

Módulo 7 - Análise detalhada do pilar de eficiência de desempenho

1.1 Boas-vindas!

Boas-vindas ao módulo sete do AWS Well-Architected: Análise detalhada do pilar de eficiência de desempenho.

1.2 Objetivos de aprendizado

Neste módulo, você terá uma visão geral do pilar de eficiência de desempenho do AWS Well-Architected Framework. Você também aprenderá os princípios de design e as práticas recomendadas do pilar de eficiência de desempenho.

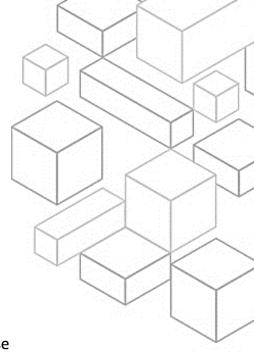
1.3 Visão geral do pilar de eficiência de desempenho

Para começar, você terá uma visão geral do pilar de eficiência de desempenho.

1.4 Pilares do Well-Architected

Atualmente, há seis pilares do Well-Architected Framework:excelência operacional, segurança, confiabilidade, eficiência de desempenho, otimização de custos e sustentabilidade. Esses pilares são os fundamentos da arquitetura de suas soluções de tecnologia na nuvem. Este módulo se concentrará no pilar de eficiência de desempenho.

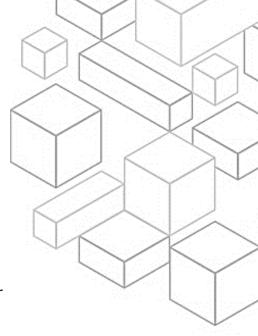
© 2023, Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.





1.5 O que é o pilar de eficiência de desempenho?

O que é o pilar de eficiência de desempenho? O pilar de eficiência de desempenho concentra-se no uso eficiente dos recursos de computação para atender aos requisitos e manter a eficiência à medida que a demanda muda e as tecnologias evoluem.



1.6 Princípios de design de eficiência de desempenho

Agora que você já sabe o que é o pilar de eficiência de desempenho, vai se aprofundar nos princípios de design do pilar de eficiência de desempenho.

1.7 Princípios de design de eficiência de desempenho

Há cinco princípios de design para a eficiência do desempenho na nuvem. O primeiro princípio de design é democratizar as tecnologias avançadas. Simplifique a implementação de tecnologia avançada para a sua equipe e delegue tarefas complexas ao seu fornecedor de nuvem. Não peça à sua equipe de TI para hospedar uma nova tecnologia. Em vez disso, considere consumi-la como um serviço. Na nuvem, as tecnologias que, de outra forma, exigiriam conhecimento especializado, como Machine Learning e bancos de dados NoSQL, tornam-se serviços que sua equipe pode usar. Isso permite que a equipe tenha mais tempo para se concentrar no desenvolvimento de produtos, em vez de gastar tempo com o provisionamento e o gerenciamento de recursos. O segundo princípio é tornar-se global em minutos. A implantação em várias Regiões ajuda a aproximar sua carga de trabalho do seu público global. Isso pode resultar em menor latência e em uma experiência melhor. Por exemplo, aproveitando

