redislabs
HOME OF REDIS

La latencia es el nuevo apagón

POR QUÉ LA VELOCIDAD ES LA NUEVA APUESTA

Resumen ejecutivo

A medida que las organizaciones aprovechan cada vez más las aplicaciones alojadas en diferentes infraestructuras, en diferentes geografías y con diferentes proveedores, la velocidad a la que los usuarios finales pueden acceder a esas aplicaciones se ve sometida a presión.

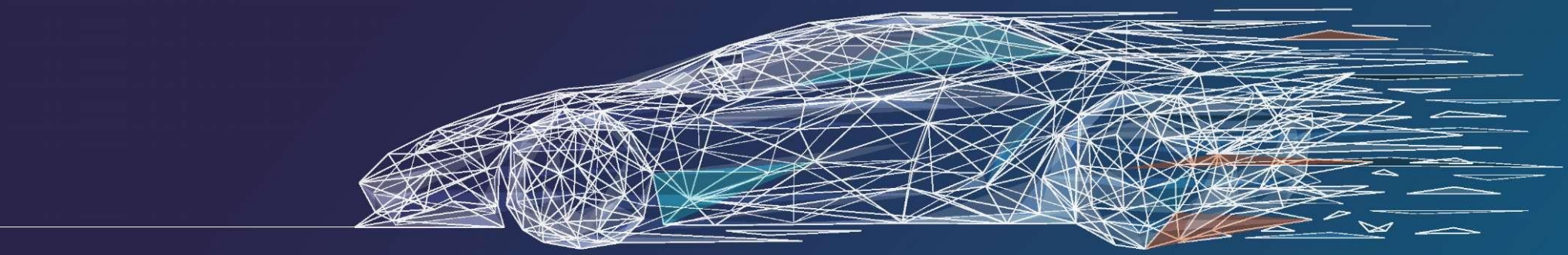
Además, las aplicaciones de hoy están compuestas por un número significativo de componentes diferentes, y usted tiene una receta para una experiencia de usuario degradada.

Estos rasgos duales (modularidad de la aplicación y complejidad de la infraestructura) pueden resultar directamente en un rendimiento deficiente de la aplicación.

Es por esta razón que la velocidad de la capa de datos, la capa horizontal común en toda la aplicación, es crítica.

Ser capaz de utilizar una capa de datos replicada geográficamente, mientras se evitan los problemas relacionados con la inconsistencia de los datos, es un desafío que todos los líderes de TI deben resolver.

Al aprovechar una capa de datos que unifica sus datos en las nubes y el mundo, las organizaciones pueden superar algunas de las limitaciones inherentes que han desafiado a los equipos de tecnología durante décadas y brindar mejores experiencias a sus usuarios finales.



Introducción

Los equipos digitales han pasado la última década asegurándose de que sus activos digitales estén disponibles en todo momento, ¡y lo han logrado en gran medida! La alta disponibilidad es ahora la norma.

Las organizaciones han logrado, en parte, este alto nivel de digitalización y alta disponibilidad al aprovechar los beneficios que brinda la nube: facilidad de escalabilidad, servicios modulares, patrones arquitectónicos más refinados. Todas estas características permiten resultados positivos, pero lo hacen con la contracara de una mayor complejidad. Esa complejidad fue inicialmente más impactante en términos de disponibilidad y dio lugar a lo que llamamos la Época de Disponibilidad. Sin embargo, a medida que las organizaciones comprenden mejor cómo ofrecer una alta disponibilidad, descubren que aún quedan otros problemas por resolver.

Pero ahora la Época de Disponibilidad está comenzando a decaer, ya que las organizaciones buscan cada vez más reducir la latencia como la próxima clave para alcanzar los resultados que buscan. Comprenden cada vez más que los productos y servicios lentos podrían no estar disponibles en lo absoluto y que la latencia es el nuevo apagón.

Desafortunadamente, resolver los problemas de latencia suele ser más difícil que crear una alta disponibilidad. Si bien la disponibilidad se puede mejorar a través de una buena ingeniería, mayores niveles de redundancia y mejor monitoreo y visibilidad, el enigma de la latencia está limitado por las mismas leyes de la física.

Para reducir la latencia tanto como sea posible, las organizaciones necesitan entender qué es la latencia y los factores que contribuyen a ella, y tener pautas claras y definitivas para reducir, tanto como sea posible, la latencia para los usuarios de sus aplicaciones y sitios web.

Si la latencia es la nueva interrupción, aquí está la inteligencia que necesita para ofrecer la latencia más baja físicamente posible.

Las organizaciones comprenden cada vez más que los productos y servicios lentos podrían no estar disponibles en lo absoluto, y que la latencia es el nuevo apagón.

No es el grande el que se come al pequeño, sino el rápido el que se come al lento

Moverse rápido es la nueva normalidad. Si bien el análisis y la prudencia fueron una vez el nombre del juego, la realidad operativa actual es que para mantenerse por delante de sus competidores, las organizaciones deben innovar más rápido que nunca. El panorama operativo de todas las organizaciones está cambiando rápidamente y el éxito se otorgará a quienes mejor puedan reaccionar ante este dinamismo.

POR QUÉ LA AGILIDAD ES TAN IMPORTANTE: EL TIEMPO ES ESENCIAL

El mundo en el que vivimos hoy es notablemente diferente al de hace unos pocos años, y la tasa de cambio sigue aumentando. En este contexto, brindar a las organizaciones la capacidad de moverse con rapidez es más importante que nunca. Es importante comprender los cambios que están ocurriendo en la sociedad para comprender mejor cuán importante es realmente moverse rápido.

