Framework Well-Architected da AWS

Pilar Eficiência de performance

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Pilar Eficiência de performance: Framework Well-Architected da AWS

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

As marcas comerciais e imagens comerciais da Amazon não podem ser usadas no contexto de nenhum produto ou serviço que não seja da Amazon, nem de qualquer maneira que possa gerar confusão entre os clientes ou que deprecie ou desprestigie a Amazon. Todas as outras marcas comerciais que não são propriedade da Amazon pertencem aos respectivos proprietários, os quais podem ou não ser afiliados, estar conectados ou ser patrocinados pela Amazon.

Table of Contents

Resumo e introdução ... 1
Introdução ... 1
Eficiência de performance ... 3
Princípios de design ... 3
Definição ... 4
Seleção de arquitetura ... 5
PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis ... 5
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas ... 8
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP03 Inclua o custo nas decisões de arquitetura ... 10
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP04 Avaliar como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura ... 12
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP05 Usar políticas e arquiteturas de referência ... 14
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP06 Usar testes comparativos para orientar decisões de arquitetura ... 16
Orientação para implementação ... 6
Recursos ... 7
PERF01-BP07 Usar uma abordagem baseada em dados para escolhas de arquitetura ... 18
Orientação para implementação ... 6
Recursos ... 7
Computação e hardware ... 22
PERF02-BP01 Selecionar as melhores opções de computação para as workloads ... 22
Orientação para implementação ... 6
Etapas de implementação ... 6
Recursos ... 7

PERF02-BP02 Entender a configuração e os recursos de computação disponíveis 26
Orientação para implementação 6
Etapas de implementação 6
Recursos 7
PERF02-BP03 Coletar métricas relacionadas à computação 30
Orientação para implementação 6
Etapas de implementação 6
Recursos 7
PERF02-BP04 Configurar e dimensionar corretamente os recursos de computação 33
Orientação para implementação 6
Recursos 7
PERF02-BP05 Dimensionar recursos de computação dinamicamente 35
Orientação para implementação 6
Recursos 7
PERF02-BP06 Usar aceleradores de computação baseados em hardware otimizados 39
Orientação para implementação 6
Recursos 7
Gerenciamento de dados 42
PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados 42
Orientação para implementação 6
Recursos 7
PERF03-BP02 Avaliar as opções de configuração disponíveis para o datastore 54
Orientação para implementação 6
Recursos 7
PERF03-BP03 Coletar e registrar métricas de performance do datastore 60
Orientação para implementação 6
Etapas de implementação 6
Recursos 7
PERF03-BP04 Implementar estratégias para melhorar a performance da consulta no datastore 63
Orientação para implementação 6
Recursos 7
PERF03-BP05 Implementar padrões de acesso a dados que utilizam cache 65
Orientação para implementação 6
Recursos 7

Rede e entrega de conteúdo 70
PERF04-BP01 Compreender como as redes afetam a performance 70
Orientação para implementação 6
Recursos 7
PERF04-BP02 Avaliar os recursos de rede disponíveis 74
Orientação para implementação 6
Recursos 7
PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload 80
Orientação para implementação 6
Recursos 7
PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos 83
Orientação para implementação 6
Recursos 7
PERF04-BP05 Escolher protocolos de rede para melhorar a performance 87
Orientação para implementação 6
Recursos 7
PERF04-BP06 Escolher o local da workload com base nos requisitos de rede 91
Orientação para implementação 6
Recursos 7
PERF04-BP07 Otimizar a configuração da rede com base em métricas 96
Orientação para implementação 6
Recursos 7
Processo e cultura 101
PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload 103
Orientação para implementação 6
Etapas de implementação 6
Recursos 7
PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica 106
Orientação para implementação 6
Recursos 7
PERF05-BP03 Definir um processo para melhorar a performance da workload 109
Orientação para implementação 6
Recursos 7
PERF05-BP04 Fazer o teste de carga da workload 111

Orientação para implementação 6
Recursos 7
PERF05-BP05 Usar automação para corrigir proativamente problemas relacionados a performance 113
Orientação para implementação 6
Recursos 7
PERF05-BP06 Manter a workload e os serviços atualizados 115
Orientação para implementação 6
Etapas de implementação 6
Recursos 7
PERF05-BP07 Revisar as métricas regularmente 117
Orientação para implementação 6
Recursos 7
Conclusão 120
Colaboradores 121
Outras fontes de leitura 122
Revisões do documento 123
Avisos 125
Glossário da AWS 126

Pilar Eficiência de performance: AWS Well-Architected Framework

Data de publicação: 6 de novembro de 2024 (Revisões do documento)

Este whitepaper destaca o pilar Eficiência de performance do AWS Well-Architected Framework. Ele fornece orientações para ajudar os clientes a aplicar as práticas recomendadas nas áreas de projeto, entrega e manutenção de ambientes da AWS.

Introdução

O AWS Well-Architected Framework ajuda a compreender os prós e os contras das decisões tomadas ao criar workloads na AWS. O uso do Framework ajuda você a aprender as práticas de arquitetura recomendadas para projetar e operar workloads confiáveis, seguras, eficientes, econômicas e sustentáveis na nuvem. Ele fornece uma maneira de você avaliar consistentemente suas arquiteturas em relação às práticas recomendadas e identificar áreas de aprimoramento. Acreditamos que ter as workloads bem arquitetadas aumenta muito a probabilidade de sucesso nos negócios.

O framework é baseado em seis pilares:

- Excelência operacional
- Segurança
- Confiabilidade
- Eficiência de performance
- Otimização de custo
- Sustentabilidade

Este documento foca a aplicação dos princípios do pilar de eficiência de performance às suas workloads. Em ambientes tradicionais on-premises, alcançar performance elevada e duradoura é algo desafiador. O uso dos princípios apresentados neste documento ajudará você a criar arquiteturas na AWS que entreguem, com eficácia, performance constante ao longo do tempo. A orientação e as práticas recomendadas deste documento estão distribuídas em cinco áreas de foco principais que servem como princípios orientadores para a criação de soluções de nuvem eficientes na AWS. Essas áreas de foco são:

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Este documento destina-se a pessoas que ocupam cargos de tecnologia, como diretores de tecnologia (CTOs), arquitetos, desenvolvedores e membros da equipe de operações. Após ler este documento, você entenderá as práticas recomendadas e as estratégias da AWS a serem usadas ao projetar arquiteturas de nuvem de alta performance.

Eficiência de performance

O pilar Eficiência de performance inclui a capacidade de usar recursos de nuvem de maneira eficiente para atender aos requisitos de performance e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem.

Tópicos

- Princípios de design
- Definição

Princípios de design

Os princípios de design a seguir podem ajudar você a alcançar e manter workloads eficientes na nuvem.

- Democratize tecnologias avançadas: facilite a implementação de tecnologia avançada para a sua equipe delegando tarefas complexas ao seu fornecedor de nuvem. Em vez de solicitar que sua equipe de TI aprenda a hospedar e executar uma nova tecnologia, avalie a possibilidade de consumir a tecnologia como um serviço. Por exemplo, bancos de dados NoSQL, transcodificação de mídia e machine learning são tecnologias que exigem altos níveis de especialização. Na nuvem, essas tecnologias se tornam serviços que sua equipe pode consumir, permitindo que a equipe se concentre no desenvolvimento de produtos, em vez de provisionamento e gerenciamento de recursos.
- Tenha alcance global em poucos minutos: a implantação da sua workload em várias regiões da AWS em todo o mundo permite oferecer menor latência e uma melhor experiência para seus clientes a um custo mínimo.
- Use arquiteturas sem servidor: as arquiteturas sem servidor eliminam a necessidade de executar e manter servidores físicos para realizar atividades tradicionais de computação. Os serviços de armazenamento sem servidor, por exemplo, podem atuar como sites estáticos (eliminando a necessidade de servidores Web) e os serviços de eventos podem hospedar código. Isso elimina o fardo operacional do gerenciamento de servidores físicos e pode reduzir os custos transacionais, pois os serviços gerenciados operam em escala de nuvem.
- Experimente com mais frequência: com recursos virtuais e automatizáveis, você pode executar rapidamente testes comparativos usando diferentes tipos de instâncias, armazenamento ou configurações.

- Considere a solidariedade mecânica: use a abordagem tecnológica mais bem alinhada aos seus objetivos. Por exemplo, avalie padrões de acesso a dados ao selecionar banco de dados ou armazenamento para a workload.

Definição

Concentre-se nas seguintes áreas para alcançar eficiência de performance na nuvem:

- Seleção de arquitetura
- Computação e hardware
- Gerenciamento de dados
- Rede e entrega de conteúdo
- Processo e cultura

Adote uma abordagem baseada em dados para criar uma arquitetura de alta performance. Reúna dados sobre todos os aspectos da arquitetura, desde o design de alto nível até a seleção e a configuração dos tipos de recursos.

Analisar suas escolhas regularmente garante que você esteja tirando proveito da evolução contínua da Nuvem AWS. O monitoramento garante que você esteja ciente de qualquer desvio em relação à performance esperada. Faça concessões em sua arquitetura visando o aprimoramento da performance, como o uso de compactação ou armazenamento em cache, ou ainda a diminuição dos requisitos de consistência.

Seleção de arquitetura

A solução ideal para uma workload específica pode variar e, muitas vezes, as soluções combinam várias abordagens. As workloads do Well-Architected usam várias soluções e permitem diferentes recursos para aprimorar a performance.

