un "objeto caliente" que se usa a nivel global (una base de datos a la cual se accede con frecuencia, como una aplicación que registra el horario de los empleados) funcionaría bien en una ubicación multirregional, lo que permite almacenar un objeto en varias ubicaciones. Posiblemente, esto pueda acercar el contenido a tus usuarios finales además de incrementar tu disponibilidad general. Otro ejemplo es la aplicación de un juego con una amplia distribución geográfica de los usuarios. Esto acerca el contenido al usuario para que tenga una mejor experiencia (menos demora) y garantiza que tu último archivo guardado se distribuya en varias ubicaciones, para que no pierdas el botín que tanto te costó ganar ante una interrupción regional.

Una cosa que debes recordar a la hora de considerar esta opción es que el almacenamiento en ubicaciones multirregionales permite un mejor rendimiento y una mayor disponibilidad, pero debe pagarse una prima y podría aumentar los cargos de salida de red, según el diseño de tu aplicación. Durante la fase de diseño de la aplicación, este es un factor importante a tener en cuenta. Otra opción cuando pensamos en el rendimiento son los depósitos en ubicaciones regionales (una buena elección si tu región está relativamente cerca de tus usuarios finales). Puedes seleccionar una región específica en la que residirán tus datos y obtener una redundancia garantizada dentro de esa región. Por lo general, este tipo de ubicación es una apuesta segura cuando tienes un equipo que trabaja en un área particular y accede a un conjunto de datos con una frecuencia relativamente alta. Según hemos observado, este es

el tipo de ubicación de almacenamiento que más se usa, ya que maneja las necesidades de la mayoría de las cargas de trabajo bastante bien. Este tipo de ubicación es de acceso rápido, es redundante dentro de la región y tiene un precio accesible, en términos generales, como almacén de objetos.

Para algo que suena tan simple como un depósito, el almacenamiento de objetos en la nube en realidad ofrece amplias posibilidades, y todas tienen diversos implicaciones en cuanto al rendimiento y los costos. Como puedes ver, hay muchas formas de ajustar las necesidades de almacenamiento de tu compañía para ahorrar algo de espacio y dinero de forma automatizada y planificada. Google Cloud proporciona muchos atributos para ayudarte a garantizar que estás sacando el máximo provecho de tu inversión en Google Cloud.

Capítulo 4
Optimización de los costos de red

Cada despliegue en la nube necesita una red a través de la cual migrar los datos. Sin una red, no puedes ver videos de gatos ni subir tus selfies, mucho menos permitir que los microservicios se comuniquen entre sí.

Google Cloud brinda una red flexible, escalable y global para tus cargas de trabajo y servicios en la nube, y el modo en que usas esa red afecta cuatro aspectos críticos de tu despliegue: costo, seguridad, rendimiento y disponibilidad. Para diseñar una arquitectura de red fiable y sólida, pero aun así rentable, es recomendable que varios equipos dentro de la compañía den su opinión con respecto a estos cuatro elementos para ayudarte a establecer tus prioridades. Los siguientes consejos hacen hincapié en algunas consideraciones que deberías tener presentes a la hora de diseñar tu solución de red.

Es recomendable que varios equipos dentro de la compañía den su opinión sobre los elementos de la arquitectura de red para ayudarte a determinar tus prioridades.

Entender los flujos de tráfico en la red

El primer paso a la hora de revisar tu estrategia general de gastos de red es comprender lo que estás usando (es decir, cuánto tráfico ingresa a tu entorno de Google Cloud y sale de él). Esto es fácil de hacer con los registros de flujo de VPC, que llevan un registro de los flujos de red enviados y recibidos por las instancias de VM. Cada entrada en el registro de flujo capta información como la IP de origen, IP de destino, y bytes enviados y recibidos por cada conexión de red (exactamente el tipo de información que se necesita para entender el tráfico de red). Estos registros son recogidos en Cloud Logging y luego los puedes exportar a BigQuery para visualizar tus tendencias.

Algunos de los casos de uso para registros de flujo de VPC son monitoreo de red, investigación forense, análisis de seguridad en tiempo real y, para el caso que nos ocupa, optimización de costos. En cuanto a la optimización de los gastos de red, la información más relevante en los registros de flujo de VPC es:

- El tráfico entre regiones y zonas
- El tráfico de Internet hacia países específicos

- Los flujos que generan la mayor cantidad de tráfico (top talkers)

Aquí encontrarás instrucciones paso a paso sobre cómo habilitar los registros de flujos de VPC.

Identifica los flujos que generan mayor cantidad de tráfico

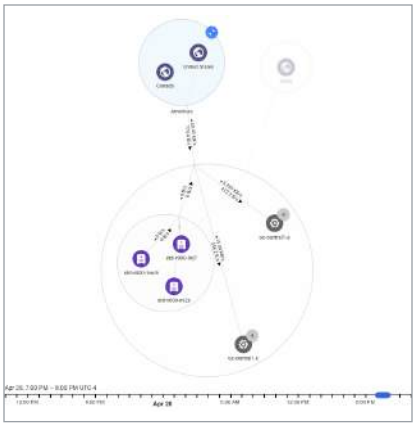
Esfuerzo ●●● **Ahorro** ●●●

La información que obtienes a partir de los registros de flujos de VPC pueden ayudarte a identificar posibles ahorros en tus costos de red actuales. Por ejemplo, la ubicación geográfica es un factor importante que debes considerar a la hora de diseñar para obtener un gasto óptimo. No todas las tarifas de red se crean de igual forma; las diversas regiones tienen diferentes costos de red.

También es importante conocer el diseño de tu red y cuánto tráfico fluye entre tus aplicaciones y usuarios. La topología de red, un módulo del Network Intelligence Center, proporciona una visibilidad integral de tu despliegue global de Google Cloud y su interacción con la Internet pública, incluida una visualización a nivel de toda la organización de la topología y las métricas asociadas de rendimiento de la red. Esto te permite identificar los despliegues ineficientes y tomar las medidas necesarias para optimizar tus tarifas de salida de red regionales e intercontinentales.

Para las tarifas de salida de Internet generales (es decir, un grupo de servidores web que proporcionan contenido a Internet), los precios pueden variar según la región en la que se encuentren esos servidores. Por ejemplo, el precio por GB en us-central1 es más bajo que el precio por GB en asia-southeast1. Otro ejemplo es el tráfico que fluye entre las regiones de Google Cloud, que puede variar significativamente dependiendo de la ubicación de esas regiones (incluso aunque no esté saliendo a Internet). Por ejemplo, el costo de sincronizar los datos entre asia-south1 (India) y asia-east1 (Taiwán) es cinco veces mayor que el de sincronizar el tráfico entre us-east1 (Carolina del Sur) y us-west1 (Oregon).

No todas las tarifas de red se crean de igual forma; las diversas regiones tienen diferentes costos de red.



Ejemplo de topología mediante topología de red, Network Intelligence Center

Además de las consideraciones regionales, es importante analizar en qué zonas están tus cargas de trabajo. Según sus necesidades de disponibilidad, quizás puedas diseñarlas para usar un tráfico de red dentro de la zona sin costo. ¡Has leído bien, sin costo! Piensa en tus VM que se comunican a través de direcciones IP externas y públicas, pero que están en la misma región o zona. Al configurarlas para que se comuniquen a través de sus direcciones IP internas, puedes ahorrar el costo de lo que habrías pagado por ese tráfico si se comunicara a través de direcciones IP externas.

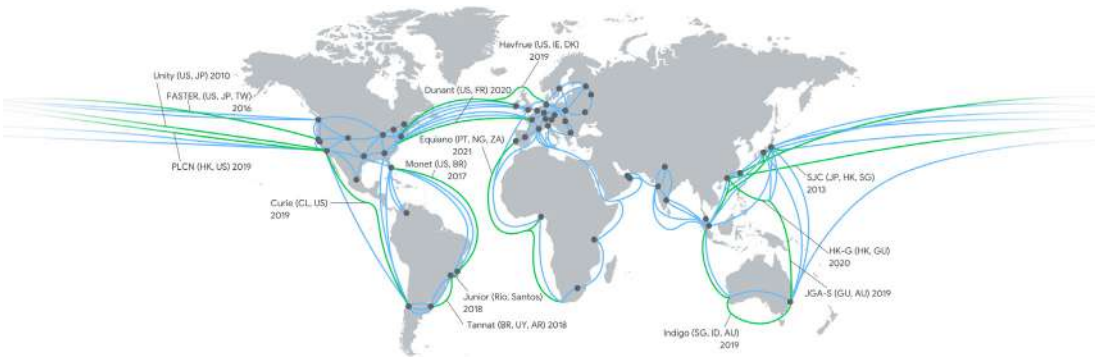
Ten en cuenta que necesitarás ponderar cualquier posible ahorro de costos de red con las implicaciones relativas a la disponibilidad de una arquitectura de una sola zona. No se recomienda desplegar en una sola zona en el caso de cargas de trabajo que requieren una alta disponibilidad, pero puede tener sentido procurar que ciertos servicios usen una red de nube privada virtual (VPC) dentro de la misma zona. Un ejemplo podría ser aplicar un enfoque de una sola zona en regiones que tienen costos más altos (Asia), pero un diseño de múltiples zonas o multirregional en América del Norte, donde los costos son inferiores.