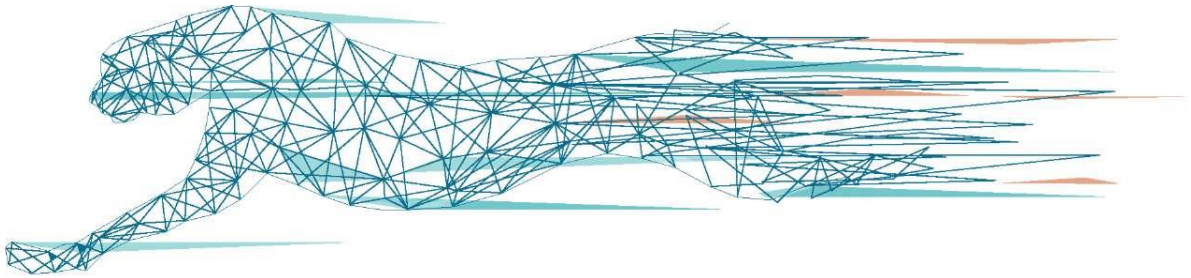
En 2011, el destacado empresario, inversor y miembro de la junta Marc Andreessen (el inventor del navegador web Netscape) escribió un artículo de opinión ahora famoso para ***The Wall Street Journal,*** donde explicaba por qué el software se está llevando el mundo por delante.

En su ensayo, Andreessen presentó su teoría sobre el tamaño, el alcance y la velocidad de este cambio, sugiriendo que: “estamos en medio de un cambio tecnológico y económico dramático y amplio en el que las empresas de software están preparadas para hacerse cargo de grandes sectores de la economía„.

Pero si bien el hecho de que este cambio es de gran importancia, es la velocidad de este cambio lo que es de mayor relevancia aquí. La capacidad de realizar cambios, aprovechar una variedad de herramientas y brindar la mejor experiencia al usuario final es fundamental para la capacidad de moverse rápido. Si todo esto se parece mucho a la forma en que funcionan las empresas de Silicon Valley, tiene sentido. El hecho es que muchas de las organizaciones que tienen éxito en la disrupción de las industrias tradicionales se ven y se sienten cada vez más como empresas de tecnología empresarial.

Hace casi diez años que Andreessen escribió ese ensayo, y muchas de sus predicciones se han cumplido. Si bien se ha convertido en un cliché usar Tesla, Uber, Lyft, Netflix y Airbnb como ejemplos de disrupción digital, es seguro decir que los ejecutivos de las empresas de taxis y hotelería se han visto afectados por una marea de proporciones sin precedentes. Sin embargo, más allá del cliché, lo importante a tener en cuenta es cuánto esfuerzo hacen estas empresas para ofrecer la experiencia de cliente más rápida posible en sus aplicaciones: la velocidad realmente importa.

Es una obviedad que moverse rápido en un entorno organizacional se basa en proporcionar experiencias digitales que muestran estos atributos de velocidad. La latencia es el nuevo bloqueador de la transformación organizacional.



Cambiar el mundo con un salto a lo digital

Una gran cantidad de ejemplos de organizaciones tradicionales tienen todas sus esperanzas de éxito del futuro en un cambio a lo digital. Vale la pena mirar algunos ejemplos para tener una idea de esta escala.

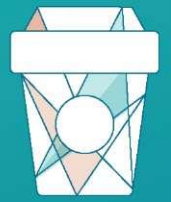
DIGITAL JOE: STARBUCKS SE CENTRA EN EL MUNDO DIGITAL

El CEO de Starbucks, Kevin Johnson, fue ejecutivo de Microsoft. Su experiencia en la gigante empresa de tecnología, también con sede en el área de Seattle, lo ayudó a aplicar el pensamiento digital a su nuevo rol en un tipo de organización muy diferente.

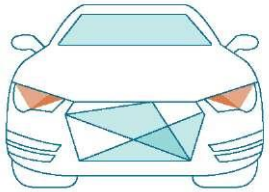
Johnson habla claramente sobre el recorrido digital de Starbucks: "Cuando otros intentan crear una aplicación móvil, Starbucks ha creado una plataforma de consumidor de un extremo a otro basada en la lealtad".

La principal innovación digital de la empresa se centra en su Aplicación móvil para pedidos y pagos. Centrarse en la aplicación es fundamentalmente una estrategia basada en el cliente, ya que aborda los deseos básicos del consumidor: conveniencia, evitación de líneas y velocidad de cumplimiento, etc. Junto con su extenso programa de lealtad, la aplicación brinda a Starbucks el lugar perfecto para vender y comercializar a los consumidores. Igual de importante es que la aplicación canaliza grandes cantidades de datos de usuario a la empresa, lo que le permite comprender mejor los hábitos y deseos de sus clientes.

Starbucks invirtió mucho en la creación de puntos de contacto digitales para sus clientes y, con su enorme presencia global, la disponibilidad de las aplicaciones, en términos de tiempo de actividad y latencia sin procesar, fue fundamental.



Centrarse en la aplicación es fundamentalmente una estrategia basada en el cliente, ya que aborda los deseos básicos del consumidor.

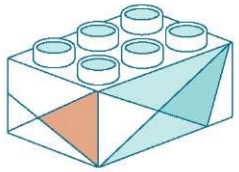


AUDI: ¿AUTOMOTRIZ O CORPORACIÓN DIGITAL?

La industria automotriz, que ya es altamente competitiva, se enfrenta a una fuerza disruptiva masiva a corto y mediano plazo. Los nuevos modelos de ventas, el auge de los vehículos eléctricos y la conducción autónoma están cambiando las reglas del juego para los fabricantes de automóviles. Ante estos retos, Audi ha cambiado la forma en que se vende sus vehículos.

Lanzado en 2012, Audi City ofrece una experiencia de marca profunda que permite a los visitantes explorar virtualmente toda la gama de productos Audi, incluso en tiendas del centro de la ciudad sin suficiente espacio para salas de exhibición.