Os recursos da AWS estão disponíveis em vários tipos e configurações, o que facilita encontrar uma abordagem que atenda melhor às suas necessidades. Também é possível encontrar opções que não são facilmente obtidas com infraestrutura on-premises. Um serviço gerenciado como o Amazon DynamoDB, por exemplo, fornece um banco de dados NoSQL totalmente gerenciado com latência de milissegundos de um dígito em qualquer escala.

Essa área de foco compartilha orientações e práticas recomendadas sobre como selecionar padrões de arquitetura e recursos de nuvem eficientes e de alta performance.

Práticas recomendadas

- PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis
- PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas
- PERF01-BP03 Inclua o custo nas decisões de arquitetura
- PERF01-BP04 Avaliar como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura
- PERF01-BP05 Usar políticas e arquiteturas de referência
- PERF01-BP06 Usar testes comparativos para orientar decisões de arquitetura
- PERF01-BP07 Usar uma abordagem baseada em dados para escolhas de arquitetura

PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis

Continue a descobrir e aprender sobre serviços e configurações disponíveis que ajudam a tomar decisões e melhorar a eficiência de performance na arquitetura da workload.

Práticas comuns que devem ser evitadas:

- Usar a nuvem como um datacenter colocalizado.

- Não modernizar a aplicação após a migração para a nuvem.
- Usar somente um tipo de armazenamento para tudo que precisa ser mantido.
- Usar tipos de instância que atendem melhor aos seus padrões atuais, mas que sejam maiores quando necessário.
- Você implanta e gerencia tecnologias disponíveis como serviços gerenciados.

Benefícios de implementar esta prática recomendada: ao pensar em novos serviços e configurações, você poderá melhorar consideravelmente a performance, reduzir custos e otimizar o esforço necessário para manter as workloads. Isso também pode ajudar a acelerar o tempo para valorização dos produtos habilitados para a nuvem.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS lança constantemente novos serviços e recursos que podem melhorar a performance e reduzir o custo das workloads na nuvem. Atualizar-se em relação a esses novos serviços e atributos é crucial para manter a eficácia da performance na nuvem. Modernizar a arquitetura da workload também ajuda a acelerar a produtividade, impulsionar a inovação e ter acesso a mais oportunidades de crescimento.

Etapas de implementação

- Faça um inventário do software e da arquitetura usados para serviços relacionados a suas workloads. Decida sobre qual categoria de produtos você quer saber mais.
- Explore as ofertas da AWS para identificar e aprender sobre os serviços e as opções de configuração relevantes que podem ajudar você a melhorar a performance e reduzir os custos e a complexidade operacional.
  - Nuvem Amazon Web Services
  - AWS Academy
  - Quais são as novidades da AWS?
  - Blog da AWS
  - AWS Skill Builder
  - Eventos e webinars da AWS
  - Treinamento da AWS and Certifications

  - Canal da AWS no Youtube
  - Workshops da AWS
  - Comunidades da AWS
- Use o Amazon Q para obter informações e conselhos relevantes sobre serviços.
- Use ambientes sandbox (sem produção) para aprender e experimentar novos serviços sem incorrer em custos adicionais.
- Aprenda constantemente sobre novos serviços e recursos de nuvem.

Recursos

Documentos relacionados:

- Visão geral da Amazon Web Services
- Recursos do Amazon EC2
- Aprenda passo a passo com um plano de aprendizado de parceiro da AWS
- AWS Training and Certification
- Meu caminho de aprendizado para me tornar um arquiteto de soluções da AWS
- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Criar aplicações modernas na AWS

Vídeos relacionados:

- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2022: Reduzir os custos operacionais e de infraestrutura com o Amazon ECS
- AWS re:Invent 2023: Compilar com a eficiência, a agilidade e a inovação da nuvem com o AWS
- AWS re:Invent 2022: Implantar modelos de ML para inferência com alta performance e baixo custo
- Esta é a minha arquitetura

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK

PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas

Use recursos disponibilizados pelo fornecedor de nuvem, como documentação, arquitetos de soluções, serviços profissionais ou parceiros apropriados, para orientar suas decisões durante a escolha da arquitetura. Eles ajudarão a analisar e melhorar sua arquitetura para alcançar a performance ideal.

Práticas comuns que devem ser evitadas:

- Você usa a AWS como um provedor de nuvem comum.
- Você usa os serviços da AWS de uma maneira para a qual eles não foram projetadas.
- Você segue todas as orientações sem considerar seu contexto de negócios.

Benefícios de implementar esta prática recomendada: usar a orientação de um provedor de nuvem ou de um parceiro apropriado pode ajudar a fazer as escolhas de arquitetura certas para as workloads e a conquistar confiança em suas decisões.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A AWS oferece uma ampla variedade de orientações, documentações e recursos que podem ajudar a criar e gerenciar workloads eficientes na nuvem. A documentação da AWS fornece exemplos de código, tutoriais e explicações detalhadas do serviço. Além da documentação, a AWS fornece programas de treinamento e certificação, arquitetos de soluções e serviços profissionais que podem ajudar os clientes a explorar diferentes aspectos dos serviços em nuvem e implementar uma arquitetura de nuvem eficiente na AWS.

Aproveite esses recursos para obter informações sobre conhecimentos valiosos e práticas recomendadas, economizar tempo e obter resultados melhores na Nuvem AWS.

Etapas de implementação

- Analise a documentação e as orientações da AWS e siga as práticas recomendadas. Esses recursos podem ajudar a escolher e configurar serviços com eficiência e obter melhor performance.
  - Documentação da AWS (como guias do usuário e whitepapers)
  - Blog da AWS
  - Treinamento da AWS and Certifications
  - Canal da AWS no Youtube
- Participe de eventos de parceiros da AWS (como os AWS Global Summits, AWS re:Invent, grupos de usuários e workshops) para ouvir dos próprios especialistas da AWS quais são as práticas recomendadas para usar os serviços da AWS.
  - Aprenda passo a passo com um plano de aprendizado de parceiro da AWS
  - Eventos e webinars da AWS
  - Workshops da AWS
  - Comunidades da AWS
- Entre em contato com a AWS para obter assistência quando precisar de mais orientações ou informações sobre produtos. AWS Os arquitetos de soluções e a AWS Professional Services fornecem orientação para a implementação de soluções. AWS Os parceiros oferecem experiência na AWS para ajudar você a desbloquear agilidade e inovação para os negócios.
- Use o Suporte se precisar de suporte técnico para otimizar o uso de um serviço. Nossos planos de suporte são projetados a fim de oferecer a combinação certa de ferramentas e acesso ao conhecimento especializado para ter sucesso com a AWS e melhorar a performance, gerenciar riscos e manter os custos sob controle.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- AWS Enterprise Support

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2023: Padrões avançados orientados a eventos com o Amazon EventBridge
- AWS re:Invent 2023: Implementar padrões de design distribuídos na AWS
- AWS re:Invent 2023: Arquitetura de aplicações como código

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK
- Arquitetura de referência de análise da AWS

PERF01-BP03 Inclua o custo nas decisões de arquitetura

Considere o custo em suas decisões de arquitetura para melhorar a utilização de recursos e a eficiência da performance de suas workloads na nuvem. Quando você está ciente das implicações de custo das suas workloads na nuvem, é mais provável que utilize recursos eficientes e reduza práticas ineficazes.

Práticas comuns que devem ser evitadas:

- Usar somente uma família de instâncias.
- Não avaliar soluções licenciadas em relação a soluções de código aberto.
- Não definir políticas de ciclo de vida de armazenamento.
- Não analisar os novos serviços e recursos da Nuvem AWS.
- Usar somente armazenamento em bloco.

Benefícios de implementar esta prática recomendada: levar em conta o custo em sua tomada de decisão permite que você use recursos mais eficientes e examine outros investimentos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Otimizar as workloads em função do custo pode melhorar a utilização dos recursos e evitar o desperdício em uma workload na nuvem. A consideração do custo nas decisões de arquitetura geralmente inclui o dimensionamento correto dos componentes da workload e a viabilização da elasticidade, o que resulta em maior eficiência da sua performance na nuvem.

Etapas de implementação

- Estabeleça objetivos de custo, como limites orçamentários para a workload na nuvem.
- Identifique os principais componentes (como instâncias e armazenamento) que impulsionam o custo da workload. É possível usar o AWS Calculadora de Preços e o AWS Cost Explorer para identificar os principais fatores de custo na workload.
- Entenda os modelos de preços na nuvem, como instâncias sob demanda, instâncias reservadas, Savings Plans e instâncias spot.
- Use as Práticas recomendadas de otimização de custos do Well-Architected para otimizar esses principais componentes em termos de custo.
- Monitore e analise constantemente os custos para identificar oportunidades de otimizar as workloads e economizar.
  - Use o AWS Budgets para receber alertas quando os custos forem inaceitáveis.
  - Use o AWS Compute Optimizer ou o AWS Trusted Advisor para receber recomendações de otimização de custos.
  - Use a Detecção de Anomalias em Custos da AWS para fazer a detecção automática de anomalias de custo e análise de causa-raiz.

