redislabs
HOME OF REDIS

Latência é a nova indisponibilidade

POR QUE A VELOCIDADE É A APOSTA DA VEZ

Sumário executivo

À medida que cada vez mais organizações usam aplicativos hospedados em diferentes infraestruturas, diferentes localizações e com diferentes provedores, a velocidade de acesso a esses aplicativos para o usuário final é colocada em xeque.

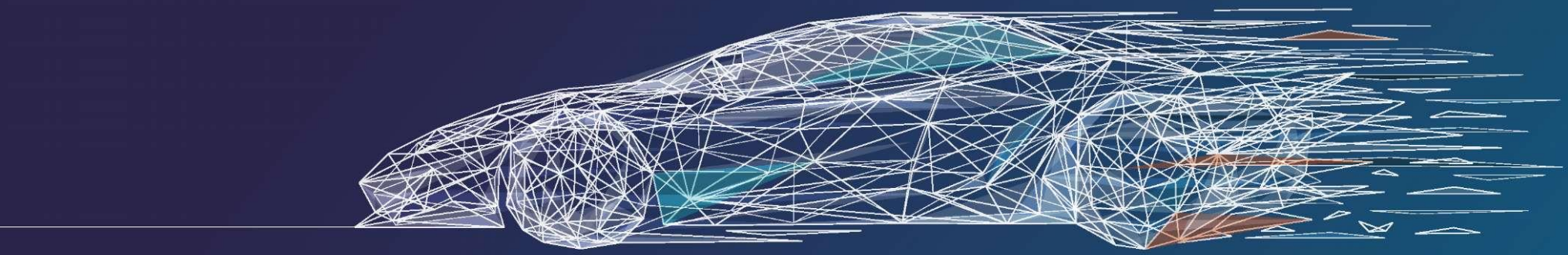
Some a isso o fato de que os aplicativos hoje são formados por um número significativo de diferentes partes de componentes e você tem a receita para uma experiência de usuário prejudicada.

Essas características duplas — modularidade do aplicativo e complexidade da infraestrutura — podem resultar diretamente em baixo desempenho do aplicativo.

É por esse motivo que a velocidade da camada de dados, a camada horizontal comum em todo o aplicativo, é crucial.

Poder usar uma camada de dados replicada geograficamente, evitando os problemas de inconsistência de dados, é um desafio que precisa ser enfrentado por todos os líderes de TI.

Ao aproveitar uma camada de dados que unifique seus dados em nuvens e ao redor do mundo, as organizações podem superar algumas das limitações inerentes que desafiam as equipes de tecnologia há décadas e oferecer melhores experiências aos usuários finais.



Introdução

As equipes digitais passaram a última década garantindo que seus ativos digitais estivessem disponíveis o tempo todo, e tiveram muito sucesso! A alta disponibilidade se tornou o padrão.

Em parte, as organizações alcançaram esse alto nível de digitalização e alta disponibilidade aproveitando os benefícios que a nuvem traz — facilidade de escalabilidade, serviços modulares, padrões de arquitetura mais refinados. Todas essas características permitem resultados positivos, mas aumentam a complexidade em contrapartida. Essa complexidade foi inicialmente mais impactante em termos de disponibilidade e deu origem ao que chamamos de “época da disponibilidade”. No entanto, à medida que as organizações entendem melhor como oferecer alta disponibilidade, elas descobrem que ainda há outros problemas a serem resolvidos.

Mas agora a época da disponibilidade está começando a diminuir, com cada vez mais organizações procurando a redução de latência como próximo passo para conseguirem os resultados que procuram. Cada vez mais elas entendem que, se for para disponibilizar produtos e serviços lentos, é melhor nem disponibilizá-los — latência é a nova indisponibilidade.

Infelizmente, resolver problemas de latência costuma ser mais difícil do que criar alta disponibilidade. Embora a disponibilidade possa ser melhorada por meio de boa engenharia, maiores níveis de redundância e melhor monitoramento e visibilidade, o enigma da latência é restringido pelas próprias leis da física.

Para reduzir a latência, as organizações precisam entender o que é latência e os fatores que contribuem para ela, além de ter diretrizes claras e definitivas para reduzir a latência ao mínimo possível para os usuários de seus aplicativos e sites.

Se a latência é a nova indisponibilidade, aqui está a inteligência de que você precisa para fornecer a menor latência fisicamente possível.

Cada vez mais organizações e serviços entendem que, se for para disponibilizar produtos e serviços lentos, é melhor nem disponibilizá-los — latência é a nova indisponibilidade.

Não é o grande que come o pequeno, é o rápido que come o lento

O ritmo acelerado é o novo normal. Embora análise e prudência já tenham sido a regra, a realidade operacional hoje é que, para ficar à frente de seus concorrentes, as organizações precisam inovar mais rápido do que nunca. O panorama operacional de cada organização está mudando rapidamente e o sucesso virá para aquelas que melhor reagirem a esse dinamismo.

POR QUE AGILIDADE É TÃO IMPORTANTE: O TEMPO É CRUCIAL

O mundo em que vivemos hoje é muito diferente daquele de apenas alguns anos atrás, e as mudanças são cada vez mais rápidas. Nesse contexto, é mais importante do que nunca que as organizações ajam com rapidez. É importante compreender as mudanças que estão ocorrendo na sociedade para entender melhor o valor de agir com rapidez.