Audi es una marca de lujo, y la decisión de la compañía de alterar su propio canal de ventas no se tomó a la ligera. Audi invirtió mucho en la creación de una experiencia minorista virtual que fuera tan auténtica como la física. Parte de este proceso incluyó la utilización de varios puntos de contacto diferentes, factores de forma de aplicación y enfoques de visualización. Hacer todo esto dentro de las expectativas de los usuarios de una experiencia rápida fue un tramo tecnológico que requirió un nuevo pensamiento.



LEGO: DE PIEZAS DE PLÁSTICO A BLOQUES DIGITALES

El grupo LEGO es el famoso fabricante danés de los juguetes para niños del mismo nombre. Pero después de un largo período de expansión de 1970 a 1991, LEGO sufrió un declive constante en su negocio de 1992 a 2004. En 2004, la empresa se encontraba al borde de la quiebra.

Al llegar a un punto de inflexión, LEGO se vio obligado a iniciar una importante reestructuración. Su transformación digital se centró en nutrir nuevas fuentes de ingresos provenientes de películas, juegos móviles y aplicaciones móviles.

Cuando LEGO se embarcó en este proceso, una de las limitaciones clave que tuvo que superar fue el impacto en el rendimiento de decenas de miles de niños que usaban simultáneamente sus diversas aplicaciones y juegos de LEGO. La administración dictaba que la velocidad, de la innovación y de la entrega de sus productos digitales era un requisito no negociable.

La administración de LEGO dictaminó que la velocidad de la innovación y de la entrega de sus productos digitales era un requisito no negociable.

Las dos épocas de la entrega digital

Las organizaciones han experimentado dos épocas en lo que respecta a la entrega digital. Primero tuvieron que lidiar con la Época de Disponibilidad. Hoy, a medida que la disponibilidad se convierte en gran parte en un problema resuelto, están entrando en la Época de la Velocidad.

LA ÉPOCA DE DISPONIBILIDAD: EL TIEMPO DE ACTIVIDAD ES CLAVE

Con la llegada del Internet y la creación de empresas como Amazon, eBay y Netflix, las corporaciones comenzaron a explorar el potencial de estas nuevas tecnologías y modelos comerciales. En los primeros días de la transformación digital, los equipos de TI perseguían principalmente una métrica singular: el tiempo de actividad. Las organizaciones que se movían hacia el mundo digital tenían un enfoque: garantizar que sus sitios web y aplicaciones estén disponibles en cualquier lugar y en cualquier momento. Estos tiempos, a los que nos referimos como la Época de Disponibilidad, se caracterizaron por herramientas y enfoques que aseguraron la confiabilidad del sitio.

La época de disponibilidad fomentó una gran cantidad de innovación, todo en un esfuerzo por aumentar el número de 9 en la métrica de porcentaje de tiempo de actividad.
La conversión de la función de desarrollo y operaciones en la función combinada de DevOps tenía la intención de acelerar el desarrollo de aplicaciones y aumentar la confiabilidad. Se crearon poderosas plataformas y herramientas de monitoreo de aplicaciones e infraestructura para lograr este santo grial: porcentajes de tiempo de actividad cada vez más altos en un entorno que evoluciona más rápidamente.

De hecho, si bien el objetivo de los "cinco nueve" es fácil de mencionar, es importante comprender lo que realmente significa el 99,999 % de tiempo de actividad: no más de unos simples 26 segundos de tiempo de inactividad por mes.A medida que más y más empresas se acercan o logran estadísticas de tiempo de actividad como esta a través de ingeniería de alta calidad y una comprensión profunda de lo que se necesita para planificar una falla, los CIO han podido enfocarse en otras áreas de mejora. En consecuencia, áreas que alguna vez fueron ignoradas ahora se están volviendo críticas.

A medida que la disponibilidad se convierte en gran parte en un problema resuelto, las organizaciones están entrando en la Época de la Velocidad.

LAS ESTADÍSTICAS QUE INDICAN EL FINAL DE LA ÉPOCA DE DISPONIBILIDAD

Las organizaciones han pasado la última década o dos diciendo que, a medida que avanzan hacia puntos de contacto más digitales con los clientes, la disponibilidad fundamental de esos puntos de contacto es clave. Toda una generación de profesionales de TI se ha obsesionado con las métricas de disponibilidad y las herramientas para mejorarlas.

Sin embargo, hay algunos factores fundamentales que cambian las reglas del juego para estos profesionales. Además del aumento de la complejidad que han creado con sus propios esfuerzos para diseñar el tiempo de actividad, también existen factores externos que impulsan los requisitos críticos para la latencia más baja posible.

A medida que los consumidores se trasladan en masa a los puntos de contacto móviles, la forma en que consumen los datos y sus requisitos de inmediatez están cambiando. Los consumidores utilizan sus dispositivos móviles para informarse mejor sobre los productos y servicios que son importantes para ellos. El 80 % de los consumidores buscan información de productos, reseñas y precios en sus teléfonos inteligentes mientras compran en una tienda física.

Y esta tendencia a consumir información es solo el comienzo; los consumidores también están realizando transacciones de nuevas formas. Un tercio de todas las compras durante la temporada de compras navideñas de 2018 se realizaron en teléfonos inteligentes.

Desafortunadamente, las organizaciones tienden a sobrestimar su propia capacidad para ofrecer buenas experiencias. La investigación de Qualtrics encontró que, si bien el 60 % de las empresas piensan que están brindando una buena experiencia móvil, solo el 22 % de los consumidores sienten lo mismo.

Todo esto apunta a esta necesidad de velocidad: la navegación móvil ocurre en diferentes contextos desde la navegación fija, mientras se camina, en la tienda y en breves descansos; todos estos contextos exigen más velocidad más que nunca.

UNA HISTORIA DE PRECAUCIÓN: EL PÉSIMO LANZAMIENTO DE DISNEY

El año pasado, Disney apostó gran parte de su éxito futuro en el lanzamiento de Disney+, el servicio de transmisión de video de alto perfil de la compañía. Al igual que muchas organizaciones que buscan causar un gran revuelo en un área fuera de su forma de entrega habitual, Disney habló sobre el lanzamiento y promocionó a los clientes por la experiencia que estaban a punto de ver.