Recursos

Documentos relacionados:

- O que é o Gerenciamento de Faturamento e Custos da AWS?
- Otimização de custos com a AWS
- Escolher uma estratégia de gerenciamento de custos na AWS
- Guia do iniciante em gerenciamento de custos na AWS
- Uma visão geral detalhada do Cost Intelligence Dashboard
- Centro de Arquitetura da AWS

- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2023: Novidades da otimização de custos com a AWS
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2023: Práticas recomendadas de otimização de custos de armazenamento na AWS
- AWS re:Invent 2023: Otimizar custos em seus ambientes com várias contas

Exemplos relacionados:

- Código de demonstração do AWS Compute Optimizer
- Workshop de otimização de custos
- Playbooks de implementação técnica de gerenciamento financeiro na nuvem
- Otimização de startups: ajustar a performance da aplicação para obter a máxima eficiência
- Workshop Otimização sem servidor (performance e custo)
- Escalar arquiteturas econômicas

PERF01-BP04 Avaliar como certas trocas (trade-offs) afetam os clientes e a eficiência da arquitetura

Ao avaliar melhorias relacionadas à performance, determine quais escolhas afetam os clientes e a eficiência das workloads. Por exemplo, se o uso de um datastore de chave-valor aumentar a performance do sistema, é importante avaliar como a alteração afetará os clientes após se tornar permanente

Práticas comuns que devem ser evitadas:

- Você pressupõe que todos os ganhos de performance devem ser implementados, mesmo que seja preciso fazer certas trocas para implementação.
- Você só avalia alterações nas workloads quando um problema de performance atinge um ponto crítico.

Benefícios de implementar esta prática recomendada: ao avaliar possíveis melhorias relacionadas à performance, você deve decidir se as concessões para as alterações são aceitáveis com os requisitos da workload. Em alguns casos, talvez seja necessário implementar controles adicionais para compensar as compensações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Identifique áreas críticas na arquitetura em termos de performance e impacto para o cliente. Determine como é possível promover aprimoramentos, quais compromissos esses aprimoramentos exigem e como eles afetam o sistema e a experiência do usuário. Por exemplo, a implementação de armazenamento de dados em cache pode ajudar a aprimorar drasticamente a performance, mas requer uma estratégia clara de como e quando atualizar ou invalidar dados em cache a fim de prevenir comportamentos incorretos do sistema.

Etapas de implementação

- Entenda os SLAs e requisitos das suas workloads.
- Defina claramente os fatores de avaliação. Os fatores podem estar relacionados a custo, confiabilidade, segurança e performance das workloads.
- Selecione arquitetura e serviços que possam atender às suas necessidades.
- Realize experiências e provas de conceitos (POCs) para avaliar os fatores e o impacto de certas trocas para os clientes e para a eficiência da arquitetura. Normalmente, workloads de alta disponibilidade, com boa performance e seguras consomem mais recursos da nuvem e, ao mesmo tempo, proporcionam uma melhor experiência ao cliente. Entenda as vantagens e desvantagens da complexidade, da performance e do custo da workload. Normalmente, priorizar dois dos fatores inviabiliza o terceiro.

Recursos

Documentos relacionados:

- Amazon Builders' Library
- KPIs do QuickSight
- Amazon CloudWatch RUM

- Documentação do X-Ray
- Entender padrões de resiliência e compromissos para arquitetar de forma eficiente na nuvem

Vídeos relacionados:

- Otimizar aplicações com o Amazon CloudWatch RUM
- AWS re:Invent 2023: Capacidade, disponibilidade, eficiência de custos: escolha três
- AWS re:Invent 2023: padrões de integração avançados e compromissos para sistemas com acoplamento fraco

Exemplos relacionados:

- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM

PERF01-BP05 Usar políticas e arquiteturas de referência

Use políticas internas e arquiteturas de referência existentes ao selecionar serviços e configurações para ser mais eficiente ao projetar e implementar a workload.

Práticas comuns que devem ser evitadas:

- Você permite uma ampla variedade de tecnologias que podem afetar os custos de gerenciamento da empresa.

Benefícios de implementar esta prática recomendada: estabelecer uma política para opções de arquitetura, tecnologia e fornecedor permite que as decisões sejam tomadas rapidamente.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Adotar políticas internas na seleção de recursos e arquitetura fornece padrões e diretrizes a serem seguidos ao fazer escolhas arquitetônicas. Essas diretrizes simplificam o processo de tomada de decisão ao escolher o serviço de nuvem certo e podem ajudar a melhorar a eficiência da performance. Implante a workload usando políticas ou arquiteturas de referência. Integre os serviços

à implantação na nuvem e, depois, use testes de performance para verificar se você pode continuar a atender aos seus requisitos de performance.

Etapas de implementação

- Entenda claramente os requisitos da sua workload na nuvem.
- Revise as políticas internas e externas para identificar as mais relevantes.
- Use as arquiteturas de referência apropriadas fornecidas pela AWS ou as práticas recomendadas do seu setor.
- Crie um continuum que consiste em políticas, padrões, arquiteturas de referência e diretrizes prescritivas para situações comuns. Isso permite que suas equipes ajam mais rapidamente. Adapte os ativos para sua vertical, se aplicável.
- Valide essas políticas e arquiteturas de referência para sua workload em ambientes de sandbox.
- Atualize-se com relação aos padrões do setor e atualizações da AWS para garantir que suas políticas e arquiteturas de referência ajudem a otimizar sua workload na nuvem.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Blog de arquitetura da AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2022: Acelerar a geração de valor para seus negócios com o SAP e a arquitetura de referência da AWS

Exemplos relacionados:

- Exemplos da AWS

- Exemplos do AWS SDK

PERF01-BP06 Usar testes comparativos para orientar decisões de arquitetura

Compare a performance de uma workload existente para entender sua performance na nuvem e orientar decisões de arquitetura com base nesses dados.

Práticas comuns que devem ser evitadas:

- Você depende de testes comparativos comuns que não são indicativos das características da workload.
- Você conta com o feedback e as percepções de clientes como seu único teste comparativo.

Benefícios de implementar esta prática recomendada: o benchmarking da sua implementação atual permite medir melhorias de performance.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use testes comparativos com testes sintéticos para avaliar a performance dos componentes da workload. O benchmarking é usado na avaliação da tecnologia para um componente específico e geralmente é mais simples de configurar do que testes de carga. Muitas vezes o benchmarking é usado no início de um novo projeto, quando ainda não há uma solução completa para o teste de carga.

Você pode criar seus próprios testes comparativos personalizados ou usar um teste padrão do setor, como o TPC-DS, para comparar suas workloads. Os benchmarks do setor são úteis ao comparar ambientes. Já os benchmarks personalizados são úteis para direcionar a tipos específicos de operações que você espera realizar em sua arquitetura.

Ao realizar testes comparativos, é importante "preaquecer" o ambiente de teste para obter resultados válidos. Execute o mesmo teste comparativo várias vezes para verificar a captura de qualquer variação ao longo do tempo.

Como normalmente é mais rápido executar testes comparativos do que testes de carga, eles podem ser usados mais cedo no pipeline de implantação e fornecer um feedback mais rápido sobre

desvios de performance. Ao avaliar uma alteração significativa em um componente ou serviço, o teste comparativo pode ser uma maneira rápida de verificar se é possível justificar a iniciativa para concretizar a alteração. O uso de testes comparativos em conjunto com testes de carga é importante porque o teste de carga informa como é a performance da workload no ambiente de produção.

Etapas de implementação

- Planeje e defina:
  - Defina os objetivos, o parâmetro de referência, os cenários de teste, as métricas (como utilização da CPU, latência ou throughput) e os KPIs para o teste comparativo.
  - Concentre-se nos requisitos do usuário em termos de experiência do usuário e em outros fatores, como tempo de resposta e acessibilidade.
  - Identifique uma ferramenta de testes comparativos adequada à workload. Você pode usar serviços da AWS como o Amazon CloudWatch ou uma ferramenta de terceiros que seja compatível com a workload.
- Configure e instrumente:
  - Prepare o ambiente e configure os recursos.
  - Implemente monitoramento e registro em log para capturar os resultados dos testes.
- Compare e monitore:
  - Execute testes comparativos e monitore as métricas durante o teste.
- Analise e documente:
  - Documente o processo de comparação e as descobertas.
  - Analise os resultados para identificar gargalos, tendências e áreas para melhoria.
  - Use os resultados do teste para tomar decisões de arquitetura e ajustar a workload. Isso pode incluir a mudança de serviços ou a adoção de novos recursos.
- Otimize e repita:
  - Ajuste as configurações e alocações de recursos com base nos testes comparativos.
  - Teste novamente a workload depois do ajuste para validar as melhorias.
  - Documente seu aprendizado e repita o processo para identificar outras áreas para melhoria.

Recursos

Documentos relacionados:

- Centro de Arquitetura da AWS
- AWS Partner Network
- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Fluxos de trabalho de genômica, Parte 5: benchmarking automatizado
- Comparar e otimizar a implantação de endpoints no Amazon SageMaker AI JumpStart

Vídeos relacionados:

- AWS re:Invent 2023: Testes comparativos de partida a frio do AWS Lambda
- Testes comparativos de serviços com estado na nuvem
- Esta é a minha arquitetura
- Otimizar aplicações com o Amazon CloudWatch RUM
- Demonstração do Amazon CloudWatch Synthetics

Exemplos relacionados:

- Exemplos da AWS
- Exemplos do AWS SDK
- Testes de carga distribuídos
- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM

PERF01-BP07 Usar uma abordagem baseada em dados para escolhas de arquitetura

Defina uma abordagem clara e baseada em dados para escolhas de arquitetura a fim de verificar se os serviços e configurações de nuvem corretos são usados para atender às suas necessidades comerciais específicas.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não deve ser atualizada ao longo do tempo.
- Suas escolhas de arquitetura são baseadas em suposições.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa.