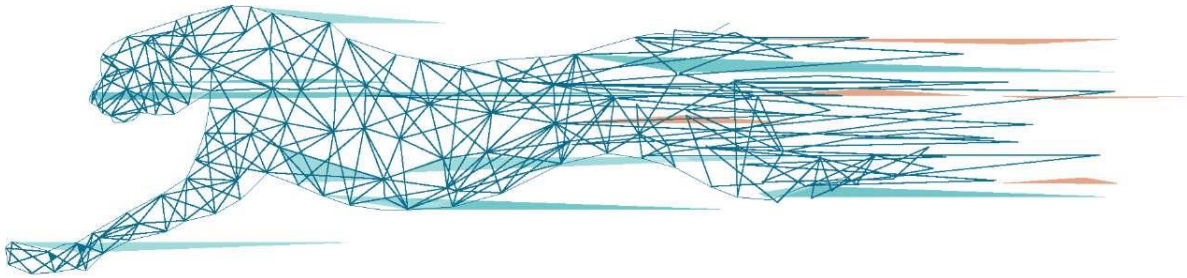
Em 2011, o notável empresário, investidor e membro do conselho Marc Andreessen (o inventor do navegador Netscape) escreveu um artigo de opinião agora famoso para o ***The Wall Street Journal,*** explicando "por que os softwares estão comendo o mundo" (título original: "Why Software Is Eating the World").

Em seu ensaio, Andreessen apresentou sua teoria sobre o tamanho, escopo e velocidade dessa mudança, sugerindo que: "Estamos no meio de uma grande e dramática mudança tecnológica e econômica, com as empresas de software prestes a assumir grandes partes da economia."

Mas, embora a ocorrência dessa mudança seja de grande importância, é a velocidade da mudança o mais relevante aqui. Fundamental para a capacidade de mover-se rapidamente é a capacidade de fazer alterações, aproveitar uma variedade de ferramentas e oferecer a melhor experiência ao usuário final. Se tudo isso lembra muito a maneira como as empresas do Vale do Silício funcionam, faz sentido. O fato é que muitas das organizações que alcançam o sucesso disruptando os setores tradicionais estão cada vez mais parecendo importantes empresas de tecnologia.

Já faz quase dez anos que Andreessen escreveu aquele ensaio e muitas de suas previsões se cumpriram. Embora tenha se tornado clichê usar Tesla, Uber, Lyft, Netflix e Airbnb como exemplos de disrupção digital, é seguro dizer que os executivos de empresas de táxi e hospitalidade foram atingidos por uma onda de proporções sem precedentes. Além do clichê, no entanto, é importante observar o quanto essas empresas se esforçam para oferecer a experiência mais ágil possível ao cliente em seus aplicativos: a velocidade realmente importa.

É evidente que mover-se rapidamente em um ambiente organizacional pressupõe o oferecimento de experiências digitais que exibam esses atributos de velocidade. A latência é o novo obstáculo para a transformação organizacional.



Mudando o mundo com a mudança para o digital

Um grande número de exemplos de organizações tradicionais está depositando suas esperanças de sucesso futuro em uma mudança para o digital. Vale a pena analisar alguns exemplos para ter uma noção da escala disso.

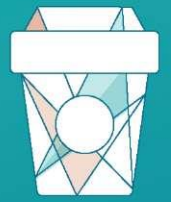
CALOURO DIGITAL: STARBUCKS NA DIANTEIRA DO DIGITAL

Kevin Johnson, o CEO da Starbucks, já foi executivo da Microsoft. Sua experiência na gigante empresa de tecnologia — também sediada na área de Seattle — o ajudou a aplicar o pensamento digital ao seu novo cargo em um tipo muito diferente de organização.

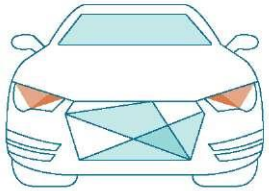
Johnson fala sobre a jornada digital da Starbucks de forma clara: “Enquanto os outros tentam construir um aplicativo móvel, a Starbucks construiu uma plataforma de consumidor ponta a ponta baseada na fidelidade.”

A principal inovação digital da empresa gira em torno da sua aplicação de pedido e pagamento móvel. O foco no aplicativo é essencialmente uma estratégia que prioriza o cliente, pois atende às necessidades básicas dele: conveniência, evitar filas e velocidade de atendimento, entre outras. Juntamente com seu extenso programa de fidelidade, o aplicativo proporciona ao Starbucks o canal perfeito para aumentar as vendas e atrair os clientes. Tão importante quanto isso é o fato de que o aplicativo direciona grandes quantidades de dados do usuário para a empresa, permitindo que ela entenda melhor os hábitos e desejos de seus clientes.

A Starbucks investiu pesado na criação de pontos de contato digitais para seus clientes e, com sua enorme presença global, a disponibilidade do aplicativo — tanto em termos de tempo de atividade bruto quanto latência — era crucial.



O foco no aplicativo é essencialmente uma estratégia que prioriza o cliente, pois atende às necessidades básicas dele.

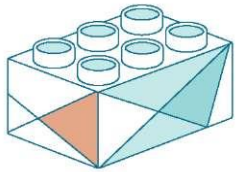


AUDI: MONTADORA OU EMPRESA DIGITAL?

A já altamente competitiva indústria automotiva enfrenta uma enorme força disruptiva no curto a médio prazo. Novos modelos de vendas, a ascensão dos veículos elétricos e a direção autônoma estão mudando o jogo para as fabricantes de automóveis. Diante desses desafios, a Audi mudou a forma como seus veículos são vendidos.