Desafortunadamente, en cuanto se lanzó Disney+, los clientes comenzaron a quejarse del mal desempeño del servicio: el almacenamiento en búfer extendido, los abandonos y la latencia general obstaculizaron lo que podría haber sido un día espléndido de lanzamiento. La crítica fue clara: un servicio que ofrece poca velocidad es tan malo como uno que no está disponible en lo absoluto.

LA ÉPOCA DE LA VELOCIDAD: EL QUE SE DUERME, PIERDE

En los últimos años, la mayoría de las empresas han adquirido un buen conocimiento del tiempo de actividad. Mientras tanto, sus proveedores de servicios han hecho mucho para incorporar múltiples redundancias en sus plataformas, garantizando que el camino hacia una disponibilidad casi perfecta sea fácil de navegar. Las herramientas de monitoreo, las prácticas de ingeniería de confiabilidad del sitio y la adopción de la resiliencia en caso de fallas inevitables han ayudado a brindar lo que los usuarios finales esperan ahora: sitios web y aplicaciones que están disponibles cuando se necesitan.

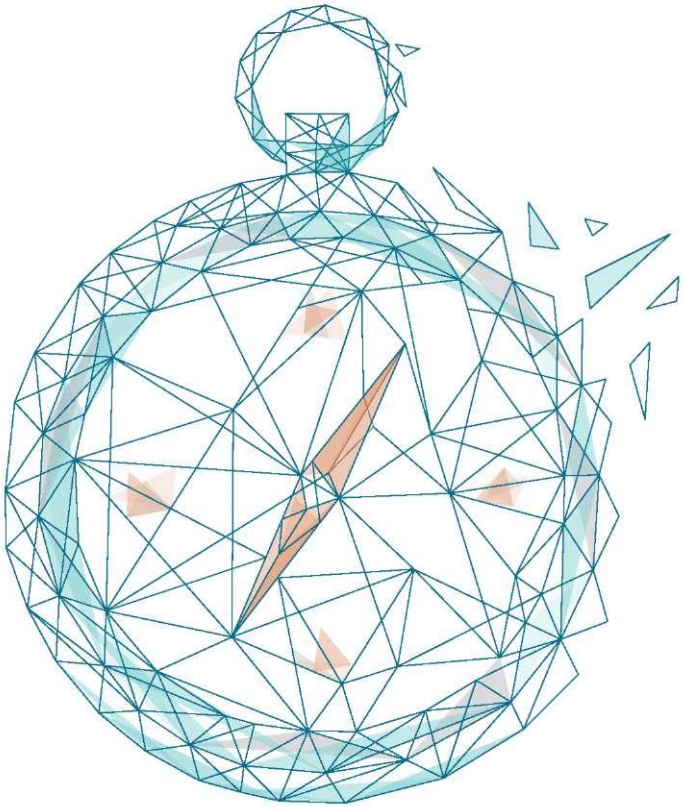
Pero toda esta ingeniería adicional y el aprovechamiento de arquitecturas cada vez más complejas en un esfuerzo por entregar las aplicaciones más resistentes, ha introducido nuevos desafíos, que son tan críticos como el tiempo de actividad.

Claramente, estamos entrando en una segunda época en la que la confiabilidad se ha convertido en algo en juego, mientras que la velocidad es ahora el diferenciador competitivo. Las decisiones de los clientes, que antes se tomaban con tiempo y análisis, se toman cada vez más en un abrir y cerrar de ojos. Y si su sitio tarda más que ese instante en cargar, o su servicio de transmisión sufre de bloqueos y pausas en el búfer, está listo para perder.

Si asume que la insatisfacción del cliente no afecta sus hábitos de consumo, piénselo de nuevo. Como se detalla en un artículo de Forbes de 2019 (¿Qué tan rápido es lo suficientemente rápido? Los tiempos de carga móvil impulsan la experiencia del cliente y las ventas de impacto): "Una página de carga lenta en un dispositivo móvil no solo pone a prueba la paciencia de los consumidores. Puede ser el "fracaso" de la experiencia del cliente lo que le cuesta una venta. Esta es la conclusión clave del Informe de velocidad de páginas de 2019. El estudio, que explora las actitudes de 1150 consumidores y empresas, expresa que la velocidad de la página es un factor decisivo en el comportamiento de compra".

Y el impacto de la baja velocidad de la página no es intrascendente: "casi el 70 % de los consumidores dice que la velocidad de la página afecta su disposición a comprar. Además, un tiempo de carga lento también reduce las posibilidades de que regresen en el futuro. Un desglose de los datos revela que el 22 % de los compradores dijo que cerraría la pestaña y el 15 % dijo que visitaría el sitio de un competidor y el 12 % le contaría a un amigo sobre su experiencia negativa".

Si la nueva época se define por la necesidad de garantizar que la latencia sea lo más baja posible, ¿qué cosas deben pensar las organizaciones para alcanzar ese objetivo?



Entrega de velocidad en un mundo complejo

En su publicación fundamental de 2013 en The Composable Enterprise™, Jonathan Murray, exdirector de tecnología de Warner Music Group, describió el futuro de la tecnología dentro del contexto de las demandas empresariales de velocidad y agilidad. Basado en la experiencia de toda su vida en la implementación de estrategias digitales de grandes corporaciones, Murray describió Composable Enterprise de esta manera: "Las funciones comerciales, los procesos, las organizaciones, las relaciones con los proveedores y la tecnología deben verse como componentes básicos que pueden reconfigurarse según sea necesario para abordar el cambiante panorama competitivo".