Una vez que hayas establecido tus costos de red para un mes promedio, es recomendable que consideres algunos enfoques diferentes para asignar mejor tus gastos. Algunos clientes rediseñan las soluciones para acercar las aplicaciones a su base de usuarios, y algunos emplean Cloud CDN para reducir el volumen de tráfico y la latencia, así como para intentar aprovechar los menores costos de Cloud CDN para brindar contenido a los usuarios. Ambas son opciones viables que pueden reducir costos o incrementar el rendimiento..

Decidir cuándo usar una VPN

Esfuerzo ●●● **Ahorro** ●●○

Lo siguiente en la lista en cuanto a analizar los gastos generales de red es el total de bytes transferidos. Con los registros de flujos de VPC, puedes ver los flujos que generan la mayor cantidad de tráfico (top talkers) dentro de tu entorno. Si estás enviando grandes cantidades de datos (piensa en TB/PB), es recomendable que te asegures de aprovechar cualquier posible descuento al que puedas tener derecho.



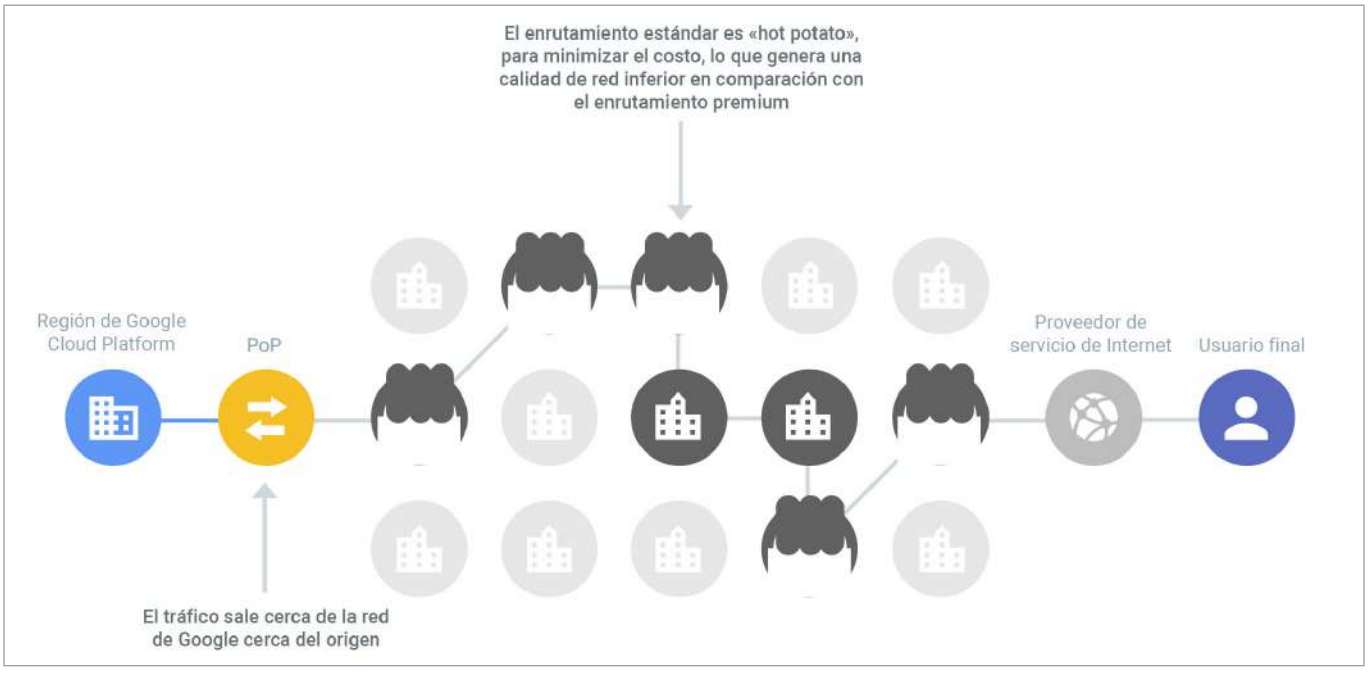
Hemos visto muchos clientes que envían grandes cantidades de datos a diario desde su entorno local a Google Cloud a través de una VPN o quizás de forma directa a través de Internet (con suerte, encriptados con SSL). Algunos clientes, por ejemplo, tienen bases de datos en hardware dedicado y on-premise, mientras que sus aplicaciones de front-end atienden solicitudes en Google Cloud. Si esto se parece a tu situación, piensa en si deberías usar una interconexión dedicada o una interconexión de Partner. Si envías grandes cantidades de datos de forma constante, puede ser más económico establecer una conexión dedicada en lugar de generar costos asociados con el tránsito de tu tráfico a través de la Internet pública o mediante una VPN.

Consulta los detalles de las consideraciones de diseño que debes revisar a la hora de seleccionar una interconexión.

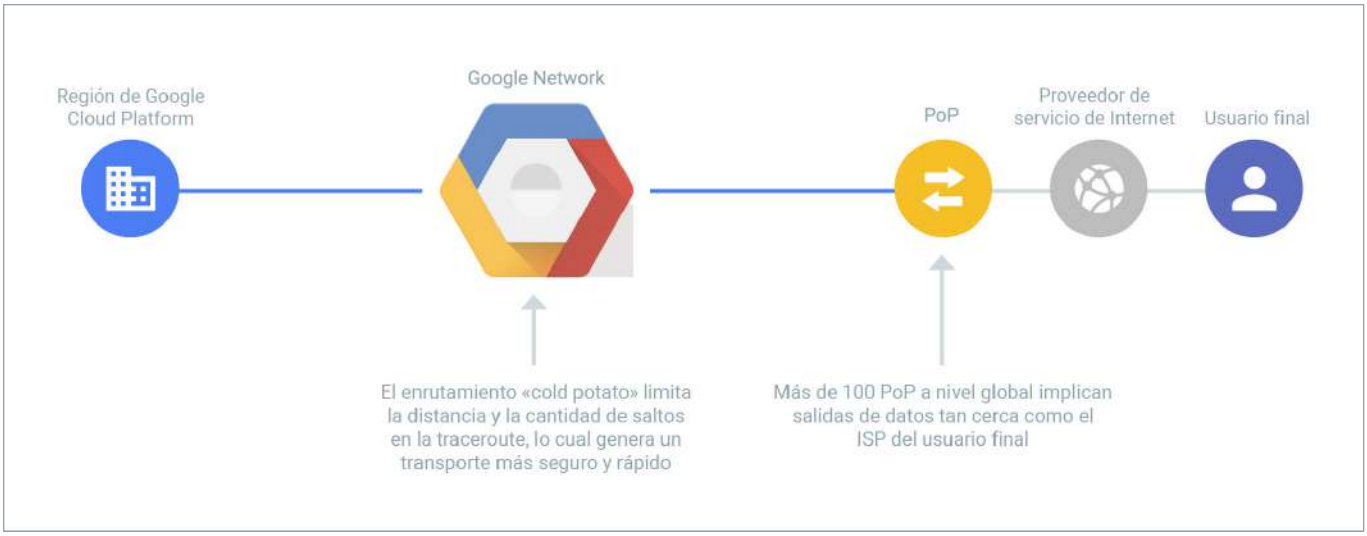
Tu red optimizada a tu manera con niveles

Esfuerzo ●○○ **Ahorro** ●●○

Uno de los principales diferenciadores de Google Cloud es que obtienes acceso a la espina dorsal de la red premium de Google, que se usa de forma predeterminada para todos los servicios. Pero quizás no necesites ese rendimiento y baja latencia para todos tus servicios. Un ejemplo podría ser la distribución de un informe de ventas diario que no necesita estar disponible de inmediato a nivel global. Para aquellos servicios respecto de los cuales estás dispuesto a compensar rendimiento y costo, ofrecemos niveles de servicio de red.