Lançada em 2012, a Audi City proporciona uma profunda experiência de marca que permite aos visitantes explorar virtualmente toda a linha de produtos Audi mesmo em lojas no centro da cidade sem espaço suficiente para showrooms.

A Audi é uma marca de luxo e a decisão da empresa de disruptar seu próprio canal de vendas não foi tomada sem critério. A Audi investiu muito na construção de uma experiência de varejo virtual tão autêntica quanto a física. Parte desse processo incluiu a utilização de vários diferentes pontos de contato, fatores de forma de aplicativo e abordagens de exibição. Fazer tudo isso atendendo às expectativas dos usuários de uma experiência veloz foi uma extensão da tecnologia que exigiu um novo pensamento.



LEGO: DOS BLOCOS DE PLÁSTICO AOS DIGITAIS

O LEGO Group é o famoso fabricante dinamarquês de brinquedos infantis com o mesmo nome. Mas após um longo período de expansão de 1970 a 1991, LEGO sofreu um declínio constante em suas vendas de 1992 a 2004. Em 2004, a empresa estava à beira da falência.

Chegando a um momento crítico, LEGO foi forçada a iniciar uma grande reestruturação. Sua transformação digital focou em nutrir novas fontes de receita provenientes de filmes, jogos para celular e aplicativos para celular.

À medida LEGO embarcava nesse processo, uma das principais limitações que precisou superar foi o impacto no desempenho causado por dezenas de milhares de crianças usando simultaneamente seus vários aplicativos e jogos LEGO. A gerência de LEGO determinou que a velocidade — de inovação e de entrega de seus produtos digitais — era um requisito inegociável.

A gerência de LEGO determinou que a velocidade — de inovação e de entrega de seus produtos digitais — era um requisito não negociável.

As duas épocas da entrega digital

As organizações passaram por duas épocas no que diz respeito à entrega digital. O primeiro desafio foi a época da disponibilidade. Hoje, com o problema da disponibilidade praticamente resolvido, as organizações estão entrando na era da velocidade.

ÉPOCA DA DISPONIBILIDADE: O TEMPO DE ATIVIDADE É O SEGREDO

Com o advento da Internet e a criação de empresas como Amazon, eBay e Netflix, as corporações começaram a explorar o potencial dessas novas tecnologias e modelos de negócios. Nos primórdios da transformação digital, o foco das equipes de TI estava em uma única métrica: o tempo de atividade. As organizações que entravam no mundo digital tinham um foco: garantindo que seus sites e aplicativos estivessem disponíveis em qualquer lugar, a qualquer hora. Esse período, que chamamos de época da disponibilidade, foi caracterizado por ferramentas e abordagens visando garantir a confiabilidade dos sites.

A época da disponibilidade promoveu muita inovação, tudo em um esforço para aumentar o valor de 9 s na métrica de porcentagem de tempo de atividade. A moldagem da função de desenvolvimento e operações na função DevOps combinada tinha o objetivo de acelerar o desenvolvimento de aplicativos e aumentar a confiabilidade. Ferramentas e plataformas poderosas de monitoramento de aplicativos e infraestrutura foram criadas para alcançar esse cenário ideal: porcentagens de tempo de atividade cada vez mais altas em um ambiente de evolução mais rápida.

Na verdade, embora a meta de “cinco-9s” seja bonita de falar, é importante entender o que realmente significa 99,999% de tempo de atividade: nada além de meros 26 segundos de inatividade por mês. À medida que mais e mais empresas aproximam-se de ou alcançam estatísticas de tempo de atividade como essa por meio de engenharia de alta qualidade e um profundo entendimento do que é necessário para se planejar contra falhas, os diretores de informática podem se concentrar em outras áreas de melhoria. Consequentemente, áreas que antes eram ignoradas agora estão se tornando cruciais.

Com o problema da disponibilidade praticamente resolvido, as organizações estão entrando na era da velocidade.

AS ESTATÍSTICAS QUE INDICAM O FIM DA ÉPOCA DA DISPONIBILIDADE

As organizações passaram as últimas duas décadas sendo informadas de que, à medida que mudam para mais pontos de contato digitais com os clientes, a disponibilidade básica desses pontos de contato é fundamental. Temos uma geração inteira de profissionais de TI obcecados por métricas de disponibilidade e ferramentas para melhorá-las.

Existem, no entanto, alguns fatores fundamentais que mudam o jogo para esses profissionais. Além da maior complexidade que seus próprios esforços de engenharia para tempo de atividade criaram, existem também fatores externos que geram requisitos críticos para a menor latência possível.

À medida que os clientes fazem a transição em massa para pontos de contato móveis, a própria maneira como eles consomem dados e suas exigências de imediatismo estão mudando. Os clientes usam seus dispositivos móveis para se informar melhor sobre os produtos e serviços que importam para eles. 80% dos clientes procuram informações, avaliações e preços dos produtos em seus smartphones enquanto fazem compras em uma loja física.

E essa tendência de consumir informação é apenas o começo; as formas de compra dos clientes também estão mudando. Um terço de todas as compras feitas na época do Natal de 2018 foram realizadas em smartphones.

Infelizmente, as organizações tendem a superestimar sua própria capacidade de entregar boas experiências. Uma pesquisa da Qualtrics descobriu que, embora 60% das empresas pensem que estão proporcionando uma boa experiência móvel, apenas 22% dos clientes sentem o mesmo.

Tudo isso aponta para essa necessidade de velocidade. A navegação móvel acontece em contextos diferentes da navegação fixa, durante a caminhada, na loja e em intervalos curtos — todos esses contextos exigem velocidade mais do que nunca.