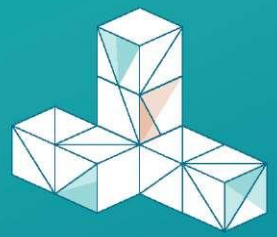
Este nuevo Modelo Operativo de Componentes (COM [Component Operating Model]) requiere un enfoque de "bloques de Lego" para diseñar e implementar procesos y las organizaciones que los respaldan. La implementación de un enfoque basado en COM tendrá un impacto profundo en la estructura de las organizaciones, la naturaleza del trabajo.

Los diseños comerciales basados en COM crearán un estrés significativo para las organizaciones y las infraestructuras de TI tradicionales. Nuestros servicios de TI actuales se crearon para servir a un modelo operativo estático y, a menudo, funcionan con el enfoque de silos. La TI necesita volverse mucho más adaptable dinámicamente para mantenerse al día con la velocidad de los negocios en la actualidad.

"Se requerirá un nuevo enfoque del Modelo de Arquitectura de Componentes (CAM [Component Architecture Model]) para la infraestructura, las aplicaciones y los servicios de TI para garantizar que la TI pueda brindar lo que la empresa necesita. El tiempo entre la identificación de una necesidad comercial y la entrega de la solución de TI requerida se convierte en horas y días en lugar de meses y años".

Escrita hace varios años, la profética publicación de Murray describe la nueva situación normal dentro de las organizaciones. Hemos visto, en los últimos años, un cambio radical en la forma en que se usa la infraestructura y se construyen las aplicaciones. Con el auge de contenedores, arquitecturas de microservicios, herramientas de aplicaciones modulares discretas y similares, mantener una aplicación en funcionamiento y garantizar que funcione bien significa hacer malabares con decenas de servicios, regiones, geografías, proveedores de servicios y más.

Entonces, si bien toda esta capacidad de composición impulsa la productividad del desarrollador y la agilidad organizacional, tiene un costo. Parecería que ofrecer una baja latencia en estas condiciones es una quimera.



Hemos visto, en los últimos años, un cambio radical en la forma en que se usa la infraestructura y se construyen las aplicaciones.

Datos rápidos en un entorno informático distribuido

Como hemos visto en el trabajo fundamental de Murray sobre componibilidad en aplicaciones e infraestructuras modernas, ya no tenemos una simple pila monolítica sobre la que se construyen las aplicaciones. Más bien, en un esfuerzo por brindar a los desarrolladores y sus organizaciones la mayor flexibilidad y la mayor velocidad posibles, aprovechamos una gran cantidad de servicios de desarrollador modulares, diferentes patrones de infraestructura, varios enfoques de alojamiento y una distribución geográfica masiva de aplicaciones. Mientras tanto, intentamos entregar estas aplicaciones lo más rápido posible a los usuarios de todo el mundo.

En esta época de enorme complejidad, sería fácil pensar que no existe un tejido común en el que las organizaciones puedan confiar: su mundo parece perpetuamente fluido y en constante cambio.

Sin embargo, hay un hilo común que se abre paso a través de todas las cosas diferentes que hace la organización, y son los datos. Al pensar en una capa de datos como un hilo coherente y unificado aprovechado por todas las demás partes de la pila, permitimos que las organizaciones le den sentido al caos. Al elegir una capa de datos diseñada para distribuir entornos que muestran los tiempos de procesamiento más rápidos, y que ofrecen la mejor resistencia de su clase, podemos ofrecer exactamente lo que necesita una empresa.

Una forma clave en que las organizaciones pueden garantizar que sus aplicaciones sean resistentes y rápidas es trabajar dentro de una estructura de capa de datos coherente. Y los datos consistentes comienzan con una base de datos que puede ofrecer objetivos aparentemente imposibles: arquitecturas distribuidas, coherencia, flexibilidad y velocidad.

Enfoques modernos para reducir la latencia

Como hemos visto, las aplicaciones se crean cada vez más utilizando microservicios: aprovechando una multitud de componentes diferentes, con diferentes enfoques de infraestructura, alojados en una variedad de ubicaciones diferentes, consumidos por personas de todas partes y distribuidos en muchas plataformas diferentes.

Con datos ubicados en tantos lugares y transmitidos a través de tantas redes diferentes, no es de extrañar que haya muchas oportunidades para que ocurran conflictos de datos. Para hacer frente a estos conflictos, se desarrollaron tipos de datos replicados sin conflictos (CRDT) para permitir que los datos se repliquen en múltiples ubicaciones.

Con los CRDT, las réplicas individuales se pueden actualizar de forma independiente y simultánea sin ninguna coordinación entre ellas. Sin los CRDT, actualizaciones simultáneas de múltiples réplicas de los mismos datos, sin coordinación entre las computadoras que alojan las réplicas, pueden producirse inconsistencias entre las réplicas.

Sin embargo, con los CRDT, se puede resolver cualquier incoherencia que resulte de este enfoque distribuido. Los CRDT se utilizaron inicialmente en situaciones en las que la distribución masiva es la norma (sistemas de chat en línea, juegos de azar en Internet y transmisión de audio y video), pero cada vez se utilizan más en aplicaciones más genéricas.

Existe una tecnología subyacente significativa que se utiliza para hacer que un CRDT funcione, pero la forma más sencilla de pensar en ello es que un CRDT proporciona una capa de datos mediante la cual las réplicas pueden actuar de forma autónoma y aun así brindar consistencia.

Con datos ubicados en tantos lugares y transmitidos a través de tantas redes diferentes, no es de extrañar que haya muchas oportunidades para que se produzcan conflictos de datos.

Superando el caché

En los modelos de base de datos tradicionales, el sitio de la base de datos está separado del caché. Piense en la base de datos como la biblioteca municipal principal y el caché como la biblioteca sucursal local, donde se guardan los libros más populares para satisfacer las demandas más comunes de los prestatarios. Si los libros más populares son consistentes, eso podría funcionar bien, pero a medida que los hábitos de lectura cambian y los libros nuevos entran y salen de su favor, eso se vuelve más difícil.