Elegir el nivel de red Estándar puede ahorrar costos.



Elegir el nivel de red Premium ofrece rendimiento y baja latencia.

Al elegir el nivel Estándar o Premium, puedes asignar la conectividad adecuada entre tus servicios, ajustar la red a las necesidades de tu aplicación y posiblemente reducir los costos de los servicios que podrían tolerar una latencia mayor y no requieren un acuerdo de nivel de servicio (SLA).

Existen algunas limitaciones a la hora de aprovechar el nivel Estándar por los beneficios de precio que ofrece. En un nivel alto, dichas limitaciones incluyen necesidades de cumplimiento respecto del tráfico que recorre la Internet pública, así como HTTP(S), proxy SSL, balanceo de cargas de proxy TCP o uso de Cloud CDN. Luego de revisar algunas de las recomendaciones, tendrás la capacidad de analizar tus servicios con tu equipo y determinar si puedes beneficiarte con los precios bajos del nivel Estándar sin afectar el rendimiento de tus servicios hacia el exterior.

Optimizar el uso para tu red

Esfuerzo ●●○ **Ahorro** ●●○

Las consideraciones anteriores son algunas de las que puedes tener en cuenta al revisar los costos de red. Sin embargo, en términos generales, deberías asegurarte de que estás aprovechando uno de los mayores beneficios de la red: pagar solo por lo que usas. Con esto en mente, te recomendamos que analices lo siguiente para asegurarte de sacar el máximo provecho de tu inversión en Google Cloud:

- **Cloud Logging:** probablemente no lo sepas, pero sí tienes control sobre la visibilidad del tráfico de la red, ya que puedes filtrar los registros que ya no necesitas. Consulta algunos ejemplos comunes de registros que

puedes excluir de forma segura. Lo mismo puede aplicarse a los registros de auditoría de acceso a datos, que pueden ser bastante grandes y generar costos adicionales. Por ejemplo, probablemente no necesites llevar esos registros en proyectos de desarrollo. Para registros de flujos de VPC y Cloud Load Balancing, también puedes habilitar el muestreo, que reduce drásticamente el volumen del tráfico de registro que se graba en la base de datos. Puedes configurarlo desde 1.0 (se guarda el 100% de las entradas del registro) hasta 0.0 (0%, no se guardan registros). Para resolver problemas o casos de uso personalizados, siempre puedes elegir recopilar telemetría para una red o subred VPC particular, o ir más a fondo y supervisar una instancia de VM o una interfaz virtual específica.

- **Acceso privado para clientes empresariales o de alto volumen:** aprovecha el Acceso privado a Google cuando sea posible para reducir costos y mejorar tu posición de seguridad.

- **Direcciones IP externas:** a partir de 2020, las direcciones IP externas que no estén dentro del nivel gratuito generarán un pequeño costo. No obstante, como buena práctica de seguridad general, es recomendable usar direcciones IP internas cuando corresponda. Para obtener más información sobre cómo migrar a IP internas, consulta nuestras guías para compilar la conectividad a Internet para VM privadas o crear un clúster privado en Google Kubernetes Engine.

Al revisar lo anterior, te asegurarás no solo de eliminar los gastos excesivos dentro de tu diseño sino, además, de estar aprovechando al máximo tu solución de nube. Equilibrar los costos con el rendimiento, la disponibilidad y la seguridad no es una tarea simple y, a menudo, requiere colaboración entre varios equipos. Nos gustaría creer que hay muchos enfoques para considerar y, la gran mayoría de las veces, la optimización de costos no es tanto una revisión de una sola vez, sino una filosofía de los equipos de tu aplicación. Encuentra los métodos que mejor funcionen para tus equipos y tu compañía

Capítulo 5
Optimización de los costos de análisis de datos con BigQuery

Ejecutar y administrar almacenes de datos heredados puede ser frustrante y demandar mucho tiempo, especialmente ahora que hay datos en todas partes y en todo lo que hacemos. Escalar los sistemas para ajustarse a este incremento en los datos ha hecho que sea incluso más difícil mantener las operaciones diarias. También está la molestia adicional de actualizar tu almacén de datos con un tiempo de inactividad mínimo y respaldar las iniciativas de aprendizaje automático e inteligencia artificial para satisfacer las necesidades empresariales. Nuestros clientes nos cuentan que eligen BigQuery, el almacén de datos empresariales sin servidores de Google Cloud, para poder concentrarse en el análisis y ser más productivos en lugar de administrar la infraestructura.

Con BigQuery, puedes ejecutar consultas a una velocidad sorprendente, obtener información en tiempo real con la transmisión de datos y comenzar a usar análisis avanzados y predictivos con funcionalidades integradas de aprendizaje automático. Un análisis de Enterprise Strategy Group (ESG) reveló que BigQuery puede proporcionar un costo de propiedad (TCO) total entre un 26% y un 34% inferior en un periodo de tres años en comparación con otras alternativas de almacenes de datos en la nube. Pero eso no implica que no puedan hacerse otras optimizaciones con tus datos alojados en BigQuery. Como el costo es uno de los factores principales detrás de las decisiones tecnológicas en esta era de la computación en la nube, las preguntas de seguimiento naturales que recibimos de nuestros clientes se relacionan con detalles de facturación y cómo optimizar continuamente los costos.

Confeccionamos esta lista de acciones que puedes realizar para ayudarte a optimizar tus costos (y, a su vez, los resultados de la empresa) con base en nuestras experiencias y en nuestro conocimiento de los productos. Un beneficio particular de optimizar los costos en BigQuery es que gracias a su diseño sin servidores, esas optimizaciones también generan un mejor rendimiento, así que no tendrás que tomar la estresante decisión de priorizar el rendimiento por sobre el costo o viceversa.

Entender los aspectos básicos de los precios en BigQuery

Observemos los precios para BigQuery y, luego, exploremos cada subcategoría de facturación para ofrecer recomendaciones que te ayudarán a reducir tu gasto en BigQuery. Para cualquier ubicación, los precios de BigQuery se desglosan de la siguiente forma (encontrarás más detalles a continuación):

- **Procesamiento de consultas**
  - A pedido: Esta opción se basa en la cantidad de datos procesados por cada consulta que ejecutas.
  - Tasa fija: Esta opción es la mejor para aquellos clientes que desean tener una previsibilidad de los costos. Los clientes compran recursos dedicados para el procesamiento de consultas y no pagan por las consultas individuales.
- **Almacenamiento**
  - Almacenamiento activo: Un cargo mensual por los datos almacenados en tablas o particiones que se hayan modificado en los últimos 90 días.
  - Almacenamiento a largo plazo: Un cargo mensual menor por los datos almacenados en tablas o particiones que no se hayan modificado en los últimos 90 días.
  - Inserciones de transmisión

Para BigQuery ML, consulta los Precios de BigQuery ML. Para el Servicio de transferencia de datos de BigQuery, consulta los Precios del Servicio de transferencia de datos de BigQuery.

Antes de analizar los precios, aquí encontrarás una lista de las operaciones de BigQuery que son gratuitas en cualquier ubicación:

- Carga de datos por lotes a BigQuery

```
-- No cost, since no tabl
DECLARE x DATE DEFAULT CU
-- Incurs the cost of sca
DECLARE y STRING DEFAULT
-- Incurs the cost of cop
-- table is created, you
-- script runs.
CREATE TEMP TABLE t AS SE
-- Incurs the cost of sca
SELECT column1 FROM t;
-- No cost, since y = 'fo
IF y = 'foo' THEN
  -- Incurs the cost of s
  -- y was equal to 'foo'
  SELECT * FROM dataset.d
ELSE
  -- Incurs the cost of s
  -- y was not equal to
  UPDATE dataset.differe
  SET col = 10
  WHERE true;
END IF;
-- Incurs the cost of sca
-- iteration of the loop
WHILE x < (SELECT MIN(dat
  -- No cost, since the e
  SET x = DATE_ADD(x, INT
  -- No cost, since the e
  IF true THEN
    -- LEAVE has no assoc
    LEAVE;
  END IF;
  -- Never executed, sinc
  -- a cost.
  SELECT * FROM dataset.b
END WHILE;
```