COMO NÃO FAZER: O DESASTROSO LANÇAMENTO DA DISNEY

No ano passado, a Disney apostou muito de seu sucesso futuro no lançamento do Disney+, o serviço de streaming de vídeo de alto nível da empresa. Assim como muitas organizações que buscam causar um grande impacto em uma área fora de seu escopo tradicional, a Disney criou muita expectativa ao redor do lançamento e animou os clientes para a experiência que eles estariam prestes a ter.

Infelizmente, assim que o Disney+ foi lançado, os clientes começaram a reclamar sobre o baixo desempenho do serviço: carregamentos demorados, quedas e latência geral dificultaram o que poderia ter sido um dia de lançamento exultante. A crítica foi clara: um serviço que oferece baixa velocidade é tão ruim quanto um que está totalmente indisponível.

A ÉPOCA DA VELOCIDADE: BOBEOU, DANÇOU

Nos últimos anos, a maioria das empresas entendeu bem a importância do tempo de atividade. Enquanto isso, seus provedores de serviço têm feito muito para criar várias redundâncias em suas plataformas, garantindo que o caminho para a disponibilidade quase perfeita seja fácil de navegar. Ferramentas de monitoramento, práticas de engenharia de confiabilidade de site e a adoção da resiliência no caso de falhas inevitáveis ajudaram a entregar o que os usuários finais agora esperam: sites e aplicativos que estejam disponíveis sempre que necessário.

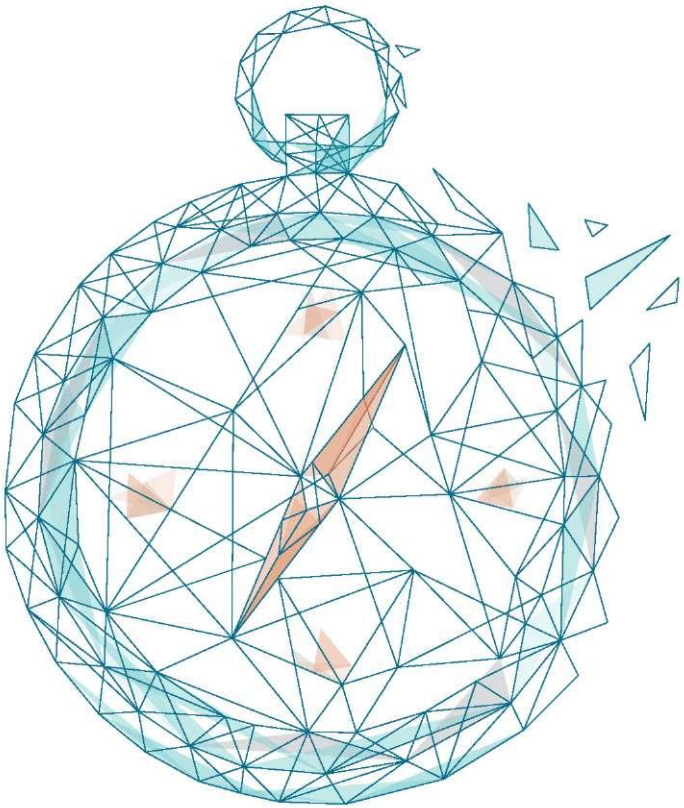
Mas toda essa engenharia adicional e o uso de arquiteturas cada vez mais complexas em um esforço para entregar os aplicativos mais resilientes apresentam novos desafios, que são tão críticos quanto o tempo de atividade.

Estamos claramente entrando em uma segunda época, com a confiabilidade se tornando a aposta da vez e a velocidade passando a ser o diferencial competitivo. As decisões dos clientes, antes feitas com tempo e análise, são cada vez mais tomadas em um piscar de olhos. E se o seu site demorar mais do que esse piscar de olhos para carregar, ou se o seu serviço de streaming sofrer com travamentos e pausas no carregamento, você tem tudo para perder.

Se você pensa que a insatisfação dos clientes não afeta seus hábitos de consumo, pense novamente. Como detalhado em um artigo da Forbes de 2019, (How Fast Is Fast Enough? Mobile Load Times Drive Customer Experience and Impact Sales (“Quão rápido é rápido o suficiente? O carregamento lento em dispositivos móveis dita a experiência do cliente e impacta as vendas”)):, “O carregamento lento em um dispositivo móvel não apenas testa a paciência do cliente, mas pode ser a falha na experiência do cliente que custará uma venda. Essa foi a principal conclusão de um relatório de velocidade de página de 2019 (...). O estudo, que explora o comportamento de 1.150 clientes e empresas, descobriu que a velocidade da página é um fator decisivo no comportamento de compra”.

E o impacto da baixa velocidade da página não é irrelevante: “Quase 70% dos clientes dizem que a velocidade da página afeta a vontade de comprar. Além disso, o carregamento lento também diminui as chances de eles voltarem no futuro. Uma análise dos dados revela que 22% dos compradores disseram que fechariam a aba, 15% disseram que visitariam o site de um concorrente e 12% contariam a um amigo sobre sua experiência negativa”.

Se a nova época é definida pela necessidade de garantir a latência mais baixa possível, no que as organizações precisam pensar para atingir esse objetivo?