Así esta noción de verificar rápidamente diferentes piezas de información de forma constante es solo la metáfora de las aplicaciones modernas: toda la componibilidad de la que habló Murray da como resultado que se tenga que acceder a los datos desde la base de datos desde muchos servicios y lugares diferentes, y en muchos tiempos diferentes.

En un mundo con servicios cada vez más discretos y, por lo tanto, cada vez más lugares donde algo puede salir mal, el modelo tradicional no es ideal (ver diagrama a continuación). Y en aplicaciones en las que los modelos de datos tienen más que ver con la transferencia de muchos fragmentos de información del tamaño de un bocado, es posible que el modelo de caché no sea la forma más rápida de llevar los datos a donde deben ir.

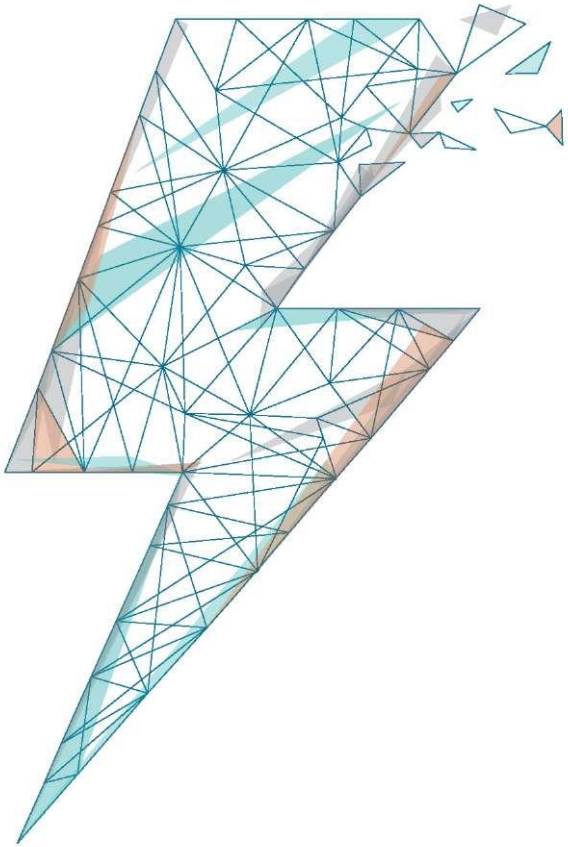
Aquí es donde entra en juego la noción de una sola capa de datos: al aprovechar una sola capa para reemplazar la combinación de base de datos/caché, se reduce la complejidad de la capa de datos. A cambio, lo que está sobrealimentado es la aplicación distribuida y modular que es la norma en la actualidad.

La ventaja adicional es que la reducción del número de partes en la capa de datos también reduce la latencia. Si bien las partes individuales de una capa de datos compleja pueden ser rápidas, tener un único almacén de datos reduce el número de saltos de red que invariablemente ralentiza las cosas.

Por tanto, en lugar del caché, muchas bases de datos modernas aprovechan las técnicas en memoria donde se usa la memoria, en lugar de los discos externos, para el almacenamiento. Esto es fundamental, ya que con todo lo almacenado en la memoria, la velocidad no está limitada por múltiples capas de almacenamiento. Con un modelo basado en caché, lo que se almacena en caché se convierte en el cuello de botella que limita la velocidad general.

En un modelo tradicional, acceder a la apariencia de los datos significa que la aplicación tiene que:

Velocidad: un byte a la vez



Las bases de datos tradicionales, como hemos visto anteriormente, dependen de la memoria externa para su caché. Hasta hace muy poco y durante los últimos 50 años, el almacenamiento se realizaba en discos giratorios físicos y, por lo tanto, la mayoría de los enfoques de bases de datos tradicionales estaban optimizados para esto.

Pero, dado que los discos duros son dispositivos físicos, tienen limitaciones creadas por el mundo físico. Para sortear estas limitaciones físicas, se crearon una serie de limitaciones operativas. Si bien es un desvío técnico, los detalles precisos de la tecnología de discos físicos tienen un gran impacto en la velocidad de la base de datos.

Sin embargo, la clave del asunto es que la línea entre el almacenamiento moderno y la memoria se está volviendo menos clara. El auge de las unidades de estado sólido (SSD) y otros nuevos enfoques de almacenamiento significan que estas soluciones de ingeniería, diseñadas para un mundo limitado por la velocidad física de los dispositivos mecánicos, ya no son necesarias. También significa que el almacenamiento se puede dividir en niveles, de modo que todos los datos se pueden guardar en un almacenamiento rápido y termina la necesidad de una memoria caché separada.

El resultado neto, para aquellos que intentan diseñar para la velocidad, es una capa de datos más rápida sobre la cual desarrollar nuestras aplicaciones.

El surgimiento de nuevos enfoques de almacenamiento significa que las soluciones de ingeniería, diseñadas para un mundo limitado por la velocidad física de los dispositivos mecánicos, ya no son necesarias.

Escalabilidad sin obstaculizar la velocidad

Está muy bien crear una aplicación que se ejecute rápidamente con un uso limitado, pero ¿qué sucede cuando su rendimiento aumenta enormemente? Este es el problema que todo desarrollador de aplicaciones, en busca de aceptación y viralidad, espera enfrentar.

Pero la escalabilidad ocurre de dos maneras: **hacia arriba,** en términos de la cantidad de datos que se transfieren a través de la capa de datos, pero también *en términos* de la gran cantidad de información que existe.

Las organizaciones necesitan construir una capa de datos que permita esta escalabilidad por etapas y sin complicaciones. Esto implica pensar en varios factores diferentes: la capacidad de ejecutar la capa de datos en múltiples ubicaciones, la capacidad de usar diferentes tipos de memoria y almacenamiento, la capacidad de clasificar los datos en niveles según su regularidad de uso y, finalmente, la capacidad de escalar globalmente.