Benefícios de implementar esta prática recomendada: ao aplicar uma abordagem bem definida para fazer escolhas de arquitetura, você usa dados para influenciar o projeto das workloads e tomar decisões conscientes ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Use a experiência interna e o conhecimento da nuvem ou de recursos externos, como casos de uso publicados ou whitepapers, para escolher recursos e serviços em sua arquitetura. Você deve ter um processo bem definido que incentive a experimentação e os testes comparativos com os serviços que podem ser usados em suas workloads.

Os atrasos de workloads críticas devem consistir não apenas em histórias de usuários que venham a oferecer funcionalidades relevantes para os negócios e usuários, mas também em histórias técnicas que formem uma base de arquitetura para as workloads. Essa base é formada por novos avanços em tecnologia e novos serviços e os adota em função de dados e justificativas adequadas. Isso verifica se a arquitetura permanece preparada para o futuro e não se torna estagnada.

Etapas de implementação

- Interaja com as principais partes interessadas para definir os requisitos das workloads, incluindo considerações de performance, disponibilidade e custo. Considere fatores como o número de usuários e o padrão de uso das workloads.
- Crie uma base de arquitetura ou uma lista de pendências de tecnologia que seja priorizada junto com a lista de pendências funcional.
- Avalie diferentes serviços em nuvem (para obter mais detalhes, consulte PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis).
- Explore diferentes padrões de arquitetura, como microsserviços ou tecnologia sem servidor, que atendem aos requisitos de performance (para obter mais detalhes, consulte PERF01-BP02 Usar a orientação do provedor de nuvem ou de um parceiro apropriado para aprender sobre padrões de arquitetura e práticas recomendadas).

- Consulte outras equipes, diagramas de arquitetura e recursos, como arquitetos de soluções da AWS, o Centro de Arquitetura da AWS e o AWS Partner Network, para obter ajuda para escolher a arquitetura certa para sua workload.

- Defina métricas de performance, como throughput e tempo de resposta, que podem ajudar você a avaliar a performance das workloads.
- Experimente e use métricas definidas para validar a performance da arquitetura selecionada.
- Monitore e faça ajustes contínuos conforme necessário para manter a performance ideal da arquitetura.
- Documente a arquitetura e as decisões selecionadas como referência para futuras atualizações e aprendizados.
- Revise e atualize constantemente a abordagem para seleção de arquitetura com base em aprendizados, novas tecnologias e métricas. Esses parâmetros podem indicar que é necessário mudar ou que há algum problema na abordagem atual.

Recursos

Documentos relacionados:

- Biblioteca de Soluções da AWS
- Centro de Conhecimentos da AWS
- Padrões de arquitetura para criar aplicações orientadas a dados fim a fim na AWS

Vídeos relacionados:

- Esta é a minha arquitetura
- AWS re:Invent 2021: Empresa orientada a dados: da visão ao valor
- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos

Exemplos relacionados:

- Exemplos da AWS

- Exemplos do AWS SDK

Computação e hardware

A opção ideal de computação para uma workload específica pode variar de acordo com o design, os padrões de uso e as definições de configuração da aplicação. As arquiteturas podem usar diferentes opções de computação para vários componentes e permitir diferentes recursos para aprimorar a performance. A seleção da opção de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Essa área de foco compartilha orientações e práticas recomendadas sobre como identificar e otimizar as opções de computação para eficiência de performance na nuvem.

Práticas recomendadas

- PERF02-BP01 Selecionar as melhores opções de computação para as workloads
- PERF02-BP02 Entender a configuração e os recursos de computação disponíveis
- PERF02-BP03 Coletar métricas relacionadas à computação
- PERF02-BP04 Configurar e dimensionar corretamente os recursos de computação
- PERF02-BP05 Dimensionar recursos de computação dinamicamente
- PERF02-BP06 Usar aceleradores de computação baseados em hardware otimizados

PERF02-BP01 Selecionar as melhores opções de computação para as workloads

Selecionar a opção de computação mais adequada para suas workloads permite melhorar a performance, reduzir os custos desnecessários de infraestrutura e reduzir os esforços operacionais necessários para mantê-las.

Práticas comuns que devem ser evitadas:

- A mesma opção de computação utilizada on-premises é usada.
- Você não tem conhecimento das opções, dos atributos e das soluções de computação em nuvem e de como essas soluções podem melhorar a performance computacional.
- Uma opção de computação existente é provisionada de forma excessiva para atender aos requisitos de ajuste de escala ou performance quando uma opção alternativa de computação se alinharia às características da workload com mais precisão.

Benefícios de implementar esta prática recomendada: ao identificar os requisitos de computação e avaliar as opções disponíveis, você pode tornar a workload mais eficiente em termos de recursos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para otimizar as workloads na nuvem quanto à eficiência de performance, é importante selecionar as opções de computação mais apropriadas para seu caso de uso e requisitos de performance. A AWS fornece uma variedade de opções de computação que atendem a diferentes workloads na nuvem. Por exemplo, você pode usar o Amazon EC2 para iniciar e gerenciar servidores virtuais, o AWS Lambda para executar código sem precisar provisionar ou gerenciar servidores, o Amazon ECS ou o Amazon EKS para executar e gerenciar contêineres ou o AWS Batch para processar grandes volumes de dados em paralelo. Com base em sua escala e necessidades de computação, você deve escolher e configurar a solução ideal para sua situação. Você também pode considerar o uso de vários tipos de soluções de computação em uma única workload, pois cada uma tem suas próprias vantagens e desvantagens.

As etapas a seguir orientam você na seleção das opções de computação certas para atender às características da workload e aos requisitos de performance.

Etapas de implementação

- Entenda os requisitos de computação das workloads. Os principais requisitos a serem considerados incluem necessidades de processamento, padrões de tráfego, padrões de acesso a dados, necessidades de ajuste de escala e requisitos de latência.
- Saiba mais sobre os diferentes serviços de computação da AWS para sua workload. Para ter mais informações, consulte PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis. Veja algumas das principais opções de computação da AWS, as características e casos de uso comuns:

Serviço da AWS	Características principais	Casos de uso comuns
Amazon Elastic Compute Cloud (Amazon EC2)	Oferece opção dedicada para hardware, requisitos de licença, grande seleção de diferentes famílias de instâncias, tipos de	Migrações do tipo mover sem alterações (lift-and-shift), aplicações monolíticas, ambientes híbridos, aplicaçõe s empresariais

Serviço da AWS	Características principais	Casos de uso comuns
	processadores e acelerado res de computação.	
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Implantação fácil, ambientes consistentes, escaláveis	Microsserviços, ambientes híbridos
AWS Lambda	Serviço de computação sem servidor que executa código em resposta a eventos e gerencia automaticamente os recursos computacionais subjacentes.	Microsserviços, aplicações orientadas a eventos
AWS Batch	Provisiona e escala de forma eficiente e dinâmica os recursos de computação do Amazon Elastic Container Service (Amazon ECS), do Amazon Elastic Kubernete s Service (Amazon EKS) e do AWS Fargate, oferecend o a opção de usar instância s sob demanda ou spot com base em seus requisitos de trabalho	HPC, treinamento de modelos de ML.
Amazon Lightsail	Aplicação Linux e Windows pré-configurada para executar pequenas workloads .	Aplicações Web simples, site personalizado.

- Avalie o custo (como cobrança por hora ou transferência de dados) e as despesas gerais de gerenciamento (como aplicação de patches e ajuste de escala) associados a cada opção de computação.

- Faça experimentos e análises comparativas em um ambiente de não produção para identificar qual opção de computação pode atender melhor às necessidades da workload.
- Depois de experimentar e identificar sua nova solução de computação, planeje a migração e valide as métricas de performance.
- Use ferramentas de monitoramento da AWS, como o Amazon CloudWatch, e serviços de otimização, como o AWS Compute Optimizer, para otimizar constantemente a computação com base em padrões de uso real.

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Funções: configuração da função do Lambda
- Recomendações para contêineres
- Recomendações para tecnologia sem servidor

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preços para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon Elastic Compute Cloud no AMS
- AWS re:Invent 2023: Novidades do Amazon Elastic Compute Cloud
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos no Amazon Elastic Compute Cloud
- AWS re:Invent 2021: Habilitar o Amazon Elastic Compute Cloud da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Otimizar a performance e os custos para sua computação na AWS
- AWS re:Invent 2019: Fundamentos da Amazon Elastic Compute Cloud
- AWS re:Invent 2022: Implantar modelos de ML para inferência com alta performance e baixo custo

- AWS re:Invent 2019: Otimizar a performance e os custos para sua computação na AWS
- Fundamentos do Amazon EC2
- Implemente modelos de ML para inferência com alta performance e baixo custo

Exemplos relacionados:

- Migrar aplicações Web para contêineres
- Executar uma aplicação Hello World sem servidor
- Workshop do Amazon EKS
- Workshop do Amazon EC2
- Workloads eficientes e resilientes com o Amazon Elastic Compute Cloud Auto Scaling
- Migrar para o AWS Graviton com serviços de contêiner

PERF02-BP02 Entender a configuração e os recursos de computação disponíveis

Entenda as opções de configuração e os recursos disponíveis para seu serviço de computação a fim de ajudar a provisionar a quantidade certa de recursos e melhorar a eficiência de performance.