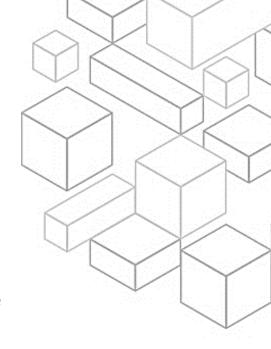


serviços como o AWS CloudFormation, você pode ativar rapidamente recursos em diferentes regiões geográficas globais com o mínimo de sobrecarga. Isso pode oferecer melhor desempenho aos usuários de aplicações e fornecer acesso a outros recursos ou funcionalidades que podem existir em diferentes Regiões. Outro princípio é usar arquiteturas sem servidor. As arquiteturas sem servidor eliminam a necessidade de executar e manter servidores físicos para atividades tradicionais de computação. Por exemplo, os serviços de armazenamento sem servidor podem atuar como sites estáticos. Isso elimina a necessidade de servidores da web, e os serviços de eventos podem hospedar códigos. E também reduz a carga operacional do gerenciamento de servidores físicos. Você pode se beneficiar de custos transacionais mais baixos porque os serviços gerenciados operam em escala de nuvem. O quarto princípio é fazer experimentos com mais frequência. Na nuvem, com recursos praticamente ilimitados, você pode comparar rapidamente as configurações de suas cargas de trabalho. Isso pode ser tão simples quanto experimentar uma instância de tamanho diferente ou um tipo de armazenamento. Ou quanto experimentar serviços totalmente diferentes, como a execução de código em uma função AWS Lambda em vez de em uma instância do Amazon Elastic Compute Cloud, ou Amazon EC2. O último princípio de design é considerar a afinidade mecânica. Alinhe sua abordagem de tecnologia a seus objetivos comerciais gerais, e não o contrário. É importante dedicar algum tempo para entender como os serviços em nuvem são consumidos e usar a abordagem tecnológica que melhor se alinhe às suas metas de carga de trabalho. Por exemplo, sempre considere os padrões de acesso aos dados ao selecionar abordagens de banco de dados ou armazenamento.



1.8 Práticas recomendadas de eficiência de desempenho

Agora que você entende os princípios de eficiência de desempenho, aprenderá sobre as práticas recomendadas de eficiência de desempenho.



1.9 Áreas de práticas recomendadas de eficiência de desempenho

O pilar de eficiência de desempenho está agrupado em quatro áreas de práticas recomendadas. Isso inclui seleção, análise, monitoramento e concessões.

1.10 Seleção

A seleção é a primeira área de práticas recomendadas de eficiência de desempenho.

1.11 Seleção da arquitetura de desempenho

Para a seleção da arquitetura de desempenho, use uma abordagem orientada por dados para selecionar os padrões e a implementação de sua arquitetura e obter uma solução econômica. Sua arquitetura provavelmente combinará várias abordagens de arquitetura. A implementação de sua arquitetura usará serviços específicos para a otimização do desempenho de sua arquitetura. Você deve conhecer os serviços e recursos disponíveis. Saiba mais sobre a ampla gama de serviços e recursos disponíveis na nuvem. Você também deve identificar os serviços e as opções de configuração relevantes para sua carga de trabalho e entender como obter o desempenho ideal. Você deve definir um processo para



escolhas arquitetônicas. Para isso, use a experiência e o conhecimento internos sobre a nuvem. Você também pode usar recursos externos, como casos de uso publicados, documentação ou whitepapers, para definir um processo que ajude a escolher recursos e serviços. Você deve definir um processo que incentive a experimentação e o benchmarking com os serviços que podem ser usados em sua carga de trabalho. Além disso, considere os requisitos de custo em suas decisões. As cargas de trabalho geralmente têm requisitos de custo para a operação. Use controles de custos internos para selecionar tipos e tamanhos de recursos com base na necessidade prevista de recursos. Determine quais componentes da carga de trabalho poderiam ser substituídos por serviços totalmente gerenciados, como bancos de dados gerenciados, caches em memória e outros serviços. Ao reduzir sua carga de trabalho operacional, você pode concentrar os recursos nos resultados comerciais. Você deve usar políticas ou arquiteturas de referência. Maximize o desempenho e a eficiência avaliando as políticas internas e as arquiteturas de referência existentes. Em seguida, use sua análise para selecionar serviços e configurações para sua carga de trabalho. Use a orientação de seu provedor de nuvem ou de um parceiro apropriado. Pesquise os recursos da empresa de nuvem, como arquitetos de soluções, serviços profissionais ou um parceiro adequado, para guiar suas decisões. Esses recursos podem ajudar a analisar e aprimorar sua arquitetura para obter um desempenho ideal. Você deve comparar o desempenho de uma carga de trabalho existente para entender como ela se comporta na nuvem. Use os dados coletados dos benchmarks para orientar as decisões de arquitetura. Use o benchmarking com testes sintéticos para gerar dados sobre o desempenho dos componentes de sua carga de trabalho. O