- Reagrupamiento automático en clústeres (no requiere configuración ni mantenimiento)
- Exportar datos de tablas
- Eliminar tablas, vistas, particiones, funciones y conjuntos de datos
- Operaciones con metadatos
- Consultas en caché
- Consultas que arrojaron un error
- Almacenamiento para los primeros 10 GB de datos por mes
- Consultar datos procesados por el primer 1 TB de datos por mes (útil para usuarios que pagan precios a pedido)

Entender la diferencia entre precios de tasa fija y a pedido

Esfuerzo ●○○ **Ahorro** ●●●

De forma predeterminada, BigQuery te cobra precios variables a pedido basados en la cantidad de bytes procesados por tus consultas. Si eres un cliente que maneja grandes volúmenes con cargas de trabajo estables, quizás te resulte más rentable abandonar los precios a pedido y pasar a pagar precios de tasa fija, lo que te permite procesar una cantidad ilimitada de bytes por un costo fijo previsible. Cuando te inscribes en un plan de precios de tasa fija, compras compromisos de ranuras; es decir, una capacidad de procesamiento de consultas específica que se mide en ranuras de BigQuery. Con la introducción de BigQuery Reservations, los clientes ahora pueden usar un autoservicio fácil y flexible para aprovechar los precios de tasa fija de BigQuery. Un buen punto de partida para decidir cuántas ranuras comprar es visualizar cuántas ranuras usaste el último mes con la ayuda de Cloud Monitoring.

Podrías caer en la tentación de creer que con una tasa fija no tienes que preocuparte por las optimizaciones de las consultas.

Nota: Si tus consultas exceden la capacidad de la tasa fija, BigQuery se ejecutará con mayor lentitud de forma proporcional hasta que las ranuras estén disponibles.

Podrías caer en la tentación de creer que con una tasa fija no tienes que preocuparte por las optimizaciones de las consultas. La realidad es que esto sigue afectando el rendimiento. Cuanto más rápidamente se ejecute tu consulta (tarea), más tareas podrás completar en la misma cantidad de tiempo que con las ranuras fijas. Si lo piensas, eso es la optimización de costos.

No comprar suficientes ranuras puede afectar el rendimiento, mientras que comprar demasiadas generará una capacidad de procesamiento ociosa y eso afectará el costo. Para encontrar el equilibrio perfecto, puedes comenzar con un plan de tasa fija mensual, que te permite una mayor flexibilidad para bajar de plan o cancelar después de 30 días. Una vez que tengas un cálculo aproximado lo suficientemente bueno sobre la cantidad de ranuras que necesitas, puedes cambiar a un plan anual de tasa fija para ahorrar más. Además, BigQuery Reservations te permite usar el plan de precios de tasa fija con aún más eficiencia y a planificar tus gastos.

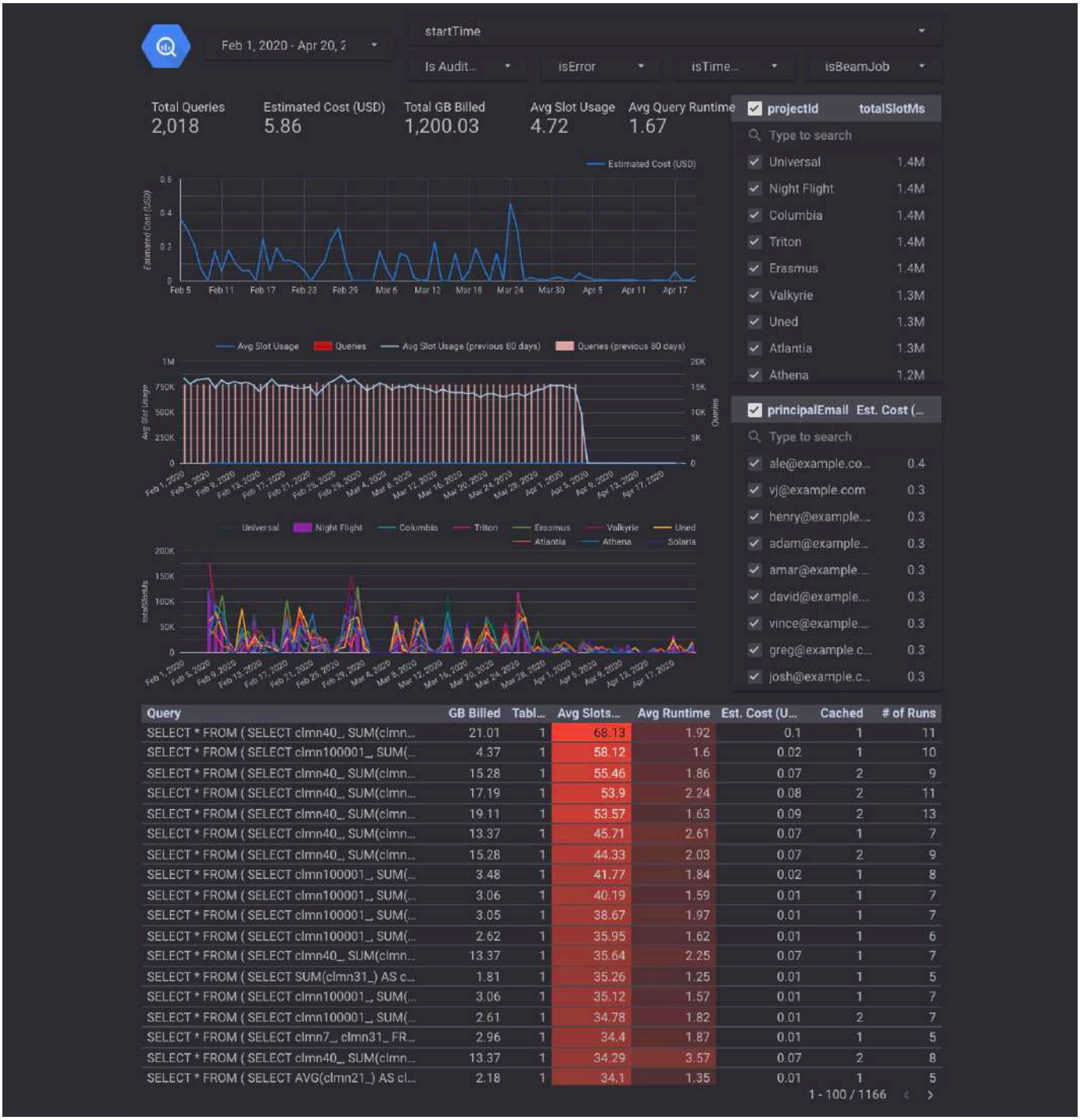
Como los requisitos empresariales cambian a un ritmo vertiginoso, hace poco presentamos las ranuras flexibles, una nueva forma de comprar ranuras de BigQuery por periodos de tiempo muy cortos, de hasta 60 segundos, además de los compromisos de tasa fija mensuales y anuales.

Las nuevas ranuras flexibles ofrecen flexibilidad para satisfacer la demanda.

Esta combinación de planes de precios a pedido y de tasa fija te permite responder de forma rápida y rentable a la cambiante demanda de análisis. Ahora que comprendes los aspectos fundamentales de los planes de precios de BigQuery, veamos cómo puedes empezar a optimizar los costos.

Visualiza tus costos de BigQuery

Una vez que comiences a entender los costos dentro de tu organización, dedica unos minutos a ejecutar un informe rápido de tu uso de BigQuery correspondiente al último mes para tener un panorama de tus costos. Puedes usar los informes de facturación en Cloud Console o simplemente exportar tus datos de facturación a BigQuery. También hay un panel detallado de Data Studio que te permite identificar las consultas caras para que puedas optimizar los costos y el rendimiento de las consultas. Además, te proporciona información sobre los patrones de uso y la utilización de recursos asociados con tu carga de trabajo. Sigue estas instrucciones paso a paso para crear un panel, como se muestra a continuación.



Los informes de BigQuery te permiten conocer tu uso.

Técnicas de optimización de costos en BigQuery: procesamiento de consultas

Es probable que consultes tus datos en BigQuery para realizar análisis y para satisfacer los casos de uso de tu empresa como análisis predictivo o administración de inventario en tiempo real.