Práticas comuns que devem ser evitadas:

- Não avaliar as opções de computação ou as famílias de instâncias disponíveis em relação às características da workload.
- Provisionar recursos de computação em excesso para atender aos requisitos de pico de demanda.

Benefícios de implementar esta prática recomendada: familiarizar-se com os atributos e as configurações de computação da AWS a fim de poder usar uma solução de computação otimizada para atender às características e às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Cada solução de computação tem configurações e recursos exclusivos disponíveis para acomodar diferentes características e requisitos das workloads. Saiba como essas opções complementam sua workload e determine quais opções de configuração são melhores para sua aplicação. Exemplos

dessas opções são famílias de instâncias, tamanhos, recursos (GPU, E/S), expansão, tempos limite, tamanhos de função, instâncias de contêineres e simultaneidade. Se a workload estiver usando a mesma opção de computação há mais de quatro semanas, e se a previsão for de que as características permanecerão as mesmas no futuro, você poderá usar o AWS Compute Optimizer para descobrir se sua opção de computação atual é adequada para as workloads de uma perspectiva de CPU e memória.

Etapas de implementação

- Entenda os requisitos da workload (como necessidade de CPU, memória e latência).
- Analise a documentação e as práticas recomendadas da AWS para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar a performance da computação. Aqui estão algumas das principais opções de configuração a serem consideradas:

Opção de configuração	Exemplos
Tipo de instância	• As instâncias otimizadas para computação são ideais para workloads que exigem uma proporção maior de vCPU/memória. • As instâncias otimizadas para memória entregam grandes quantidades de memória para oferecer compatibilidade com as workloads com uso intenso de memória. • As instâncias otimizadas para armazenamento são projetadas para workloads que exigem alta leitura sequencial e acesso de gravação (IOPS) no armazenamento local.
Modelo de definição de preços	• As instâncias sob demanda permitem usar a capacidade de computação por hora ou segundo sem uma confirmação de longo prazo. Essas instâncias são ideais para expansões acima das necessidades de performance da linha de base. • Os Savings Plans oferecem economias significativas em relação às instâncias sob

Opção de configuração	Exemplos
	demanda em troca do compromisso de usar uma quantidade específica de potência computacional por um período de um ou três anos. • As instâncias spot permitem que você aproveite a capacidade de instância não utilizada com um desconto para as workloads sem estado e tolerantes a falhas.
Auto Scaling	Use a configuração de Auto Scaling para combinar recursos computacionais com padrões de tráfego.
Dimensionamento	• Use o Compute Optimizer para obter uma recomendação de machine learning sobre a configuração de computação que correspon de melhor às características da computaçã o. • Use o AWS Lambda Power Tuning para selecionar a melhor configuração para a função do Lambda.
Aceleradores de computação baseados em hardware	• As instâncias com computação acelerada executam funções como processamento gráfico ou correspondência de padrões de dados com mais eficiência do que as alternativas baseadas em CPU. • Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1

Recursos

Documentos relacionados:

- Computação em nuvem com a AWS
- Tipos de instância do Amazon EC2
- Controle do estado do processador para sua instância do Amazon EC2
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Funções: configuração da função do Lambda

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon EC2 no AWS Management Console
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2
- AWS re:Invent 2022: Otimizar o Amazon EKS para performance e custo na AWS

Exemplos relacionados:

- Código de demonstração do Compute Optimizer
- Workshop sobre instâncias spot do Amazon EC2
- Workloads eficientes e resilientes com o Amazon EC2 AWS Auto Scaling
- Workshop de desenvolvedores para Graviton
- Dia de imersão em workloads da AWS para Microsoft
- Dia de imersão em workloads da AWS para Linux
- Código de demonstração do AWS Compute Optimizer

- Workshop do Amazon EKS

PERF02-BP03 Coletar métricas relacionadas à computação

Registre e acompanhe métricas relacionadas à computação para entender melhor a performance dos seus recursos e melhorar sua performance e utilização.

Práticas comuns que devem ser evitadas:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só usa as métricas padrão registradas pelo software de monitoramento.
- Você só revisa as métricas quando há um problema.

Benefícios de implementar esta prática recomendada: a coleta de métricas relacionadas à performance ajudará você a alinhar a performance da aplicação aos requisitos empresariais para garantir que você atenda às necessidades da workload. Isso também pode ajudar a melhorar constantemente a performance e a utilização dos recursos na workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

As workloads na nuvem podem gerar grandes volumes de dados, como métricas, logs e eventos. Na Nuvem AWS, coletar métricas é uma etapa essencial para melhorar a segurança, a eficiência de custos, a performance e a sustentabilidade. A AWS oferece uma ampla variedade de métricas relacionadas à performance usando serviços de monitoramento, por exemplo, o Amazon CloudWatch, para fornecer insights valiosos. Métricas como utilização de CPU, utilização de memória, E/S de disco e entrada e saída da rede podem fornecer informações sobre os níveis de utilização ou gargalos de performance. Use essas métricas como parte de uma abordagem impulsionada por dados para ajustar e otimizar ativamente os recursos de sua workload. Em um caso ideal, você deve coletar todas as métricas relacionadas aos recursos de computação em uma única plataforma com políticas de retenção implementadas para apoiar as metas operacionais e de custo.

Etapas de implementação

- Identifique quais métricas relacionadas à performance são relevantes para a workload. Você deve coletar métricas sobre a utilização de recursos e a forma como a workload na nuvem está operando (como tempo de resposta e throughput).
  - Métricas padrão do Amazon EC2
  - Métricas padrão do Amazon ECS
  - Métricas padrão do Amazon EKS
  - Métricas padrão do Lambda
  - Métricas de memória e disco do Amazon EC2
- Escolha e configure a solução certa de registro e monitoramento para a workload.
  - Observabilidade nativa da AWS
  - AWS Distro para OpenTelemetry
  - Amazon Managed Service for Prometheus
- Defina o filtro e a agregação necessários para as métricas com base nos requisitos da workload.
  - Quantificar méticas de aplicações personalizadas com o Amazon CloudWatch Logs e filtros métricos
  - Coletar métricas personalizadas com a marcação com tags estratégica do Amazon CloudWatch
- Configure políticas de retenção de dados para que as métricas correspondam às metas operacionais e de segurança.
  - Retenção de dados padrão para métricas do CloudWatch
  - Retenção de dados padrão para CloudWatch Logs
- Se necessário, crie alarmes e notificações para as métricas a fim de ajudar a reagir proativamente a problemas relacionados à performance.
  - Criar alarmes para métricas personalizadas usando a detecção de anomalias do Amazon CloudWatch
  - Criar métricas e alarmes para páginas da Web específicas com o Amazon CloudWatch RUM
- Use a automação para implantar os agentes de agregação de métricas e logs.
  - Automação do AWS Systems Manager
  - Coletor do OpenTelemetry

Recursos

Documentos relacionados:

- Monitoramento e observabilidade
- Práticas recomendadas: implementar a observabilidade com a AWS
- Documentação do Amazon CloudWatch
- Coletar métricas e logs de instâncias do Amazon EC2 e servidores on-premises com o agente do CloudWatch
- Acessar o Amazon CloudWatch Logs para AWS Lambda
- Usar o CloudWatch Logs com Instâncias de contêiner
- Publicar métricas personalizadas
- AWS Answers: log centralizado
- Serviços da AWS que publicam métricas do CloudWatch
- Monitorar o Amazon EKS no AWS Fargate

Vídeos relacionados:

- AWS re:Invent 2023 [LANÇAMENTO]: Monitoramento de aplicações para workloads modernas
- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS re:Invent 2023: Observabilidade direta com o AWS Distro para OpenTelemetry
- Gerenciamento da performance de aplicações na AWS

Exemplos relacionados:

- Dia de imersão em workloads da AWS para Linux - Amazon CloudWatch
- Monitorar clusters e contêineres do Amazon ECS
- Monitorar com painéis do Amazon CloudWatch
- Workshop do Amazon EKS

PERF02-BP04 Configurar e dimensionar corretamente os recursos de computação

Configure e dimensione corretamente os recursos de computação para atender aos requisitos de performance das workloads e evitar que recursos sejam subutilizados ou usados em excesso.

Práticas comuns que devem ser evitadas:

- Ignorar os requisitos de performance das workloads, o que ocasiona recursos computacionais superprovisionados ou subprovisionados.
- Você escolhe somente a maior ou a menor instância disponível para todas as workloads.
- Você usa apenas uma família de instâncias para facilitar o gerenciamento.
- Você ignora as recomendações do AWS Cost Explorer ou do Compute Optimizer para o dimensionamento correto.
- Você não reavalia a workload quanto à adequação dos novos tipos de instância.
- Você certifica apenas um pequeno número de configurações de instâncias para sua organização.

Benefícios de implementar esta prática recomendada: o dimensionamento correto dos recursos computacionais garante a operação ideal na nuvem, evitando o provisionamento excessivo e o subprovisionamento de recursos. O dimensionamento adequado dos recursos de computação normalmente resulta em melhor performance e melhor experiência do cliente, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

O dimensionamento correto permite que as organizações operem a infraestrutura de nuvem de forma eficiente e econômica ao mesmo tempo que atendem às suas necessidades comerciais. O provisionamento excessivo de recursos na nuvem pode gerar custos extras, enquanto o provisionamento insuficiente pode resultar em baixa performance e em uma experiência negativa para o cliente. A AWS fornece ferramentas como AWS Compute Optimizer e AWS Trusted Advisor que usam dados históricos para fornecer recomendações para dimensionar corretamente seus recursos computacionais.