Entregando velocidade em um mundo complexo

Em sua publicação seminal de 2013 intitulada The Composable Enterprise™, ("A Empresa Modular", em português), Jonathan Murray, ex-CTO do Warner Music Group, descreveu o futuro da tecnologia dentro do contexto das demandas corporativas por velocidade e agilidade. Baseado em sua experiência de vida entregando estratégias digitais para grandes corporações, Murray descreveu a "empresa modular" desta forma: "As funções, processos, organizações, relacionamentos com fornecedores e tecnologia da empresa precisam ser vistas como componentes que podem ser reconfigurados conforme necessário para lidar com o cenário competitivo em constante mudança".

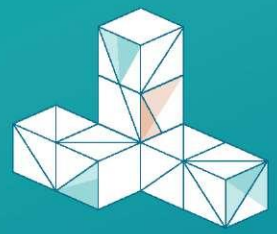
O novo modelo operacional de componentes (COM, do inglês "Component Operating Model") requer uma abordagem de "bloco de Lego" para projetar e implementar processos e as organizações que os apoiam. A implementação de uma abordagem baseada no COM terá impactos profundos na estrutura das organizações e na natureza do trabalho.

Os projetos de empresa baseados no COM criarão pressão significativa para organizações e infraestruturas de TI tradicionais. Nossos serviços de TI atuais foram desenvolvidos para servir a um modelo operacional estático — e muitas vezes funcionalmente isolado. A TI precisa se tornar muito mais dinamicamente adaptável para acompanhar a velocidade das empresas hoje.

"Um nova abordagem com o modelo de arquitetura de componentes (CAM, do inglês 'Component Architecture Model') para infraestrutura, aplicativos e serviços de TI será necessária para garantir que a TI possa entregar o que as empresas precisam. O tempo entre a identificação da necessidade de uma empresa e a entrega da solução de TI necessária precisa passar a ser horas e dias em vez de meses e anos."

Escrito há vários anos, a profética publicação de Murray descreve o novo normal dentro das organizações. Vimos, nos últimos anos, uma mudança sísmica na forma como a infraestrutura é usada e os aplicativos são construídos. Com o surgimento de contêineres, arquiteturas de microsserviços, ferramentas de aplicativos modulares reservados e semelhantes, manter um aplicativo funcionando e garantir que funcione bem requer malabarismos com dezenas de serviços, regiões, localizações, provedores de serviços e muito mais.

Portanto, embora toda essa modularidade impulsione a produtividade dos desenvolvedores e a agilidade organizacional, isso tem um custo. Parece que entregar baixa latência nessas condições é um sonho impossível.



Vimos, nos últimos anos, uma mudança sísmica na forma como a infraestrutura é usada e os aplicativos são construídos.

Dados rápidos em um ambiente de computação distribuído

Como vimos no trabalho inspirador de Murray sobre modularidade em aplicativos e infraestruturas modernas, não temos mais um stack monolítico simples sobre a qual os aplicativos são construídos. Em vez disso, em um esforço para dar aos desenvolvedores e suas organizações a maior flexibilidade e velocidade possível, usamos um grande número de serviços de desenvolvedor modulares, diferentes padrões de infraestrutura, várias abordagens de hospedagem e distribuição geográfica massiva de aplicativos. Estamos o tempo todo tentando entregar esses aplicativos o mais rápido possível para usuários em todo o mundo.

Nessa época de enorme complexidade, seria fácil pensar que não existe uma estrutura comum na qual as organizações possam confiar — o mundo delas parece perpetuamente fluido e em constante mudança.

Existe um fio em comum, entretanto, que permeia todas as diferentes coisas que uma organização faz: os dados. Ao pensar na camada de dados como um processo consistente e unificado aproveitado por todas as outras partes do stack, as organizações podem se organizar em meio ao caos. E ao escolher uma camada de dados projetada para ambientes distribuídos que apresenta tempos de processamento muito rápidos e oferece a melhor resiliência da categoria, podemos oferecer exatamente o que uma empresa precisa.

Uma das principais maneiras pelas quais as organizações podem garantir que seus aplicativos sejam resilientes e rápidos é trabalhar com uma estrutura de camada de dados consistente. E dados consistentes começam com um banco de dados que possa entregar objetivos aparentemente impossíveis: arquiteturas distribuídas, consistência, flexibilidade e velocidade.

Abordagens modernas para reduzir a latência

Como vimos, os aplicativos são cada vez mais desenvolvidos usando microsserviços: alavancando uma infinidade de diferentes partes componentes, com diferentes abordagens de infraestrutura, hospedados em vários locais diferentes, consumidos por pessoas em todos os lugares e distribuídos em muitas plataformas diferentes.

Com dados localizados em tantos lugares e transmitidos por tantas redes diferentes, não é de se surpreender que existam muitas oportunidades para a ocorrência de conflitos de dados. Para lidar com esses conflitos, os tipos de dados replicados sem conflito (CRDTs) foram desenvolvidos para permitir que os dados sejam replicados em vários locais.

Com os CRDTs, réplicas individuais podem ser atualizadas de forma independente e simultânea, sem qualquer coordenação auxiliar entre elas. Sem os CRDTs, atualizações simultâneas em várias réplicas dos mesmos dados, sem coordenação entre os computadores que hospedam as réplicas, podem resultar em inconsistências entre as réplicas.

Com os CRDTs, no entanto, quaisquer inconsistências resultantes dessa abordagem distribuída podem ser resolvidas. Inicialmente, os CRDTs eram usados em situações em que a distribuição em massa é a norma — sistemas de bate-papo online, jogos de azar na internet e streaming de áudio e vídeo —, mas cada vez mais vemos o uso em aplicativos mais genéricos.