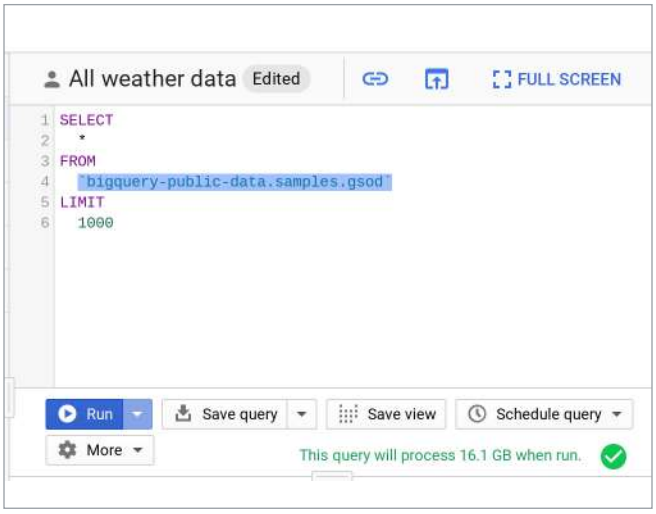
El plan de precios a pedido es el que eligen la mayoría de los usuarios y las empresas cuando comienzan a usar BigQuery. Lo que pagas es la cantidad de bytes procesados, independientemente de los datos alojados en BigQuery o las fuentes de datos externos involucradas. Hay algunas formas en las que puedes reducir la cantidad de bytes procesados. Repasemos las buenas prácticas para reducir el costo de ejecutar tus consultas, como comandos de SQL, tareas, funciones definidas por el usuario y mucho más.

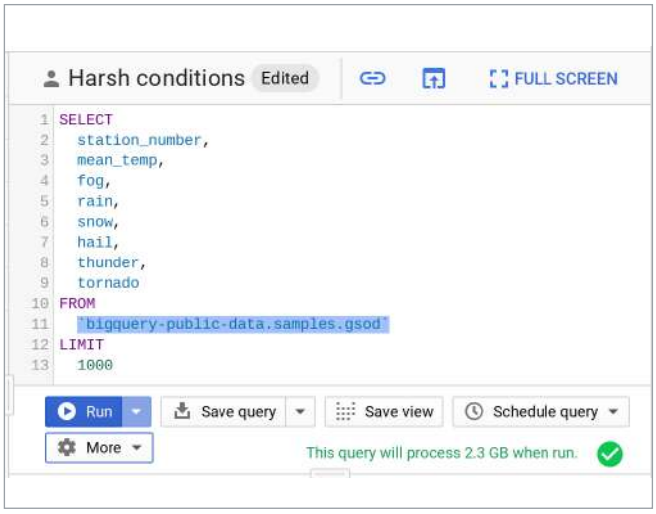
1. Solo consulta los datos que necesites. (¡Lo decimos en serio!)

Effort ●○○ **Savings** ●○○

BigQuery puede proporcionar un rendimiento increíble porque almacena los datos como una estructura de datos en columnas. Esto significa que SELECT * es la forma más cara de consultar datos, ya que realizará un escaneo de consultas completo en cada columna de la(s) tabla(s), incluso en aquellas que tal vez no necesitas. (Conocemos esa sensación de culpa que surge cuando sumamos la cantidad de veces que hemos usado SELECT * el último mes).

Miremos un ejemplo de cuántos datos procesará una consulta. Aquí estamos consultando uno de los conjuntos de datos públicos del clima disponibles en BigQuery:





Consultar conjuntos de datos públicos del clima en BigQuery

Como puedes ver, al seleccionar las columnas necesarias, podemos reducir los bytes procesados casi ocho veces, lo que representa una forma rápida de optimizar costos. Observa también que aplicar la cláusula LIMIT a tu consulta no tiene efecto en el costo. Si necesitas explorar los datos y entender su semántica, siempre puedes usar la opción de vista previa de datos sin cargo.

Recuerda que se te cobran los bytes procesados en la primera etapa de la ejecución de la consulta. Evita crear una consulta compleja de varias etapas solo para optimizar la cantidad de bytes procesados en las etapas intermedias, ya que eso no afecta el costo (aunque sí podrías lograr mejoras en el rendimiento).

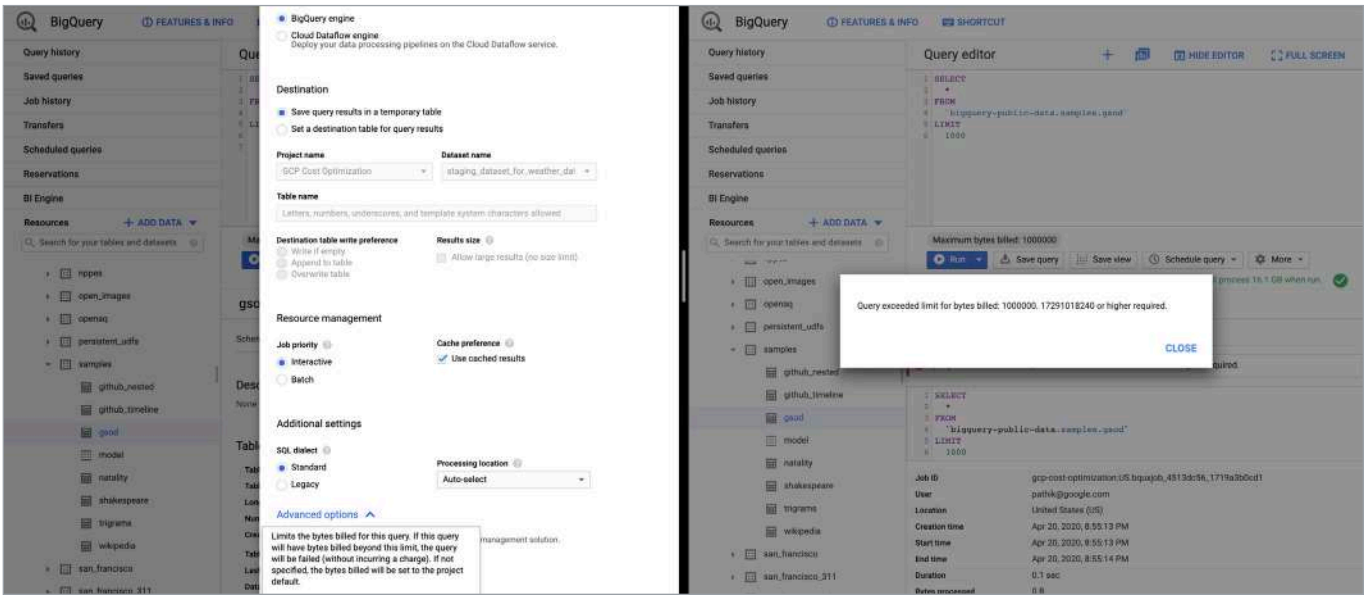
Filtra tu consulta lo antes posible y con la mayor frecuencia que puedas para reducir costos y aumentar el rendimiento en BigQuery.

2. Configura controles para evitar errores humanos accidentales.

Esfuerzo ●○○ **Ahorro** ●●○

La consulta anterior alcanzó los GB, un contratiempo que puede costarte algunos centavos, algo que es aceptable para la mayoría de las empresas. No obstante, cuando tienes tablas de conjuntos de datos que rondan los TB o PB y a las que varias personas acceden, consultar todas las consultas accidentalmente podría generar un costo de consulta significativo.

En este caso, usa la configuración de cantidad máxima de bytes facturados para limitar el costo de la consulta. Si se supera ese máximo, la consulta arrojará un error y no generará ningún costo, como se muestra a continuación.