Etapas de implementação

- Escolha um tipo de instância que melhor atenda às suas necessidades:

  - Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
  - Seleção de tipo de instância baseada em atributos para o Amazon EC2 Fleet
  - Criar um grupo do Auto Scaling usando seleção de tipo de instância baseada em atributos
  - Otimizar seus custos de computação do Kubernetes com a consolidação do Karpenter
- Analise as várias características de performance da sua workload e como elas se relacionam a uso de memória, rede e CPU. Use esses dados para escolher os recursos que melhor correspondam ao perfil e às metas de performance da workload.
- Monitore o uso de recursos usando ferramentas de monitoramento da AWS, como o Amazon CloudWatch.
- Selecione a configuração correta para os recursos computacionais.
  - Para workloads efêmeras, avalie as métricas da instância do Amazon CloudWatch, como `CPUUtilization`, para identificar se a instância está subutilizada ou superutilizada.
  - Para workloads estáveis, verifique as ferramentas de dimensionamento correto da AWS, como AWS Compute Optimizer e AWS Trusted Advisor em intervalos regulares para identificar oportunidades de otimizar e dimensionar corretamente o recurso de computação.
- Teste as alterações na configuração em um ambiente de não produção antes de implementá-las em um ambiente ativo.
- Reavalie constantemente novas ofertas de computação e compare-as com as necessidades da workload.

Recursos

Documentos relacionados:

- Computação na nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Funções: configuração da função do Lambda
- Controle do estado do processo para sua instância do Amazon EC2

Vídeos relacionados:

- Fundamentos do Amazon EC2
- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa do Amazon EC2 no AWS Management Console
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2

Exemplos relacionados:

- Código de demonstração do AWS Compute Optimizer
- Workshop do Amazon EKS
- Recomendações de dimensionamento correto

PERF02-BP05 Dimensionar recursos de computação dinamicamente

Use a elasticidade da nuvem para aumentar ou diminuir os recursos de computação dinamicamente a fim de atender às suas necessidades e evitar provisionamento excessivo ou insuficiente da capacidade para a workload.

Práticas comuns que devem ser evitadas:

- Reagir a alarmes aumentando a capacidade manualmente.
- Usar as mesmas diretrizes de dimensionamento (geralmente infraestrutura estática) do ambiente on-premises.
- Manter a capacidade aumentada após um evento de ajuste de escala, em vez de reduzi-la novamente.

Benefícios de implementar esta prática recomendada: configurar e testar a elasticidade dos recursos computacionais pode ajudar você a economizar dinheiro, manter os benchmarks de performance e melhorar a confiabilidade à medida que o tráfego muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS oferece a flexibilidade de aumentar ou diminuir seus recursos dinamicamente por meio de uma variedade de mecanismos de ajuste de escala a fim de atender às mudanças na demanda. Combinado com métricas relacionadas à computação, um ajuste de escala dinâmico permite que as workloads respondam automaticamente às mudanças e usem o conjunto ideal de recursos computacionais para atingir sua meta.

Você pode usar diversas abordagens diferentes para corresponder a oferta de recursos com a demanda.

- Abordagem de monitoramento de meta: monitore a métrica de ajuste de escala e aumente ou diminua automaticamente a capacidade conforme necessário.
- Ajuste de escala preditivo: aumente ou reduza a escala em antecipação às tendências diárias e semanais.
- Abordagem baseada em cronograma: defina seu próprio cronograma de ajuste de escala de acordo com as mudanças de carga previsíveis.
- Ajuste de escala de serviços: escolha serviços (como de tecnologia sem servidor) que sejam escalados automaticamente de acordo com o projeto.

É necessário garantir que as implantações de workload possam lidar com eventos de expansão e redução da escala.

Etapas de implementação

- Instâncias, contêineres e funções de computação oferecem mecanismos para elasticidade, seja em combinação com o ajuste de escala automático ou como um recurso do serviço. Veja alguns exemplos de mecanismos de ajuste de escala automático:

Mecanismo de ajuste de escala automático	Onde usar
Amazon EC2 Auto Scaling	Ajuda a garantir que você tenha o número correto de instâncias do Amazon EC2 disponíveis para processar a carga da aplicação.
Application Auto Scaling	Para escalar automaticamente os recursos para serviços da AWS individuais além do Amazon EC2, como funções do AWS Lambda ou serviços do Amazon Elastic Container Service (Amazon ECS).
Kubernetes Cluster Autoscaler/Karpenter	Para escalar automaticamente os clusters do Kubernetes.

- O ajuste de escala geralmente é discutido em relação a serviços de computação, como instâncias do Amazon EC2 ou funções do AWS Lambda. Não se esqueça de considerar também a configuração de serviços não computacionais, como AWS Glue, para atender à demanda.
- Verifique se as métricas de ajuste de escala correspondem às características da workload que está sendo implantada. Se você estiver implantando uma aplicação de transcodificação de vídeo, espera-se que a utilização da CPU seja de 100%, e essa não deve ser sua métrica principal. Use a profundidade da fila de trabalhos de transcodificação. Você pode usar uma métrica personalizada para sua política de ajuste de escala, se necessário. Para escolher as métricas certas, considere a seguinte orientação para o Amazon EC2:
  - A métrica deve ser uma métrica de utilização válida e descrever o quanto uma instância está ocupada.
  - O valor da métrica deve aumentar e diminuir em proporção ao número das instâncias no grupo do Auto Scaling.
- Certifique-se de usar o ajuste de escala dinâmico em vez do ajuste de escala manual para seu grupo do Auto Scaling. Também recomendamos usar políticas de ajuste de escala de rastreamento de metas em seu ajuste de escala dinâmico.
- Verifique se as implantações da workload podem lidar com os dois eventos de ajuste de escala (aumento e redução). Como exemplo, você pode usar o Histórico de atividades para verificar uma atividade de ajuste de escala em um grupo do Auto Scaling.

- Avalie sua workload em relação a padrões previsíveis e, ao antecipar alterações previstas e planejadas na demanda, ajuste a escala proativamente. Com o ajuste de escala preditivo, é possível eliminar a necessidade de superprovisionar a capacidade. Para obter mais informações, consulte Ajuste de escala preditivo com o Amazon EC2 Auto Scaling.

Recursos

Documentos relacionados:

- Computação na nuvem com a AWS
- Tipos de instância do Amazon EC2
- Contêineres do Amazon ECS: instâncias de contêiner do Amazon ECS
- Contêineres do Amazon EKS: nós de processamento do Amazon EKS
- Funções: configuração da função do Lambda
- Controle do estado do processo para sua instância do Amazon EC2
- Mergulho profundo no ajuste de escala automático de clusters do Amazon ECS
- Introdução ao Karpenter: um dimensionador automático de clusters do Kubernetes de código aberto e alta performance

Vídeos relacionados:

- AWS re:Invent 2023: AWS Graviton: a melhor performance de preço para suas workloads da AWS
- AWS re:Invent 2023: Novos recursos de IA generativa no Amazon EC2 no console de gerenciamento da AWS
- AWS re:Invent 2023: Novidades do Amazon EC2
- AWS re:Invent 2023: Economias inteligentes: estratégias de otimização de custos com o Amazon EC2
- AWS re:Invent 2021: Capacitar o Amazon EC2 da próxima geração: mergulho profundo no Nitro System
- AWS re:Invent 2019: Fundamentos do Amazon EC2

Exemplos relacionados:

- Exemplos de grupos do Amazon EC2 Auto Scaling

- Workshop do Amazon EKS
- Escalar suas workloads do Amazon EKS executando-as em IPv6

PERF02-BP06 Usar aceleradores de computação baseados em hardware otimizados

Use aceleradores de hardware para executar determinadas funções com mais eficiência do que as alternativas baseadas em CPU.

Práticas comuns que devem ser evitadas:

- Em sua workload, você não compara uma instância de uso geral com uma instância criada para um propósito específico capaz de oferecer maior performance e menor custo.
- Você está usando aceleradores de computação baseados em hardware para tarefas que podem ser eficientes com o uso de alternativas baseadas em CPU.
- Você não está monitorando o uso da GPU.

Benefícios de implementar esta prática recomendada: ao usar aceleradores baseados em hardware, como unidades de processamento gráfico (GPUs) e matrizes de portas programáveis em campo (FPGAs), você pode executar determinadas funções de processamento com mais eficiência.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

As instâncias com computação acelerada fornecem acesso a aceleradores de computação baseados em hardware, como GPUs e FPGAs. Esses aceleradores de hardware executam certas funções, como processamento gráfico ou correspondência de padrões de dados, com mais eficiência do que alternativas baseadas em CPU. Muitas workloads aceleradas, como renderização, transcodificação e machine learning, são altamente variáveis em termos de uso de recursos. Execute esse hardware apenas pelo tempo necessário e desative-o com automação quando não precisar mais dele para melhorar a eficiência da performance geral.

Etapas de implementação

- Identifique quais instâncias com computação acelerada podem atender às suas necessidades.