Existe uma tecnologia significativa por trás do funcionamento do CRDT, mas a maneira mais simples de pensar nisso é que um CRDT fornece uma camada de dados na qual as réplicas podem agir de forma autônoma e ainda oferecer consistência.

Com dados localizados em tantos lugares e transmitidos por tantas redes diferentes, não é de se surpreender que existam muitas oportunidades para a ocorrência de conflitos de dados.

Superando o cache

Em modelos de banco de dados tradicionais, o local do banco de dados é separado do cache. Pense no banco de dados como a biblioteca municipal e no cache como a biblioteca local, onde os livros mais populares são mantidos para atender às demandas mais comuns dos leitores. Se os livros mais populares forem consistentes, isso pode funcionar bem, mas à medida que os hábitos de leitura mudam e novos livros entram e saem das tendências, isso se torna mais difícil.

E essa noção de verificar rapidamente diferentes peças de informação de maneira constante é apenas a metáfora para aplicativos modernos; toda a modularidade que Murray abordou resultados em dados do banco de dados tendo que ser acessados a partir de muitos serviços e locais diferentes, e em muitos períodos diferentes.

Em um mundo com serviços cada vez mais reservados e, portanto, cada vez mais lugares onde algo pode dar errado, o modelo tradicional não é o ideal (veja o diagrama abaixo). E em aplicativos cujos modelos de dados estejam centrados na transferência de vários pequenos pedaços de informações, o modelo de cache pode não ser a maneira mais rápida de mandar os dados para onde eles precisam ir.

É aqui que entra a noção de uma única camada de dados; usando uma camada única para substituir a combinação de banco de dados/cache, a complexidade da camada de dados é reduzida. Em troca, o que é turbinado é a aplicação distribuída e modular que é a regra hoje.

A vantagem adicional é que reduzir o número de componentes na camada de dados também reduz a latência. Embora componentes individuais de uma camada de dados complexa possam ser rápidos, ter um único armazenamento de dados reduz o número de saltos de rede que invariavelmente atrasam as coisas.

No lugar do cache, portanto, muitos bancos de dados modernos usam técnicas de memória em que a memória, em vez de discos externos, é usada para armazenamento. Isso é crucial, já que com tudo armazenado na memória, a velocidade não é limitada por várias camadas de armazenamento. Com um modelo baseado em cache, o que é armazenado no cache torna-se o gargalo que limita a velocidade geral.

Em um modelo tradicional, para acessar os dados, o aplicativo precisa:

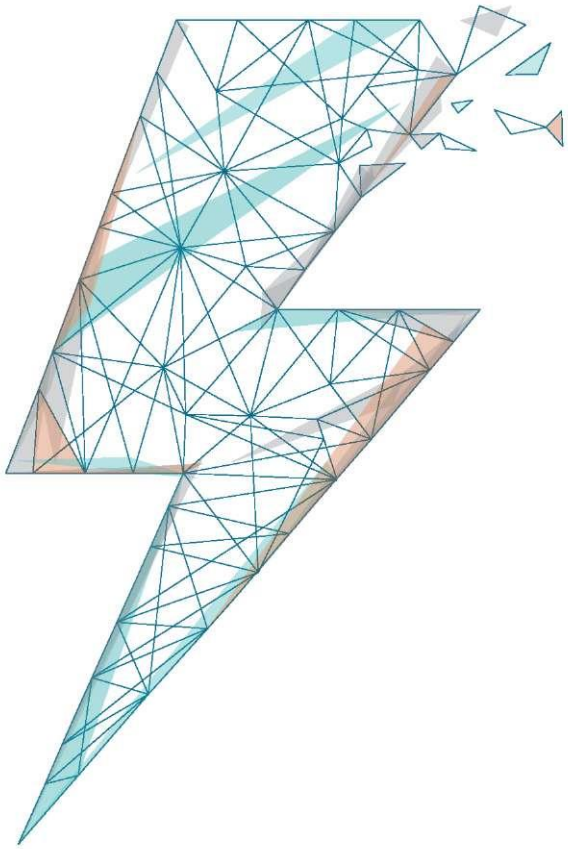
Velocidade: um byte por vez

Os bancos de dados tradicionais, como vimos acima, dependem de memória externa para seu cache. Até muito recentemente e nos últimos 50 anos, o armazenamento acontecia em discos físicos giratórios, então a maioria das abordagens tradicionais de banco de dados eram otimizadas para isso.

Mas como os discos rígidos são dispositivos físicos, eles têm restrições criadas pelo mundo físico. Para contornar essas restrições físicas, várias restrições operacionais foram criadas. Embora seja um desvio técnico, as minúcias da tecnologia de disco físico têm um grande impacto na velocidade do banco de dados.

O xis da questão, no entanto, é que a linha entre o armazenamento moderno e a memória está ficando menos nítida. O surgimento das unidades de drives sólidos (SSDs) e outras novas abordagens de armazenamento significam que essas soluções alternativas de engenharia, projetadas para um mundo limitado pela velocidade física dos dispositivos mecânicos, não são mais necessárias. Significam também que o armazenamento pode ser hierarquizado, fazendo com que todos os dados possam ser mantidos em um armazenamento rápido e não haja mais a necessidade de um cache separado.

O resultado prático, para aqueles que estão em busca de velocidade, é uma camada de dados mais rápida para construirmos nossos aplicativos.