benchmarking geralmente é mais rápido de configurar do que o teste de carga e é usado para avaliar a tecnologia de um componente específico. Por fim, teste a carga de trabalho. Implante sua arquitetura de carga de trabalho mais recente na nuvem usando diferentes tipos e tamanhos de recursos. Monitore a implantação para capturar métricas de desempenho que identifiquem gargalos ou excesso de capacidade. Use essas informações de desempenho para projetar ou aprimorar sua arquitetura e seleção de recursos.

1.12 Seleção da arquitetura de computação

Ao selecionar uma arquitetura de computação, escolha recursos de computação que atendam a seus requisitos e necessidades de desempenho e, ao mesmo tempo, ofereçam grande eficiência de custo e esforço. Isso pode ajuda você a realizar mais com o mesmo número de recursos. Porém, a solução de computação ideal para uma carga de trabalho sempre variará com base no design da aplicação, nos padrões de uso e nas definições de configuração. Você deve avaliar as opções de computação disponíveis. Entenda as características de desempenho das opções relacionadas à computação disponíveis para você. Saiba como funcionam as instâncias, os contêineres e as funções e quais são as vantagens e desvantagens que eles trazem para a sua carga de trabalho. Na AWS, a computação está disponível em três formas: instâncias, contêineres e funções. Além disso, entenda as opções de configuração de computação disponíveis. Como as várias opções complementam sua carga de trabalho e quais opções de configuração são melhores para seu sistema. Os exemplos incluem a família de instâncias, tamanhos, recursos como GPU ou E/S, tamanhos de

^{© 2023,} Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.



funções, instâncias de contêineres e único em comparação com multi-tenancy. Você também deve coletar métricas relacionadas à computação. Para entender o desempenho dos seus recursos de computação, é preciso registrar e acompanhar a utilização de vários sistemas. Você pode usar esses dados para determinar com mais precisão os requisitos de recursos. Determinar a configuração necessária por meio do dimensionamento correto. Analise as características de desempenho de sua carga de trabalho e como essas características estão relacionadas à memória, à rede, à E/S e ao uso da CPU. Use esses dados para escolher os recursos que melhor atendem ao perfil de sua carga de trabalho. Por exemplo, uma carga de trabalho com uso intensivo de memória, como um banco de dados, pode se beneficiar de uma proporção maior de memória por núcleo. No entanto, uma carga de trabalho com uso intenso de computação pode precisar de uma contagem e frequência de núcleos mais altas, mas pode ser satisfeita com uma quantidade menor de memória por núcleo. Você também pode usar a elasticidade disponível dos recursos. A nuvem oferece a flexibilidade de expandir e reduzir seus recursos dinamicamente por meio de uma variedade de mecanismos para atender às mudanças na demanda. Combinando essa elasticidade com métricas relacionadas à computação, uma carga de trabalho pode responder automaticamente às alterações para usar os recursos de que precisa e somente os recursos de que precisa. Avalie continuamente as necessidades de computação com base em métricas. Use uma abordagem orientada por dados para avaliar e otimizar os recursos de computação para sua carga de trabalho ao longo do tempo.