- Para workloads de machine learning, utilize hardware específico para sua workload, como AWS Trainium, AWS Inferentia e Amazon EC2 DL1. AWS Instâncias do Inferentia, como instâncias Inf2, oferecem performance até 50% melhor por watt em relação a instâncias comparáveis do Amazon EC2.
- Colete métricas de uso para as instâncias com computação acelerada. Por exemplo, você pode usar o agente do CloudWatch para coletar métricas como `utilization_gpu` e `utilization_memory` e para suas GPUs, conforme mostrado em Coletar métricas de GPU NVIDIA com o Amazon CloudWatch.
- Otimize o código, a operação de rede e as configurações dos aceleradores de hardware para garantir que o hardware subjacente seja totalmente utilizado.
  - Otimizar as configurações da GPU
  - Monitoramento e otimização da GPU na AMI de aprendizado profundo
  - Otimizar a E/S para ajuste de performance da GPU de treinamento de aprendizado profundo no Amazon SageMaker AI
- Use as mais recentes bibliotecas de alta performance e drivers de GPU.
- Use automação para liberar instâncias de GPU quando não estiverem em uso.

Recursos

Documentos relacionados:

- Trabalhar com GPUs no Amazon Elastic Container Service
- Instâncias de GPU
- Instâncias com AWS Trainium
- Instâncias com o AWS Inferentia
- Vamos arquitetar! Como arquitetar com chips e aceleradores personalizados

- Computação acelerada
- Instâncias VT1 do Amazon EC2
- Como faço para escolher o tipo de instância do Amazon EC2 apropriado para minha workload?
- Escolher o melhor acelerador de IA e compilação de modelos para inferência de visão computacional com o Amazon SageMaker AI

Vídeos relacionados:

- AWS re:Invent 2021: Como selecionar instâncias de GPU do Amazon Elastic Compute Cloud para aprendizado profundo
- AWS re:Invent 2022 [NOVO LANÇAMENTO!]: Introdução as instâncias Inf2 do Amazon EC2 baseadas no AWS Inferentia2
- AWS re:Invent 2022: Acelerar o aprendizado profundo e inovar com mais rapidez com o AWS Trainium
- AWS re:Invent 2022: Aprendizado profundo na AWS com a NVIDIA: do treinamento à implantação

Exemplos relacionados:

- Amazon SageMaker AI e NVIDIA GPU Cloud (NGC)
- Usar o SageMaker AI com Trainium e Inferentia para workloads otimizadas de treinamento e de inferência em aprendizado profundo
- Otimizar modelos de PLN com instâncias Inf1 do Amazon Elastic Compute Cloud no Amazon SageMaker AI

Gerenciamento de dados

A solução de gerenciamento de dados ideal para um sistema específico varia conforme o tipo de dados (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o throughput necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. As workloads do Well-Architected usam datastores específicos que permitem que recursos diferentes melhorem a performance.

Essa área de foco compartilha orientações e práticas recomendadas para otimizar o armazenamento de dados, os padrões de movimentação e acesso, e a eficiência de performance dos armazenamentos de dados.

Práticas recomendadas

- PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados
- PERF03-BP02 Avaliar as opções de configuração disponíveis para o datastore
- PERF03-BP03 Coletar e registrar métricas de performance do datastore
- PERF03-BP04 Implementar estratégias para melhorar a performance da consulta no datastore
- PERF03-BP05 Implementar padrões de acesso a dados que utilizam cache

PERF03-BP01 Usar um datastore com propósitos específicos que melhor atenda aos requisitos de acesso e armazenamento de dados

Entenda as características dos dados (como possibilidade de compartilhamento, tamanho, tamanho do cache, padrões de acesso, latência, throughput e persistência dos dados) a fim de selecionar os datastores com propósito específico (armazenamento ou banco de dados) para sua workload.

Práticas comuns que devem ser evitadas:

- Utilizar um único datastore porque há experiência e conhecimento internos de um tipo específico de solução de banco de dados.
- Você pressupõe que todas as workloads têm requisitos de acesso e armazenamento de dados semelhantes.

- Você não implementou um catálogo de dados para criar um inventário de seus ativos de dados.

Benefícios de implementar esta prática recomendada: entender as características e os requisitos de dados permite que você determine a tecnologia de armazenamento mais eficiente e com melhor performance adequada às necessidades da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Ao selecionar e implementar o armazenamento de dados, certifique-se de que as características de consulta, ajuste de escala e armazenamento atendam aos requisitos de dados da workload. A AWS fornece várias tecnologias de armazenamento de dados e banco de dados, incluindo armazenamento em blocos, armazenamento de objetos, armazenamento de streaming, sistema de arquivos, bancos de dados relacionais, de chave-valor, de documentos, na memória, de grafos, de séries temporais e ledger. Cada solução de gerenciamento de dados tem opções e configurações disponíveis para compatibilidade com seus casos de uso e modelos de dados. Ao compreender as características e os requisitos dos dados, você pode se separar da tecnologia de armazenamento monolítico e das abordagens restritivas e únicas para se concentrar no gerenciamento adequado dos dados.

Etapas de implementação

- Realize um inventário dos vários tipos de dados que existem na workload.
- Entenda e documente as características e os requisitos dos dados, incluindo:
  - Tipo de dados (não estruturados, semiestruturados, relacionais)
  - Volume e crescimento de dados
  - Durabilidade dos dados: persistentes, efêmeros, transitórios
  - Requisitos de ACID (atomicidade, consistência, isolamento, durabilidade)
  - Padrões de acesso a dados (com muita leitura ou gravação)
  - Latência
  - Throughput
  - IOPS (operações de entrada/saída por segundo)
  - Período de retenção de dados
- Conheça os diferentes datastores (serviços de armazenamento e banco de dados) disponíveis para a workload na AWS que podem atender às características dos dados (conforme descrito em

PERF01-BP01 Conhecer e compreender os serviços e recursos de nuvem disponíveis. Alguns exemplos de tecnologias de armazenamento da AWS e suas principais características incluem:

Tipo	Serviços da AWS	Características principais
Armazenamento de objetos	Amazon S3	Escalabilidade ilimitada, alta disponibilidade e várias opções de acessibilidade. A transferência e o acesso a objetos dentro e fora do Amazon S3 podem usar um serviço, como o Transfer Acceleration ou Access Points, para oferecer suporte a sua localização, necessida des de segurança e padrões de acesso.
Armazenamento de arquivamento	Amazon S3 Glacier	Desenvolvido para arquivame nto de dados.
Armazenamento de streaming	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Ingestão e armazenam ento eficientes de dados de streaming.
Sistema de arquivos compartilhado	Amazon Elastic File System (Amazon EFS)	Sistema de arquivos montável que pode ser acessado por vários tipos de soluções de computação.

Tipo	Serviços da AWS	Características principais
Sistema de arquivos compartilhado	Amazon FSx	Baseia-se nas soluções de computação mais recentes da AWS para oferecer compatibilidade com quatro sistemas de arquivos usados com frequência: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. A latência, o throughput e as IOPS do Amazon FSx variam de acordo com o sistema de arquivos e devem ser consideradas ao selecionar o sistema de arquivos certo para as necessidades da sua workload.
Armazenamento em bloco	Amazon Elastic Block Store (Amazon EBS)	Serviço de armazenamento em blocos fácil de usar, escalável e de alta performance projetado para o Amazon Elastic Compute Cloud (Amazon EC2). O Amazon EBS inclui armazenamento baseado em SSD para workloads transacionais de alto throughput e em HDD para workloads trabalho de alto throughput.

Tipo	Serviços da AWS	Características principais
Banco de dados relacional	Amazon Aurora, Amazon RDS, Amazon Redshift.	Projetados para oferecer compatibilidade com transações ACID (atomicidade, consistência, isolamento, durabilidade) e manter a integridade referencial e uma forte consistência de dados. Muitas aplicações tradicionais, planejamento de recursos empresariais (ERP), gerenciamento de relacionamentos com o cliente (CRM) e comércio eletrônico usam bancos de dados relacionais para armazenar os dados.
Banco de dados de chave-valor	Amazon DynamoDB	Otimizados para padrões de acesso comuns, normalmente visando armazenar e recuperar grandes volumes de dados. Aplicações web de alto tráfego, sistemas de comércio eletrônico e aplicações de jogos são os casos de uso habituais para bancos de dados de chave-valor.

Tipo	Serviços da AWS	Características principais
Banco de dados de documentos	Amazon DocumentDB	Projetado para armazenar dados semiestruturados, como documentos do tipo JSON. Esses bancos de dados ajudam os desenvolv edores a criar e atualizar rapidamente aplicações de gerenciamento de conteúdo, catálogos e perfis de usuário, por exemplo.
Banco de dados na memória	Amazon ElastiCache , Amazon MemoryDB for Redis	Usados para aplicações que exigem acesso em tempo real aos dados, latência mais baixa e throughput mais alto. É possível usar bancos de dados na memória para armazenamento em cache de aplicações, gerenciam ento de sessões, tabelas de classificação de jogos, arquivo de atributos de ML de baixa latência, sistema de mensagens de microsser viços e um mecanismo de streaming de alto throughput.