O surgimento das novas abordagens de armazenamento significa que as soluções alternativas de engenharia, projetadas para um mundo limitado pela velocidade física dos dispositivos mecânicos, não são mais necessárias.

Escalabilidade sem prejudicar a velocidade

É ótimo criar um aplicativo que seja executado rapidamente com uso limitado, mas o que acontece quando há um enorme aumento na sua taxa de transferência? Esse é o problema que todo desenvolvedor de aplicativos em busca de aceitação e viralidade deve esperar enfrentar.

Mas o processo de escalar pode acontecer de duas formas: **para cima,** em termos de quantos dados estão sendo transferidos pela camada de dados, mas também *para fora* em termos da quantidade de informações existentes.

As organizações precisam construir uma camada de dados que permita a escalabilidade contínua em etapas. Isso envolve pensar sobre uma série de diferentes fatores: a capacidade de executar a camada de dados em vários locais, a capacidade de usar diferentes tipos de memória e armazenamento, a capacidade de agrupar dados dependendo de sua regularidade de uso e, por fim, a capacidade de escalar globalmente.

Vamos focar nessa última área. Toda essa capacidade de armazenar e processar na memória é boa, mas ao espalhar o aplicativo pelo mundo, será possível usufruir do mesmo baixo nível de latência?

É UM MUNDO MULTINÚCLEOS

O processamento moderno acontece cada vez mais com um contato multinúcleos. “Multinúcleo” se refere simplesmente à existência de duas ou mais unidades de processamento individuais em uma CPU. As instruções enviadas para a CPU podem ser processadas em núcleos separados ao mesmo tempo, aumentando a velocidade geral.

Aproveitar as arquiteturas multinúcleo pode ser um desafio. As organizações que desejem usar uma camada de dados que possa ser dimensionada da maneira mais eficiente precisam pensar sobre isso. Sua camada de dados é capaz de escalar horizontalmente em um único cluster para fornecer a melhor escala com a menor latência?

ESCALAR PARA CIMA ESCALAR PARA FORA

As organizações precisam construir uma camada de dados que permita a escalabilidade contínua em etapas.

UM PASSEIO PELO ROTA DO CAP

Já que este documento será inevitavelmente usado por aqueles que pretendem construir aplicativos distribuídos globalmente que exibam desempenho semelhante ao local, vale a pena examinar algumas restrições em torno das camadas de dados distribuídos.

Há cerca de 20 anos, o cientista da computação Eric Brewer desenvolveu o Teorema CAP, que se aplica a aplicativos distribuídos e, especificamente, aos dados que esses aplicativos criam e consomem.

O teorema CAP, nos termos mais simples, afirma que qualquer sistema de dados compartilhados em rede pode ter apenas duas das três propriedades desejáveis; consistência (C), equivalente a ter uma única cópia atualizada dos dados; alta disponibilidade (A) desses dados (para atualizações); e tolerância a partições de rede (P).

E nesses primórdios em que a busca por velocidade era tudo, o teorema CAP significava que as abordagens mais prováveis de oferecer as velocidades mais rápidas e disponibilidade de aplicativos (partições de rede e alta disponibilidade) também resultariam em inconsistência de dados.

Nas décadas desde a introdução do teorema CAP, no entanto, novas abordagens para lidar com sistemas distribuídos que permitem essa façanha teoricamente impossível foram desenvolvidas: consistência de dados, disponibilidade e tolerância de partição. O surgimento de novas abordagens de dados significa que podemos ter baixa latência sem abrir mão da consistência dos dados.

Embora este não seja o lugar para linguagem extremamente técnica, é importante para aqueles com responsabilidade pelos aplicativos de suas organizações compreender os princípios de funcionamento dos aplicativos modernos.

Como observado, em um mundo onde os aplicativos são, por necessidade, distribuídos, haverá vários nós incluídos em muitos aplicativos individuais. Nesse cenário multinós, existem duas opções gerais: **dados ativo-passivo** ou **dados ativo-ativo**.

O TEOREMA CAP

O surgimento de novas abordagens de dados significa que podemos ter baixa latência sem abrir mão da consistência dos dados.

CAMADA DE DADOS UNIFICADA

Os planos de dados, a parte do software que processa as solicitações de dados, podem ser ativo-ativo ou ativo-passivo.

Na abordagem ativo-ativo (também conhecida como dual-active), cada nó tem acesso a um banco de dados replicado, permitindo a cada nó o acesso e uso de um único aplicativo.

Essa tecnologia é o que permite manter os dados consistentes para seus aplicativos em diferentes ambientes (servidores, híbrido, multinuvem) e até mesmo em aplicativos distribuídos globalmente. Em um sistema ativo-ativo, todas as solicitações são distribuídas de acordo com a capacidade de processamento disponível. Quando ocorre uma falha em um nó, outro nó na rede toma seu lugar.

Um cluster ativo-ativo normalmente é composto de pelo menos dois nós, ambos executando ativamente o mesmo tipo de serviço simultaneamente. Como há mais nós disponíveis para uso, também haverá uma melhoria significativa no rendimento e nos tempos de resposta em comparação a uma abordagem ativo-passivo.

ATIVO-PASSIVO

Um cluster ativo-passivo também é composto por pelo menos dois nós. No entanto, como o nome “ativo-passivo” indica, nem todos os nós são ativos. Em uma Em um cluster com dois nós, por exemplo, se o primeiro nó já estiver ativo, o segundo nó deve estar passivo ou em espera. O nó passivo (também conhecido como failover) serve como reserva, pronto para assumir o controle assim que o servidor ativo (também conhecido como primário) for desconectado ou não puder atender.