1.13 Seleção da arquitetura de armazenamento

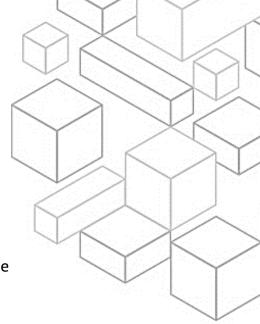
Para a seleção da arquitetura de armazenamento, a solução

de armazenamento ideal para um sistema varia com base no tipo de método de acesso, que pode ser bloco, arquivo ou objeto, e com base no fato de os padrões de acesso serem aleatórios ou sequenciais. A solução de armazenamento ideal varia de acordo com o throughput necessário e as restrições de disponibilidade e durabilidade. Também pode ser diferente com base na frequência de acesso ou na frequência de atualização. Ao selecionar a solução de armazenamento para sua arquitetura, você pode seguir algumas práticas recomendadas. Primeiro, entenda as características e os requisitos do armazenamento. Identifique e documente as necessidades de armazenamento da carga de trabalho e defina as características de armazenamento de cada local. Exemplos de características de armazenamento incluem acesso compartilhável, tamanho do arquivo, taxa de crescimento, throughput, operações de entrada e saída por segundo ou IOPS, latência, padrões de acesso e persistência de dados. Use essas características para avaliar se os serviços de armazenamento em bloco, arquivo, objeto ou instância são a solução mais eficiente para suas necessidades de armazenamento. Em seguida, avalie as opções de configuração disponíveis. Avalie as diversas características e opções de configuração e como se relacionam ao armazenamento. Entenda onde e como usar IOPS provisionado, unidade de estado sólido ou SSDs, armazenamento magnético, armazenamento de objetos, armazenamento de arquivos ou armazenamento temporário para otimizar o espaço de armazenamento e o desempenho de sua carga de trabalho. A última prática recomendada é tomar decisões com base em padrões e métricas de acesso. Escolha sistemas de armazenamento com base nos padrões de acesso de

^{© 2023,} Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.



sua carga de trabalho e configure-os determinando como a carga de trabalho acessa os dados. Aumente a eficiência do armazenamento escolhendo o armazenamento em objetos em vez do armazenamento em blocos. Configure as opções de armazenamento escolhidas de acordo com seus padrões de acesso aos dados.



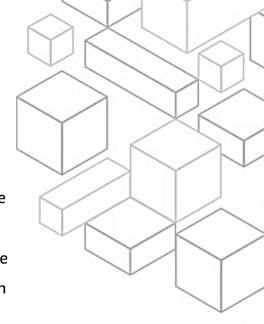
1.14 Seleção da arquitetura do banco de dados

Com relação à seleção da arquitetura do banco de dados, escolher a solução e os recursos errados para um sistema pode reduzir a eficiência do desempenho. Portanto, ao decidir por uma solução de banco de dados ideal, é importante considerar os requisitos. Os exemplos incluem disponibilidade, consistência, tolerância a partições, latência, durabilidade, dimensionamento e capacidade de consulta. Você pode seguir as práticas recomendadas ao selecionar a solução de banco de dados ideal para sua carga de trabalho. Primeiro, entenda as características dos dados. Escolha suas soluções de gerenciamento de dados para corresponder de forma ideal às características, aos padrões de acesso e aos requisitos de seus conjuntos de dados de carga de trabalho. Ao selecionar e implementar uma solução de gerenciamento de dados, certifique-se de que as características de consulta, scaling e armazenamento sejam compatíveis com os requisitos de dados da carga de trabalho. Saiba como as opções de banco de dados correspondem aos seus modelos de dados e quais opções de configuração são melhores para o seu caso de uso. Depois, avalie as opções disponíveis. Entenda as opções de banco de dados disponíveis e como elas podem otimizar seu desempenho antes de selecionar sua solução de gerenciamento de dados. Use o teste de carga para identificar as métricas do banco de dados que são

^{© 2023,} Amazon Web Services, Inc. ou suas afiliadas. Todos os direitos reservados.



importantes para sua carga de trabalho. Enquanto você explora as opções de banco de dados, considere as opções de grupos de parâmetros, armazenamento, memória, computação, réplica de leitura, consistência eventual, pool de conexões e armazenamento em cache. Faça experiências com essas opções de configuração para aprimorar as métricas.