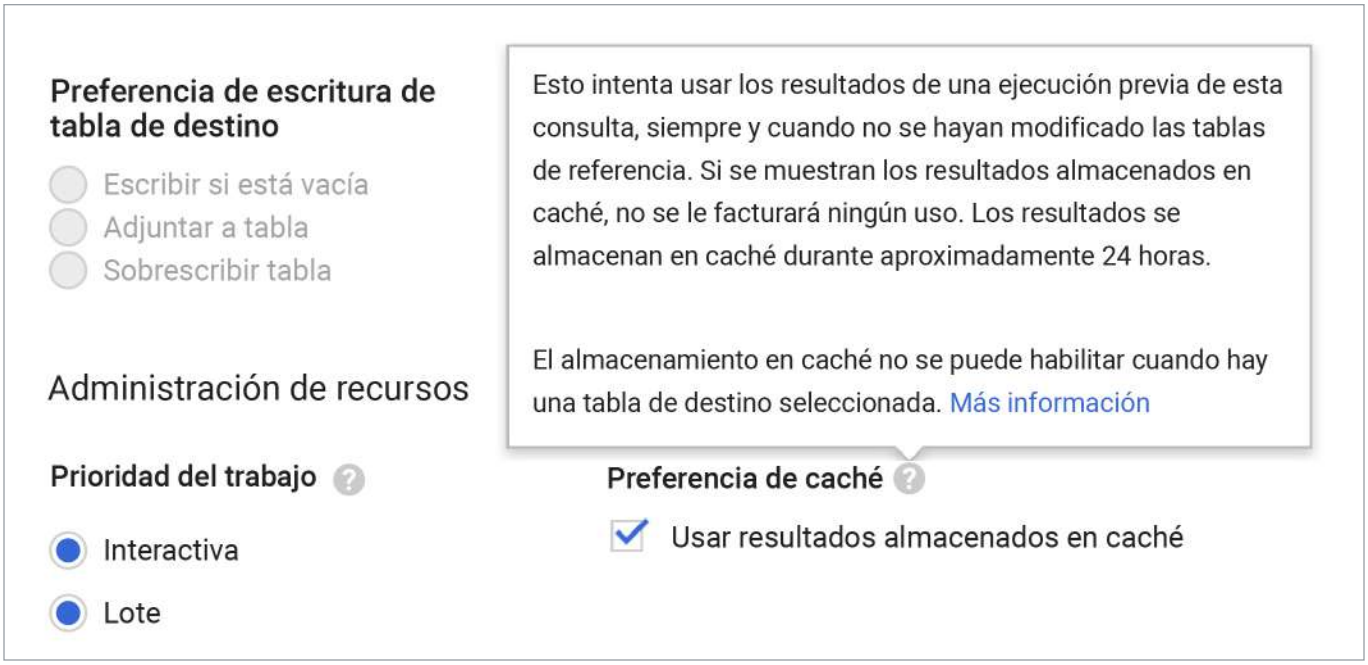
Una vez un cliente preguntó por qué el control personalizado es tan importante. Para poner las cosas en perspectiva, supongamos que tienes 10 TB de datos en una ubicación de EE. UU. (multirregional), cuyo almacenamiento te cuesta aproximadamente USD200 mensuales. Si 10 usuarios barren todos los datos con [SELECT * ..] 10 veces al mes, tu factura de BigQuery ahora será de aproximadamente USD5,000 porque estás barriendo 1 PB de datos por mes. Aplicar límites pensados cuidadosamente puede ayudar a prevenir este tipo de consultas accidentales. Ten en cuenta que cancelar una consulta en ejecución podría generar hasta la totalidad del costo de la consulta como si se hubiera completado.

Además de habilitar el control de costos a nivel de consulta, puedes aplicar una lógica similar a nivel de usuario o proyecto.

3. Usa el almacenamiento en caché con inteligencia.

Esfuerzo ●○○ **Ahorro** ●○○

Con pocas excepciones, el almacenamiento en caché en realidad puede aumentar el rendimiento de tus consultas, y no se te cobrarán los resultados obtenidos de las tablas almacenadas en caché. La preferencia de almacenamiento en caché está activada de forma predeterminada. Para ver la configuración del almacenamiento en caché en tu consola de Google Cloud, haz clic en Más -> Configuración de consulta en tu editor de consultas, como se muestra aquí: Además, ten en cuenta que el almacenamiento en caché es por usuario y por proyecto.



El almacenamiento en caché puede mejorar el rendimiento de tus consultas.

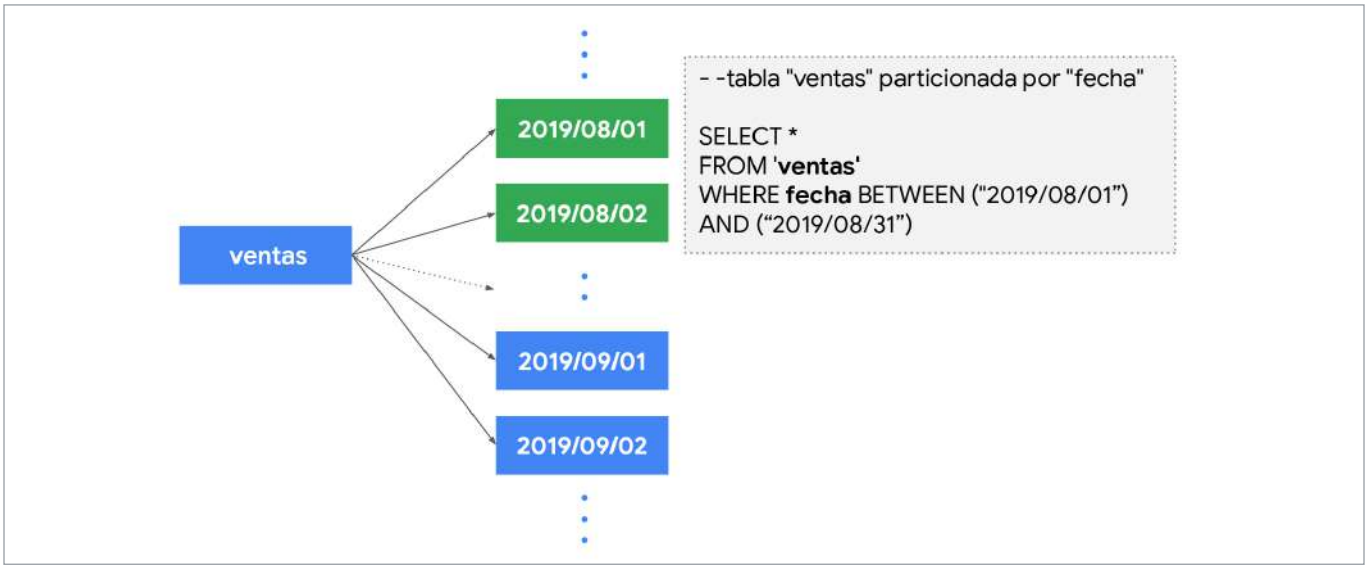
Tomemos un ejemplo de la vida real, donde tienes un panel de Data Studio respaldado por BigQuery y al que acceden cientos o incluso miles de usuarios. Esto te mostrará de inmediato que hay una necesidad de almacenar en caché tus consultas de varios usuarios de forma inteligente.

Para aumentar significativamente el uso de los resultados almacenados en caché para varios usuarios, usa una sola cuenta de servicio para realizar consultas en BigQuery, o usa conectores de comunidad, como puedes ver en esta demostración de Next '19.

4. Divide tus tablas en particiones.

Esfuerzo ●○○ **Ahorro** ●●○

Hacer particiones en tus tablas, siempre que sea posible, puede ayudar a reducir el costo de procesar consultas además de mejorar el rendimiento. En la actualidad, puedes hacer particiones en una tabla según la hora de transferencia, la fecha o incluso la columna de marca de tiempo. Supongamos que divides en particiones una tabla de ventas que contiene datos de los últimos 12 meses. Esto genera particiones más pequeñas que contienen datos para cada día, como se muestra a continuación.



Dividir tus tablas en particiones puede reducir el costo de procesar las consultas.

Ahora, cuando realizas una consulta para analizar los datos de `ventas` del mes de agosto, solo pagas por los datos procesados en esas 31 particiones, no por la tabla completa.

Un beneficio más es que cada partición se considera por separado para el almacenamiento a largo plazo, como se indicó anteriormente. Piensa en nuestro ejemplo anterior, los datos de `ventas` a menudo están cargados y modificados para los últimos meses. Así que todas las particiones que no se modificaron en los últimos 90 días ya te están ahorrando parte de los costos de almacenamiento. Para obtener realmente los beneficios de consultar una tabla particionada, deberías filtrar la tabla con una columna de partición.

A la hora de crear o actualizar tablas particionadas, puedes habilitar Requerir filtro de partición, que forzará a los usuarios a incluir una cláusula WHERE que especifique la columna de partición, o la consulta arrojará un error.