Pasemos a esta última área. Toda esta capacidad para almacenar y procesar en memoria es buena, pero si su aplicación necesita extenderse por todo el mundo, ¿puede seguir disfrutando de este mismo bajo nivel de latencia?

ES UN MUNDO DE MULTINÚCLEOS

El procesamiento moderno ocurre cada vez más con un contacto multinúcleo. Multinúcleo se refiere a la informática en la que existen dos o más unidades de procesamiento individuales dentro de una CPU. Las instrucciones que se envían a la CPU se pueden procesar en núcleos separados al mismo tiempo, lo que aumenta la velocidad general.

Aprovechar las arquitecturas de múltiples núcleos puede ser un desafío. Las organizaciones que deseen aprovechar una capa de datos que pueda escalar de la manera más eficiente deben pensar en esto. ¿Su capa de datos puede escalar horizontalmente en un solo clúster para ofrecer la mejor escala con la latencia más baja?

AMPLIACIÓN ESCALABILIDAD

Las organizaciones necesitan construir una capa de datos que permita esta escalabilidad por etapas y sin complicaciones.

UN PASEO POR LA CALLE CAP

Dado que este documento será inevitablemente utilizado por aquellos que aspiran a crear aplicaciones distribuidas globalmente que muestren un rendimiento similar al de las localizadas, vale la pena analizar algunas limitaciones en torno a las capas de datos distribuidos.

Hace unos 20 años, el científico informático Eric Brewer desarrolló el teorema CAP, que se relaciona con las aplicaciones distribuidas y, específicamente, los datos que esas aplicaciones crean y consumen.

El teorema CAP, en términos más simples, afirma que cualquier sistema de datos compartidos en red puede tener solo dos de las tres propiedades deseables; consistencia (C) equivalente a tener una única copia actualizada de los datos; alta disponibilidad (A) de esos datos (para actualizaciones); y tolerancia a las particiones de la red (P).

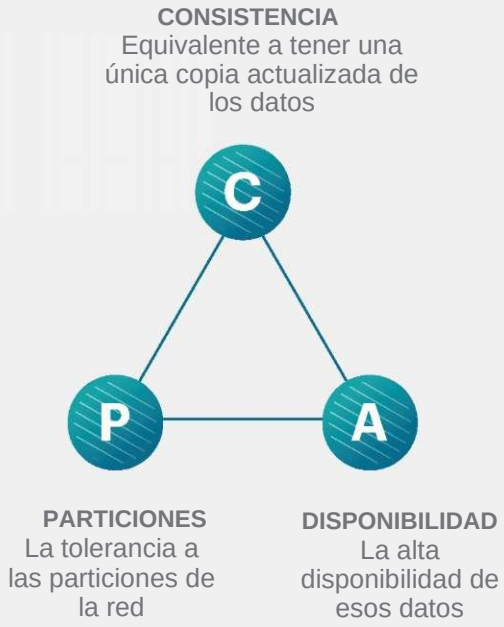
Y con ello, en esos primeros días en los que la búsqueda de la velocidad era todo, el teorema de CAP significaba que los enfoques con mayor probabilidad de dar las velocidades más rápidas y la disponibilidad de la aplicación (particiones de red y alta disponibilidad) también daría lugar a inconsistencias de datos.

Sin embargo, en las décadas transcurridas desde la introducción del teorema CAP, se han desarrollado nuevos enfoques para manejar sistemas distribuidos que permiten esa hazaña teóricamente imposible: consistencia, disponibilidad y tolerancia a la partición de los datos. El surgimiento de nuevos enfoques de datos significa que podemos tener una latencia baja sin renunciar a la coherencia de los datos.

Si bien este no es el lugar para un tratado técnico definitivo, es importante que quienes tienen la responsabilidad de la aplicación de sus organizaciones comprendan los rudimentos de cómo funcionan las aplicaciones modernas.

Como se señaló, en un mundo donde las aplicaciones están, por necesidad, distribuidas, habrá múltiples nodos incluidos en muchas aplicaciones individuales. En esta situación multinodal, hay dos amplias opciones: **datos activos-pasivos** o **datos activos-activos.**

EL TEOREMA CAP



El surgimiento de nuevos enfoques de datos significa que podemos tener una latencia baja sin renunciar a la coherencia de los datos.

CAPA DE DATOS UNIFICADOS

Los planos de datos, la parte del software que procesa las solicitudes de datos, pueden ser Activo-Activo o Activo-Pasivo.

Activo-Activo (también llamado a veces activo-dual) es un enfoque mediante el cual cada nodo tiene acceso a una base de datos replicada que le da a cada nodo acceso y uso de una sola aplicación. Esta tecnología es lo que permite la capacidad de mantener los datos consistentes para sus aplicaciones en diferentes entornos (servidores, nubes híbridas, multinube) e incluso aplicaciones que se distribuyen en todo el mundo. En un sistema activo-activo, todas las solicitudes se equilibran en carga en toda la capacidad de procesamiento disponible. Cuando ocurre una falla en un nodo, otro nodo de la red ocupa su lugar.

Un clúster Activo-Activo generalmente se compone de al menos dos nodos, ambos ejecutando activamente el mismo tipo de servicio simultáneamente. Debido a que hay más nodos disponibles para servir, también habrá una mejora notable en el rendimiento y los tiempos de respuesta en comparación con un enfoque Activo-Pasivo.

ACTIVO-PASIVO

Un clúster activo-pasivo también consta de al menos dos nodos. Sin embargo, como implica el nombre “Activo-Pasivo”, no todos los nodos están activos. En una clúster con dos nodos, por ejemplo, si el primer nodo ya está activo, el segundo nodo debe ser pasivo o estar en espera. El nodo pasivo (también conocido como conmutación por error) sirve como respaldo que está listo para asumir el control tan pronto como el servidor activo (también conocido como primario) se desconecte o no pueda dar servicio.