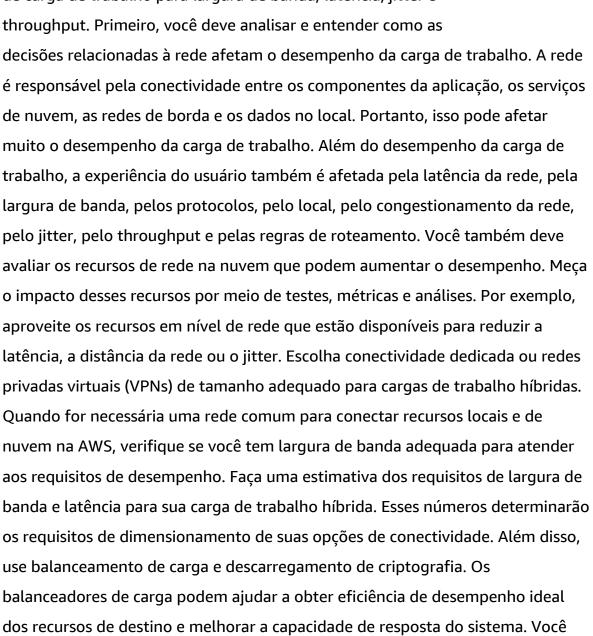
Além disso, colete e registre as métricas de desempenho do banco de dados para entender o desempenho dos sistemas de gerenciamento de dados. Essas métricas o ajudarão a otimizar os recursos de gerenciamento de dados para garantir que os requisitos da carga de trabalho sejam atendidos e que você tenha uma visão geral clara do desempenho da carga de trabalho. Use ferramentas, bibliotecas e sistemas que registram medições de desempenho para o desempenho do banco de dados. Escolha o armazenamento de dados com base em padrões de acesso. Use os padrões de acesso da carga de trabalho e os requisitos das aplicações para decidir sobre as tecnologias e os serviços de dados ideais a serem usados. Por fim, otimize o armazenamento de dados com base em padrões e métricas de acesso. Use características de desempenho e padrões de acesso que otimizem a forma como os dados são armazenados ou consultados para obter o melhor desempenho. Avalie como as otimizações, como indexação, distribuição de chaves, projeto de data warehouse ou estratégias de cache, afetam o desempenho do sistema ou a eficiência geral.

1.15 Seleção de arquitetura de rede

A seleção da arquitetura de rede pode ter grandes impactos (positivos e negativos) no desempenho e no comportamento da carga de trabalho, porque a rede está entre todos os componentes da carga de trabalho. Há também cargas



de trabalho que dependem muito do desempenho da rede, como a computação de alto desempenho, ou HPC, em que o conhecimento profundo da rede é importante para aumentar o desempenho do cluster. Você deve determinar os requisitos de carga de trabalho para largura de banda, latência, jitter e throughput. Primeiro, você deve analisar e entender como as





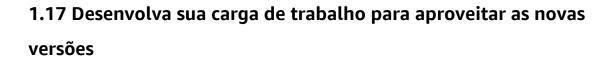
deve avaliar os requisitos de desempenho da sua carga de trabalho e escolher os protocolos de rede que otimizam o desempenho geral da carga de trabalho. Há uma relação entre a latência e a largura de banda para atingir o throughput. Por exemplo, se a sua transferência de arquivos estiver usando TCP, latências mais altas reduzirão o throughput geral. Há abordagens para corrigir isso com o ajuste do TCP e protocolos de transferência otimizados. Algumas abordagens usam o Protocolo de datagrama de usuário, ou UDP. O protocolo Scalable Reliable Datagram (SRD), ou Datagrama confiável e dimensionável, é um protocolo de transporte de rede criado pela AWS para Elastic Fabric Adapters (EFAs) que fornece entrega confiável de datagramas. Ao contrário do protocolo TCP, o SRD pode reordenar os pacotes e entregá-los fora de ordem. Esse mecanismo de entrega fora de ordem do SRD envia pacotes em paralelo por caminhos alternativos, aumentando o throughput. Você também deve escolher o local da sua carga de trabalho com base nos requisitos da rede. Avalie as opções de posicionamento de recursos para reduzir a latência da rede e melhorar o throughput, proporcionando uma experiência do usuário ideal ao reduzir o tempo de carregamento da página e de transferência de dados. Por fim, você deve otimizar a configuração da rede com base nas métricas. A configuração inadeguada da rede pode afetar o desempenho, a eficiência e o custo da rede. Em ambientes de rede comuns, para concluir rapidamente a implantação em estágio inicial, a configuração adequada da rede não é totalmente considerada em termos de desempenho da rede. Para otimizar a configuração da rede, primeiro é preciso ter visibilidade e dados sobre o ambiente de rede. Para entender o desempenho dos seus recursos de rede, colete e analise dados para tomar decisões informadas sobre a otimização da