Quando os clientes se conectam a um cluster de dois nós na configuração ativo-passivo, eles se conectam a apenas um servidor. Em outras palavras, todos os clientes se conectam ao mesmo servidor. Como na configuração ativo-ativo, é importante que os dois servidores tenham exatamente as mesmas configurações. Isso é chamado de redundância e garante que os dados possam ser replicados perfeitamente entre os nós.

Se forem feitas alterações nas configurações do servidor primário, essas alterações devem ser propagadas para o servidor de failover. Dessa forma, quando o failover entrar em ação, os clientes não conseguirão perceber a diferença.

Se a latência é a nova indisponibilidade, claramente quanto mais próximo um nó estiver do usuário do aplicativo, menores serão os valores de latência. Portanto, precisamos encontrar uma maneira de distribuir os aplicativos globalmente (já que distribuir nós próximos aos usuários do aplicativo reduz a latência) e ao mesmo tempo garantir a consistência. Felizmente, temos um auxílio nesse sentido.

CONSTRUÍDA COM VELOCIDADE EM MENTE

A replicação livre de conflitos é uma noção que permite que várias cópias (réplicas) de dados existam em vários locais de maneira consistente. É um método muito importante para garantir baixa latência para aplicativos distribuídos, mas há outros aspectos a serem considerados. Conforme observado acima, bancos de dados modernos projetados para entregar a latência mais baixa para aplicativos modernos armazenam dados na memória. Ao eliminar a necessidade de um cache externo, podemos reduzir a quantidade de tráfego de dados necessária.

Enquanto os bancos de dados tradicionais foram projetados para casos de uso em que o tempo de processamento de dezenas ou centenas de milissegundos era aceitável, no mundo de hoje, com a demanda por aplicativos com tempos de resposta instantâneos, o desempenho de menos de um milissegundo é uma necessidade.

TUDO BEM FALHAR, DESDE QUE SEJA RÁPIDO

O failover é um sistema automatizado que garante que, no caso de falha de um nó por qualquer motivo, outro nó replicado compense a falha. Embora seja fácil de montar um failover, a velocidade desse failover é o que determina os impactos da falha no usuário final.

Para garantir a menor latência em um mundo no qual as falhas em nós podem ser inevitáveis, é importante que a camada de dados multinós possa ativar o failover o mais rápido possível.

Sumário

No mundo moderno, as organizações, dedicadas a fornecer experiências digitais, precisam garantir que seus clientes possam usar os aplicativos quando e onde quiserem. Mas os usuários de hoje exigem não apenas acesso contínuo, mas também desempenho virtualmente instantâneo. Em um mundo em transição da época da disponibilidade para a época da velocidade, a latência pode ser tão ruim quanto a indisponibilidade do aplicativo.

Felizmente, temos opções hoje que simplesmente não estavam disponíveis há uma década. Vários obstáculos para a entrega de aplicativos rápidos — entre eles o teorema CAP — foram superados. E agora, as organizações podem aproveitar uma camada de dados livre de conflitos, independentemente de quantas réplicas sejam usadas.

Usando bancos de dados que funcionam inteiramente na memória e executando-os de maneira ativa-ativa, oferecemos bancos de dados mais rápidos do que os disponíveis anteriormente e entregamos a baixa latência exigida pelos usuários de aplicativos de hoje.

Isso deve ser prioridade para todas as organizações: seus concorrentes e disruptores estão entregando aplicativos rápidos conforme a exigência dos seus clientes; você não tem o luxo do tempo.

Sobre o autor: Ben Kepes

Ben Kepes é analista de tecnologia, comentarista e consultor e, na última década e meia, conquistou um bom público como um especialista reconhecido globalmente nas áreas de computação em nuvem, tecnologia empresarial e transformação digital.

Os comentários de Ben foram amplamente publicados em veículos como Forbes, Wired e The Guardian, fazendo com que ele fosse convidado para palestrar em diversas conferências de tecnologia, negócios e interesse geral.

@benkepes

Sobre a Redis Labs

As empresas modernas dependem do poder dos dados em tempo real. Com a Redis Labs, as organizações entregam experiências instantâneas de uma maneira altamente confiável e escalável.

A Redis Labs é o lar do Redis, o banco de dados em memória mais popular do mundo, e fornecedora comercial do Redis Enterprise, que oferece desempenho superior, confiabilidade incomparável e flexibilidade inigualável para soluções de personalização, aprendizado de máquina, IoT, pesquisa, comércio eletrônico, comunidade e medição em todo o mundo.

A Redis Labs, consistentemente classificada como líder nos principais relatórios de analistas de NoSQL, banco de dados em memória, banco de dados operacional e banco de dados como serviço (DBaaS), tem a confiança de mais de 7.400 clientes corporativos, incluindo cinco das maiores empresas do mundo (segundo a lista Fortune 10), três das quatro emissoras de cartão de crédito, três das cinco maiores empresas de comunicação, três das cinco maiores empresas de saúde, seis das oito maiores empresas de tecnologia e quatro das sete maiores varejistas.

O Redis Enterprise, disponível como serviço em nuvens públicas e privadas, como software baixável, em contêineres e para implementações híbridas (nuvem e local), possibilita casos de uso populares do Redis, como transações de alta velocidade, gerenciamento de tarefas e fila, armazenamento de sessão de usuário, ingestão de dados em tempo real, notificações, armazenamento em cache de conteúdo e dados de série temporal.

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

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

Siga-nos