Cuando los clientes se conectan a un clúster de dos nodos en configuración Activo-Pasivo, se conectan a un solo servidor. En otras palabras, todos los clientes se conectan al mismo servidor. Al igual que en la configuración Activo-Activo, es importante que los dos servidores tengan exactamente la misma configuración. Esto se denomina redundancia y garantiza que los datos se puedan replicar sin problemas entre los nodos.

Si se realizan cambios en la configuración del servidor principal, esos cambios deben conectarse en cascada al servidor de conmutación por error. Entonces, cuando la conmutación por error actúe, los clientes no podrán notar la diferencia.

Si la latencia es el nuevo apagón, es evidente que cuanto más cerca esté un nodo del usuario de la aplicación, menores serán las cifras de latencia. Por lo tanto, necesitamos encontrar una manera de distribuir aplicaciones globalmente (ya que distribuir nodos cerca de los usuarios de la aplicación reduce la latencia), sin dejar de garantizar la coherencia. Afortunadamente, contamos con algo de ayuda en este sentido.

CONSTRUIDO CON LA VELOCIDAD EN MENTE

La replicación libre de conflictos es una noción que permite que existan múltiples copias (réplicas) de datos en múltiples ubicaciones de manera consistente. Es un método muy importante para asegurar una baja latencia para aplicaciones distribuidas, pero hay otros aspectos a considerar. Como se señaló anteriormente, las bases de datos modernas diseñadas para ofrecer la latencia más baja para las aplicaciones modernas almacenan datos en la memoria. Al eliminar la necesidad de un caché externo, podemos reducir la cantidad de tráfico de datos requerida.

Si bien las bases de datos tradicionales se diseñaron para casos de uso en los que el tiempo de procesamiento de 10 o 100 milisegundos era aceptable, en el mundo actual, donde se requieren tiempos de respuesta de aplicaciones instantáneos, el rendimiento en menos de milisegundos es una necesidad.

FALLAR ESTÁ BIEN SI SE HACE RÁPIDO

La conmutación por error, como su nombre lo indica, es un sistema automatizado mediante el cual, en caso de que un nodo falle por alguna razón, otro nodo replicado toma el relevo. Si bien la conmutación por error es fácil de diseñar, la velocidad de esa conmutación es lo que determina los impactos de la falla en el usuario final.

Para garantizar la latencia más baja en un mundo en el que las fallas de nodos pueden ser inevitables, es importante que la capa de datos multinodales pueda ofrecer una conmutación por error lo más rápido posible.

Resumen

En el mundo moderno, las organizaciones, impulsadas por ofrecer experiencias digitales, deben asegurarse de que sus partes interesadas puedan utilizar las aplicaciones cuando y donde lo deseen. Pero los usuarios de hoy exigen no solo un acceso continuo, sino también un rendimiento virtualmente instantáneo. La latencia, en un mundo que pasa de la época de la disponibilidad a la época de la velocidad, puede ser tan mala como la falta de disponibilidad de las aplicaciones.

Afortunadamente, hoy tenemos opciones que simplemente no estaban disponibles hace una década. Se han superado muchos impedimentos para entregar aplicaciones rápidas, entre ellos el teorema CAP. Y ahora, las organizaciones tienen la capacidad de aprovechar una capa de datos que está libre de conflictos independientemente de cuántas réplicas se utilicen.

Al aprovechar las bases de datos que funcionan completamente en la memoria y ejecutarlas de manera activa-activa, entregamos bases de datos más rápidas que las que estaban disponibles anteriormente y brindamos la baja latencia que exigen los usuarios de aplicaciones de hoy.

Esto debe considerarse un asunto urgente para todas las organizaciones: sus competidores y sus disruptores están entregando aplicaciones rápidas y sus clientes las exigen; usted no puede darse el lujo del tiempo.

Acerca del autor - Ben Kepes

Ben Kepes es analista de tecnología, comentarista y consultor. Durante la última década y media, ha acumulado un gran número de seguidores como experto en la materia. Es reconocido mundialmente en las áreas de computación en la nube, tecnología empresarial y transformación digital.

Los comentarios de Ben han sido ampliamente publicado en medios como Forbes, Wired y The Guardian y ha sido invitado a hablar en diversas conferencias de tecnología, negocios e interés general.

@benkepes

redislabs
HOME OF REDIS

Sobre Redis Labs

Las empresas modernas dependen del poder de los datos en tiempo real. Con Redis Labs, las organizaciones brindan experiencias instantáneas de una manera altamente confiable y escalable.

Redis Labs es el hogar de Redis, la base de datos en memoria más popular del mundo y el proveedor comercial de Redis Enterprise, que ofrece un rendimiento superior, una confiabilidad incomparable y una flexibilidad inigualable para la personalización, aprendizaje automático, internet de las cosas, búsqueda, comercio electrónico, soluciones sociales y de medición en todo el mundo.

Redis Labs, constantemente clasificado como líder en informes de analistas superiores sobre NoSQL, bases de datos en memoria, bases de datos operativas y base de datos como servicio (DBaaS), es confiable por más de 7400 clientes empresariales, incluidas cinco compañías Fortune 10, tres de los cuatro emisores de tarjetas de crédito, tres de las cinco principales empresas de comunicaciones, tres de las cinco principales empresas de atención médica, seis de las ocho principales empresas de tecnología y cuatro de las principales siete minoristas.

Redis Enterprise, disponible como servicio en nubes públicas y privadas, como software descargable, en contenedores y para implementaciones en la nube híbrida/local, potencia los casos de uso populares de Redis, como transacciones de alta velocidad, administración de trabajos y colas, almacenamiento de sesiones de usuario, ingesta de datos en tiempo real, notificaciones, almacenamiento en caché de contenido y datos de series de tiempo.

Sede corporativa

700 E El Camino Real Suite 250
Mountain View, CA 94040

Tel.: +1 (415) 930-9666
redislabs.com

Síganos