configuração da rede. Meça o impacto dessas mudanças e use as medições de impacto para tomar decisões futuras.

1.16 Análise

A análise é a segunda área de práticas recomendadas de eficiência de desempenho.

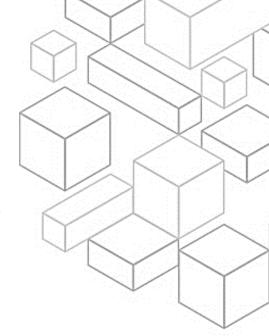


Desenvolva sua carga de trabalho para aproveitar as vantagens das novas versões. Quando você arquiteta cargas de trabalho, há um número limitado de opções que você pode escolher. No entanto, ao longo do tempo, novas tecnologias e abordagens que se tornam disponíveis poderiam melhorar o desempenho de sua carga de trabalho. Existem práticas recomendadas que você pode seguir para revisar e desenvolver continuamente suas cargas de trabalho para aproveitar as vantagens das novas versões. Primeiro, avalie as maneiras de melhorar o desempenho à medida que novos serviços, padrões de design e ofertas de produtos forem disponibilizados. Determine quais deles poderiam melhorar o desempenho ou aumentar a eficiência da carga de trabalho por meio de avaliação, discussão interna ou análise externa. Além disso, defina um processo para melhorar o desempenho da carga de trabalho, avaliando novos serviços, padrões de projeto, tipos de recursos e configurações à medida que forem disponibilizados. Por exemplo, execute testes de desempenho existentes em novas ofertas de instância para determinar seu potencial de melhorar sua carga de trabalho. Por fim, desenvolva o desempenho da carga de trabalho ao





longo do tempo. Como organização, use as informações coletadas por meio do processo de avaliação para promover ativamente a adoção de novos serviços ou recursos quando eles estiverem disponíveis.



1.18 Monitoramento

O monitoramento é a terceira área de práticas recomendadas de eficiência de desempenho.

1.19 Monitorar os recursos para garantir o desempenho esperado

Monitore os recursos para garantir o desempenho esperado após a implementação da carga de trabalho para corrigir qualquer problema ou desvio dos níveis de desempenho esperados. Para isso, primeiro use um serviço de monitoramento e observabilidade para registrar métricas relacionadas ao desempenho. Por exemplo, registre transações de banco de dados, consultas lentas, latência de E/S, throughput de solicitações HTTP, latência de serviço ou outros dados importantes. Você também deve analisar as métricas quando ocorrerem eventos ou incidentes, usando painéis ou relatórios de monitoramento para entender e diagnosticar o impacto. Essas exibições fornecem informações sobre quais partes da carga de trabalho não estão funcionando conforme o esperado. Estabeleça os principais indicadores de desempenho, ou KPIs, para medir o desempenho da carga de trabalho. Por exemplo, uma carga de trabalho baseada em API pode usar a latência geral da resposta como uma indicação do desempenho geral. Um site de comércio eletrônico pode optar por usar o número de compras como seu KPI. Em seguida,