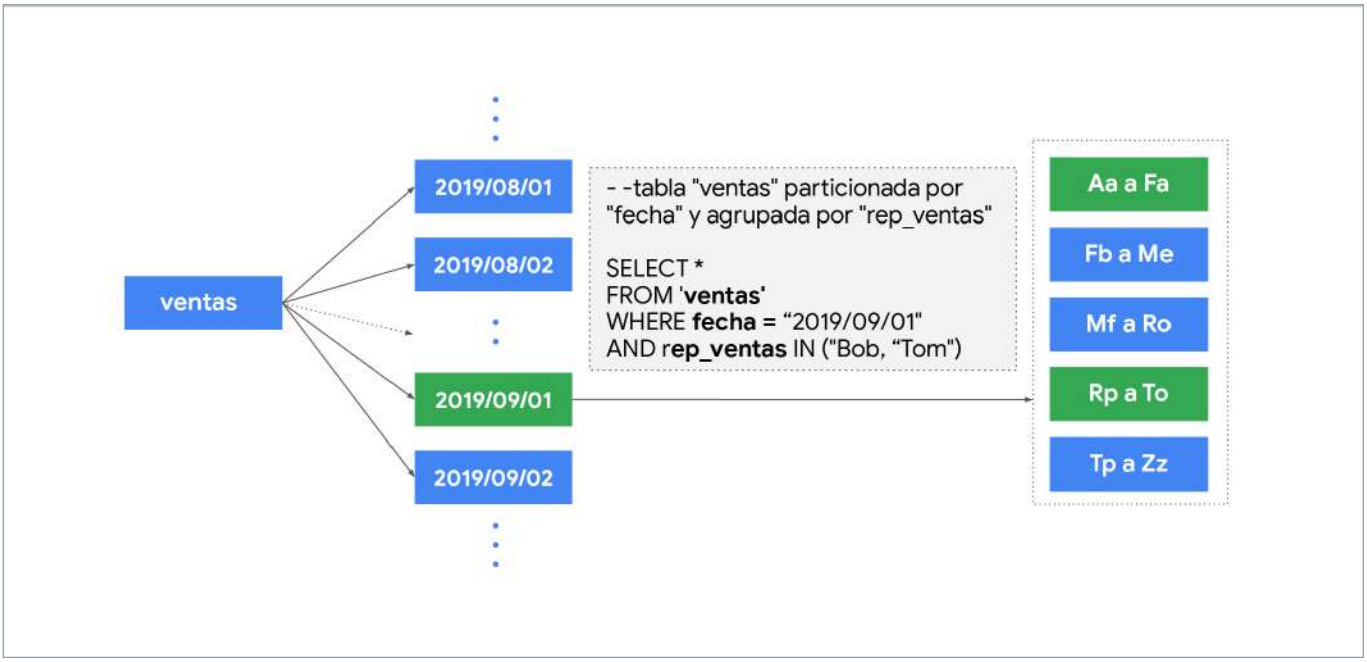
5. Reduce aún más el barrido de tus datos con el agrupamiento en clústeres.

Esfuerzo ●○○ **Ahorro** ●○○

Después de dividir en particiones, puedes agrupar tus tablas en clústeres, lo cual organiza tus datos según el contenido para hasta cuatro columnas. Posteriormente, BigQuery ordena los datos con base en el orden de las columnas especificado y los organiza en un bloque. Cuando usas los filtros de consultas con estas columnas, BigQuery analiza solo los bloques relevantes mediante un proceso que se conoce como reducción de bloques.

Por ejemplo, a continuación, el líder del equipo de ventas necesita un panel que muestre las métricas relevantes para representantes de ventas específicos. Habilitar el agrupamiento en clústeres en la columna rep_ventas es una buena estrategia, ya que se utilizará a menudo como filtro. Como se muestra abajo, puedes ver que BigQuery solo

analiza una partición (2019/09/01) y los dos bloques donde se puede encontrar a los representantes de ventas Bob y Tom. El resto de los bloques en esa partición se reducen. Esto disminuye la cantidad de bytes procesados y, en consecuencia, el costo de consulta asociado.



Habilitar el agrupamiento en clústeres puede ayudar a reducir costos.

El agrupamiento en clústeres está permitido solamente en datos particionados. Siempre puedes usar la división en particiones según los datos de transferencia o ingresar una columna de fecha o marca de tiempo falsa para permitir el agrupamiento en clústeres en tu tabla.

Puedes encontrar mucha más información sobre agrupamiento en clústeres aquí. Y explora las nuevas vistas materializadas para tener un rendimiento y una eficiencia mejorados, además de ahorrar costos.

Técnicas de optimización de costos en BigQuery: almacenamiento

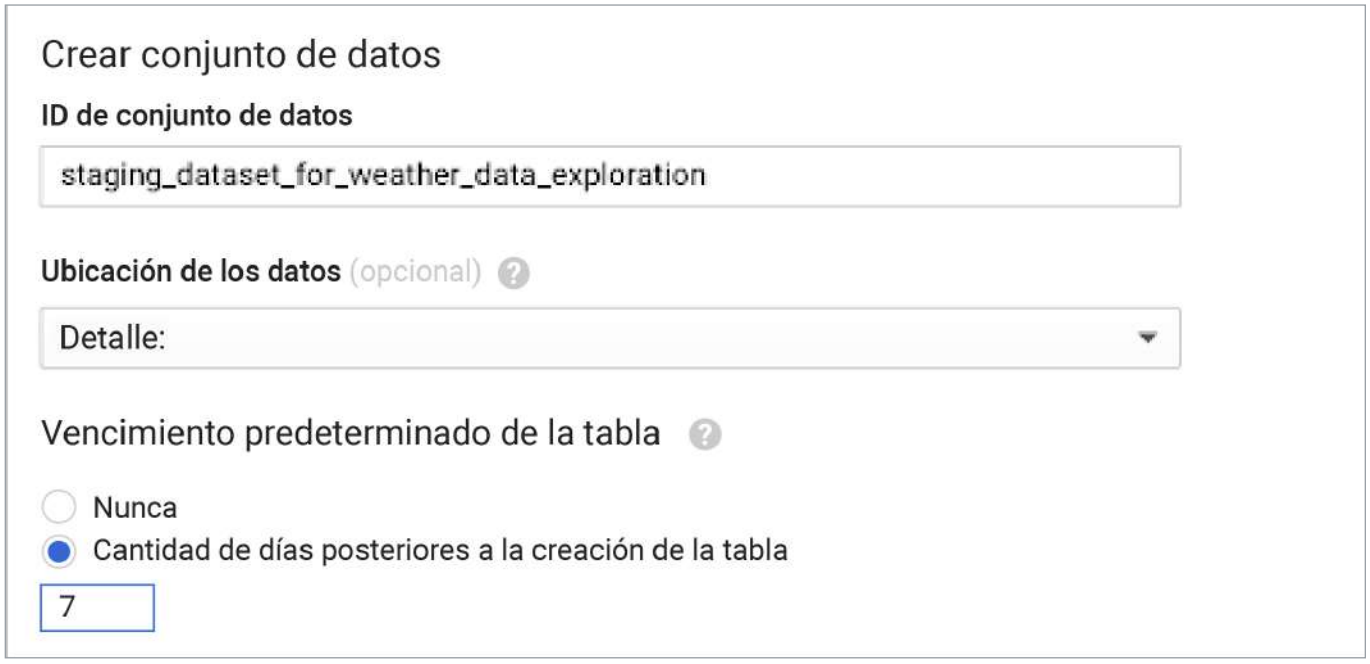
Una vez que los datos están cargados en BigQuery, los cambios se basan en la cantidad de datos almacenados en tus tablas por segundo. A continuación, encontrarás algunas recomendaciones para optimizar tus costos de almacenamiento de BigQuery.

1. Guarda tus datos solo durante el tiempo que los necesites.

Esfuerzo ●○○ **Ahorro** ●●○

De forma predeterminada, los datos almacenados en el formato de datos en columnas de Capacitor de BigQuery ya están encriptados y comprimidos. Configura el vencimiento predeterminado de la tabla en tu conjunto de datos para datos temporales de preproducción que no necesitas conservar.

En este ejemplo, solo necesitamos consultar el conjunto de datos de preproducción del clima hasta que el trabajo posterior limpie los datos y los mueva a un conjunto de datos de producción. Podemos establecer siete días para el vencimiento predeterminado de la tabla.



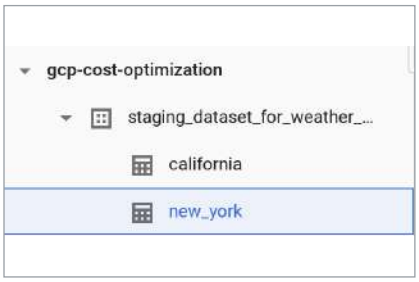
Las fechas de vencimiento de las tablas pueden ayudar a ahorrar recursos.

Observa que si no vas a actualizar el vencimiento predeterminado de la tabla para un conjunto de datos, solo se aplicará a las nuevas tablas creadas. Usa una Declaración DLL para modificar tus tablas existentes.