com os KPIs relacionados ao desempenho que você definiu, use um sistema de monitoramento que gere alarmes automaticamente quando essas medições estiverem fora dos limites esperados. Você também deve analisar as métricas em intervalos regulares. Como manutenção de rotina ou em resposta a eventos ou incidentes, analise quais métricas são coletadas. Use essas revisões para identificar quais métricas foram fundamentais para resolver os problemas e quais métricas adicionais, se estivessem sendo monitoradas, ajudariam a identificar, resolver ou evitar problemas. Por fim, monitore e emita alarmes de forma proativa. Use KPIs, combinados com sistemas de monitoramento e alerta, para tratar proativamente de problemas relacionados ao desempenho. Use alarmes para automatizar ações e corrigir problemas sempre que possível. Escale o alarme para aqueles capazes de responder se a resposta automática não for possível. Por exemplo, você pode ter um sistema capaz de prever os valores esperados de KPI e enviar um alarme quando eles violarem determinados limites. Você também pode ter uma ferramenta para interromper ou reverter automaticamente as implantações se os KPIs estiverem fora dos valores esperados.

1.20 Concessões

As concessões são a última área de práticas recomendadas de eficiência de desempenho.

1.21 Usar concessões para melhorar o desempenho

O uso de concessões para melhorar o desempenho ao arquitetar soluções



permite que você selecione uma abordagem ideal. Muitas vezes, é possível melhorar o desempenho trocando consistência, durabilidade e espaço por tempo e latência. Essas práticas recomendadas o ajudarão a responder à pergunta sobre como usar as concessões para melhorar o desempenho. Primeiro, entenda e identifique as áreas em que o aumento do desempenho de sua carga de trabalho terá um impacto positivo na eficiência ou na experiência do cliente. Por exemplo, um site que tenha uma grande quantidade de interação com o cliente pode se beneficiar do uso de serviços de borda para aproximar a entrega de conteúdo dos clientes. Outra prática recomendada é pesquisar e entender os vários padrões de design e serviços que ajudam a melhorar o desempenho da carga de trabalho. Como parte da análise, identifique o que você poderia negociar para obter maior desempenho. Por exemplo, o uso de um serviço de cache pode ajudar a reduzir a carga colocada nos sistemas de banco de dados. No entanto, isso pode exigir alguma engenharia para implementar o cache seguro ou a possível introdução de consistência eventual em algumas áreas. Além disso, ao avaliar as melhorias relacionadas ao desempenho, determine quais opções afetarão seus clientes e a eficiência da carga de trabalho. Por exemplo, se o uso de um armazenamento de dados de valor-chave aumentar o desempenho do sistema, é importante avaliar como a natureza consistente desse armazenamento afetará os clientes. Meça o impacto das melhorias de desempenho. À medida que forem feitas alterações para melhorar o desempenho, avalie as métricas e os dados coletados. Use essas informações para determinar o impacto que a melhoria de desempenho teve sobre a carga de trabalho, os componentes da carga de trabalho e seus clientes. Essa medição ajuda a entender as melhorias resultantes da concessão. Ela



também pode ajudar a determinar se houve algum efeito colateral negativo. A última prática recomendada é usar várias estratégias relacionadas ao desempenho, quando aplicável.

Por exemplo, você pode usar estratégias como o armazenamento de dados em cache para evitar chamadas excessivas à rede ou ao banco de dados. É possível usar réplicas de leitura para mecanismos de banco de dados para melhorar as taxas de leitura. Você pode experimentar estratégias como sharding ou compactação de dados sempre que possível para reduzir os volumes de dados. Você também pode usar os resultados de armazenamento em buffer e streaming à medida que estiverem disponíveis para evitar bloqueios.

1.25 Resumo

Neste módulo, você aprendeu sobre o pilar de eficiência de desempenho. Iniciamos com uma visão geral e incluímos uma discussão aprofundada sobre a proposta de valor, os princípios de design e as práticas recomendadas do pilar de eficiência de desempenho.

1.26 Agradecemos sua atenção

Agradecemos sua participação!