BigQuery también ofrece la flexibilidad para proporcionar diferentes fechas de vencimiento de las tablas dentro del mismo conjunto de datos. Entonces, esta tabla llamada nueva_york en el mismo conjunto de datos necesita que los datos se retengan durante más tiempo.

Como se muestra en las imágenes de la siguiente página, nueva_york retendrá sus datos durante seis meses, y

como no especificamos un vencimiento de la tabla para california, su vencimiento será el predeterminado de siete días.

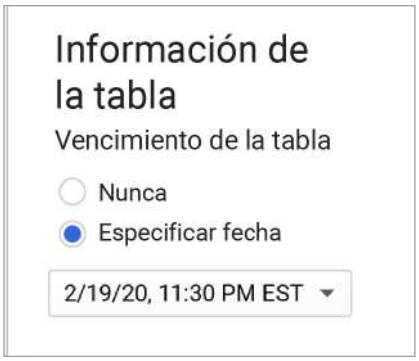


Al igual que el vencimiento a nivel de conjunto de datos y a nivel de tabla, también puedes configurar un vencimiento a nivel de partición. Consulta nuestra documentación pública para ver los comportamientos predeterminados.

2. Sé consciente de cómo editas tus datos.

Esfuerzo ●○○ **Ahorro** ●●○

Si tu tabla o partición de una tabla no se ha editado durante 90 días, el precio de los datos almacenados en la tabla se reducirá de forma automática en aproximadamente un 50%. No hay una disminución en el rendimiento, la durabilidad, la disponibilidad ni ninguna otra funcionalidad cuando una tabla o partición pasa a considerarse para almacenamiento a largo plazo.



Para sacarle el máximo provecho al almacenamiento a largo plazo, ten en cuenta cualquier acción que edite los datos de tus tablas, como transmitir, copiar o cargar datos, incluso cualquier acción de DML o DDL. Esto devolverá tus datos a un almacenamiento activo y reiniciará el cronómetro de 90 días. Para evitarlo, puedes considerar cargar el nuevo lote de datos a una nueva tabla o a una partición de una tabla si tiene sentido para tu caso de uso.

Consultar los datos de las tablas, además de algunas otras acciones, no reinicia el cronómetro de 90 días y los precios continúan considerándose como almacenamiento a largo plazo.

En la mayoría de los casos, mantener los datos en BigQuery es provechoso a menos que estés seguro de que se accederá a los datos en la tabla al menos una vez por año, como el almacenamiento de archivos por razones legales o normativas. En ese caso, explora la opción de exportar los datos de la tabla a la clase coldline de un depósito de Cloud Storage por un precio incluso mejor que el del almacenamiento a largo plazo de BigQuery.

3. Evita las copias duplicadas de datos.

Esfuerzo ●○○ **Ahorro** ●○○

BigQuery usa un modelo de acceso a datos federados que te permite consultar los datos directamente desde fuentes de datos externas como Cloud Bigtable, Cloud Storage, Google Drive y Cloud SQL. Esto es útil para evitar las copias duplicadas de datos, reduciendo así los costos de almacenamiento. También es útil para leer datos en una pasada desde una fuente externa o acceder a una pequeña cantidad de datos que se modifican con frecuencia y que no es necesario que se carguen en BigQuery cada vez que se cambian.

Elige esta técnica para aquellos casos de uso en los que tenga más sentido. Por lo general, las consultas que se ejecutan en fuentes externas no tienen un rendimiento tan bueno en comparación con las consultas ejecutadas sobre los mismos datos almacenados en BigQuery, ya que los datos almacenados en BigQuery tienen un formato de columnas que genera un rendimiento mucho mayor.

4. Verifica si estás usando la inserción de transmisión para cargar tus datos.

Esfuerzo ●○○ **Ahorro** ●●○

Puedes cargar los datos a BigQuery de dos formas: como tarea de carga por lotes o a través de una transmisión en tiempo real, mediante las inserciones de transmisión. A la hora de optimizar tus costos de BigQuery, lo primero que debes hacer es revisar tu factura y ver si estás pagando las inserciones de transmisión. Si la respuesta es afirmativa, pregúntate, "¿Necesito que los datos estén inmediatamente disponibles (segundos en lugar de horas) en BigQuery?" y "¿Usaré estos datos para cualquier otro caso de uso en tiempo real una vez que los datos estén disponibles en BigQuery?" Si la respuesta a cualquiera de estas preguntas es no, entonces te recomendamos usar la carga de datos por lotes, que es gratis.

5. Entiende los procesos de copia de seguridad y recuperación de desastres de BigQuery.

Esfuerzo ●○○ **Ahorro** ●○○

BigQuery conserva un historial de siete días con los cambios realizados a tu tabla, lo que te permite consultar una instantánea de tus datos en un punto en el tiempo. Esto significa que puedes revertir los datos sin restaurar a partir de las copias de seguridad de recuperación. Si se elimina la tabla, su historial se elimina después de dos días.

Para encontrar la cantidad de filas de una tabla de hace una hora en una instantánea, usa la siguiente consulta:

```
Select COUNT(*) FROM [Project _ ID:Dataset.Table@-3600000]
```

Encuentra más ejemplos en la documentación.

En el caso de datos que sean críticos para la empresa, sigue la guía de Situaciones de recuperación ante desastres para datos con el objetivo de hacer una copia de seguridad de datos, especialmente si estás usando BigQuery en una ubicación regional.

Celebra tu éxito

Si lo usas correctamente, BigQuery puede satisfacer todas tus necesidades modernas de almacenamiento de datos a un precio muy razonable. Una vez que se hayan implementado las acciones de optimización de costos, deberías ver una reducción considerable en tu factura de BigQuery (a menos que hayas seguido las buenas prácticas desde el primer día). ¡En cualquiera de los dos casos, celebra tu éxito! Te lo mereces.

Optimización de costos en acción

Si sigues estos principios podrías ahorrar en costos y abrir la capacidad en el corto plazo (pero el retorno de la inversión a largo plazo paga bastantes dividendos). Los clientes con los que hablamos han reducido drásticamente los costos de datos y han ahorrado en infraestructura, lo que les permite agregar funcionalidades totalmente nuevas y explorar soluciones innovadoras como el aprendizaje automático. También han obtenido un mejor rendimiento y una mayor simplicidad en sus entornos de TI. Además, escalar según sea necesario se vuelve mucho más fácil.

Por ejemplo, un estudio determinó que la base de datos Spanner de Google Cloud acarrea un costo total de propiedad (TCO) que es un 78% menor que el de las bases de datos on-premise. Optiva migró sus bases de datos de Oracle heredadas a Google Cloud en busca de una eficiencia de costos y logró una mejor escala.

Además, los clientes de diversas industrias han podido hacer mucho más con la nube, y han ayudado a sus equipos a ser más productivos, en comparación con otros proveedores de nube o tecnologías heredadas. Esto puede conducir a mejores experiencias de cliente, innovación en nuevos productos y la incorporación de nuevas iniciativas empresariales. El proveedor de la plataforma de aprendizaje automático MD Insider solía tener un funcionamiento más lento de la red, y las tasas de errores y costos crecieron en consecuencia. Usar Google Cloud para los servicios de datos hizo que se quintuplicara el rendimiento y se acelerara el tiempo de salida al mercado hasta en un 30%. Los desarrolladores del sitio sobre desempeño médico ahora son mucho más productivos, y el tiempo que anteriormente dedicaban a administrar la infraestructura ahora lo usan para desarrollar productos.

Raycatch redujo los costos de infraestructura para su tecnología de energía renovable basada en IA en un 80% con Compute Engine, Cloud BigTable y otras tecnologías. La compañía agregó flexibilidad y estabilidad a sus sistemas, además de abrir la capacidad de su nube anterior, lo que permite a los desarrolladores trabajar más rápido. Raycatch pudo reducir 60 veces las horas de trabajo necesarias para supervisar una tarea de análisis; por eso, ahora el equipo de TI puede hacer un valioso trabajo de optimización mucho mayor.

Hay muchas estrategias de administración de costos y buenas prácticas que pueden ayudarte a sacar el máximo provecho de la nube. Usa lo que funcione mejor para tu equipo y organización. Con un poco de esfuerzo, la nube ofrece eficiencias en el presente y el potencial de un ROI significativo en el futuro.

Obtén más información sobre la optimización de costos

Descubre nuevas formas de reducir y optimizar tu gasto de TI para un crecimiento inmediato y a largo plazo. Habla con un experto de Google Cloud.

Google Cloud

cloud.google.com