Tipo	Serviços da AWS	Características principais
Banco de dados de grafos	Amazon Neptune	Utilizado para aplicações que precisam navegar e consultar milhões de relacionamentos entre conjuntos de dados de grafos altamente conectado s com latência de milissegu ndos em grande escala. Muitas empresas usam bancos de dados gráficos para detecção de fraudes, redes sociais e mecanismos de recomendação.
Banco de dados de séries temporais	Amazon Timestream	Utilizado para coletar, sintetizar e gerar com eficiência insights de dados que mudam ao longo do tempo. Aplicações de IoT, DevOps e telemetria industria l podem utilizar bancos de dados de séries temporais.
Coluna ampla	Amazon Keyspaces (for Apache Cassandra)	Usa tabelas, linhas e colunas, mas ao contrário de um banco de dados relacional, os nomes e o formato das colunas podem variar de linha para linha na mesma tabela. Normalmente, você vê um repositório de coluna ampla em aplicações industriais de alta escala para manutenção de equipamentos, gerenciam ento de frotas e otimização de rotas.

Tipo	Serviços da AWS	Características principais
ledger	Amazon Quantum Ledger Database (Amazon QLDB)	Oferece uma autoridade centralizada e confiável para manter um registro escalável, imutável e criptograficamente verificável de transações para cada aplicação. Vemos os bancos de dados de livro-razão empregados em sistemas de registro, cadeia de suprimentos, inscrições e até mesmo transações bancárias.

- Se você estiver criando uma plataforma de dados, utilize a arquitetura de dados moderna na AWS para integrar seu data lake, data warehouse e datastores específicos.
- As principais questões que você precisa considerar ao escolher um datastore para sua workload são as seguintes:

Pergunta	Fatos a serem considerados
Como os dados são estruturados?	• Se os dados não forem estruturados, considere um armazenamento de objetos, como o Amazon S3, ou um banco de dados NoSQL, como o Amazon DocumentDB. • Para dados de valor-chave, considere o DynamoDB, o Amazon ElastiCache (Redis OSS) ou o Amazon MemoryDB.
Que nível de integridade referencial é necessário?	• Para restrições de chave estrangeira, bancos de dados relacionais como Amazon RDS e Aurora podem fornecer esse nível de integridade. • Normalmente, em um modelo de dados NoSQL, você desnormalizaria os dados

Pergunta	Fatos a serem considerados
	em um único documento ou coleção de documentos para serem recuperados em uma única solicitação em vez de unir documentos ou tabelas de diferentes locais.
A conformidade com ACID (atomicidade, consistência, isolamento, durabilidade) é necessária?	• Se as propriedades ACID associadas aos bancos de dados relacionais forem necessárias, pense em um banco de dados relacional, como o Amazon RDS e o Aurora. • Se uma consistência forte for necessária para o banco de dados NoSQL, você pode usar leituras altamente consistentes com o DynamoDB.
Como as necessidades de armazenamento serão alteradas ao longo do tempo? Como isso afeta a escalabilidade?	• Os bancos de dados sem servidor, como o DynamoDB e o Amazon Quantum Ledger Database (Amazon QLDB), serão escalados dinamicamente. • Os bancos de dados relacionais têm limites superiores em armazenamento provisionado e devem ser particionados horizontalmente usando mecanismos, como fragmentação, quando atingem esses limites.
Qual é a proporção de consultas de leitura em relação a consultas de gravação? O armazenamento em cache melhoraria a performance?	• Workloads com muitas operações de leitura poderão se beneficiar de uma camada de cache, como ElastiCache ou DAX, se o banco de dados for o DynamoDB. • As leituras também podem ser descarreg adas em réplicas de leitura com bancos de dados relacionais, como o Amazon RDS.

Pergunta	Fatos a serem considerados
O armazenamento e a modificação (OLTP – Processamento de transações on-line) ou a recuperação e a geração de relatórios (OLAP – Processamento analítico on-line) têm uma prioridade mais alta?	• Para um processamento transacional de throughput alto de leitura no estado em que se encontra, considere um banco de dados NoSQL, como o DynamoDB. • Para padrões de leitura complexos e de throughput alto (como junção) com consistência use o Amazon RDS. • Para consultas analíticas, considere usar um banco de dados colunar, como o Amazon Redshift, ou exportar os dados para o Amazon S3 e realizar análises usando o Athena ou o Amazon QuickSight.
Que nível de durabilidade os dados exigem?	• O Aurora replica automaticamente os dados entre três zonas de disponibilidade em uma região, o que significa que seus dados terão mais durabilidade com menos chance de serem perdidos. • O DynamoDB é automaticamente replicado entre várias zonas de disponibilidade, fornecendo alta disponibilidade e durabilid ade aos dados. • O Amazon S3 fornece 11 noves de durabilid ade. Muitos serviços de banco de dados, como o Amazon RDS e o DynamoDB, são compatíveis com a exportação de dados para o Amazon S3 para retenção de longo prazo e arquivamento.

Pergunta	Fatos a serem considerados
Você quer se livrar de mecanismos de bancos de dados comerciais ou custos de licenças?	• Considere usar mecanismos de código aberto, como PostgreSQL e MySQL no Amazon RDS ou Aurora. • Utilize o AWS Database Migration Service e o AWS Schema Conversion Tool para realizar migrações de mecanismos de bancos de dados comerciais para código aberto
Qual é a expectativa operacional para o banco de dados? A migração para serviços gerenciados é uma preocupação importante?	• Utilizar o Amazon RDS em vez do Amazon EC2 e o DynamoDB ou o Amazon DocumentDB em vez de um host automátic o de um banco de dados NoSQL pode reduzir a sobrecarga operacional.
Como o banco de dados é acessado no momento? Ele é acessado apenas por aplicações ou há usuários de inteligência de negócios (BI) e outras aplicações prontas para uso conectadas?	• Se houver dependências de ferramentas externas, talvez seja necessário manter a compatibilidade com os bancos de dados que elas suportam. O Amazon RDS é totalmente compatível com as diferentes versões do mecanismo a que ele oferece suporte, incluindo Microsoft SQL Server, Oracle, MySQL e PostgreSQL.

- Faça experimentos e testes comparativos em um ambiente de não produção para identificar qual datastore pode atender às necessidades da workload.

Recursos

Documentos relacionados:

- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx para Lustre

- Performance do Amazon FSx para Windows File Server
- Amazon S3 Glacier: documentação do S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Armazenamento na nuvem com a AWS
- Características de E/S do Amazon EBS
- Bancos de dados na nuvem com a AWS
- Cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Práticas recomendadas do Amazon DynamoDB
- Escolher entre o Amazon EC2 e o Amazon RDS
- Práticas recomendadas de implementação do Amazon ElastiCache

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a eficiência do Amazon Elastic Block Store e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon Simple Storage Service
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon Simple Storage Service
- AWS re:Invent 2022: Construir arquiteturas de dados modernos na AWS
- AWS re:Invent 2022: Construir arquiteturas de data mesh na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon Aurora e suas inovações
- AWS re:Invent 2023: Modelagem de dados com o Amazon DynamoDB
- AWS re:Invent 2022: Modernizar aplicações com bancos de dados com propósito específico
- Mergulho profundo no Amazon DynamoDB: padrões de design avançados (DAT403-R1)

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS
- Bancos de dados para desenvolvedores
- Dia de imersão na arquitetura de dados moderna na AWS
- Criar um data mesh na AWS
- Exemplos do Amazon S3
- Otimizar o padrão de dados usando o compartilhamento de dados do Amazon Redshift
- Migrações de bancos de dados
- MS SQL Server: demonstração da replicação do AWS Database Migration Service (AWS DMS)
- Workshop prático de modernização de bancos de dados
- Exemplos do Amazon Neptune

PERF03-BP02 Avaliar as opções de configuração disponíveis para o datastore

Entenda e avalie os vários atributos e opções de configuração disponíveis para seus datastores a fim de otimizar o espaço de armazenamento e a performance da workload.

Práticas comuns que devem ser evitadas:

- Você só usa um tipo de armazenamento, como o Amazon EBS, para todas as workloads.
- Você usa as IOPS provisionadas para todas as workloads sem testes reais em todos os níveis de armazenamento.
- Você não sabe quais são as opções de configuração da solução de gerenciamento de dados escolhida.
- Você conta somente com o aumento do tamanho da instância sem examinar outras opções de configuração.
- Você não testa as características de ajuste de escala do datastore.

Benefícios de implementar esta prática recomendada: a exploração e a experimentação das configurações de datastore permitem que você reduza o custo da infraestrutura, melhore a performance e diminua o esforço necessário para manter as workloads.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Uma workload pode ter um ou mais datastores usados com base nos requisitos de armazenamento e acesso aos dados. Para otimizar a eficiência de performance e custos, é necessário avaliar os padrões de acesso aos dados para determinar as configurações apropriadas do datastore. Ao explorar as opções de datastore, leve em consideração vários aspectos, como opções de armazenamento, memória, computação, réplica de leitura, requisitos de consistência, grupo de conexões e opções de armazenamento em cache. Experimente essas várias opções de configuração para melhorar as métricas de eficiência de performance.

Etapas de implementação

- Entenda as configurações atuais (como tipo de instância, tamanho do armazenamento ou versão do mecanismo de banco de dados) do datastore.
- Analise a documentação e as práticas recomendadas da AWS para saber mais sobre as opções de configuração indicadas que podem ajudar a melhorar a performance do datastore. As principais opções de datastore a serem consideradas são:

Opção de configuração	Exemplos
Descarregar leituras (como réplicas de leitura e cache)	• Nas tabelas do DynamoDB, é possível descarregar leituras usando o DAX para armazenamento em cache. • É possível criar um cluster do Amazon ElastiCache (Redis OSS) e configurar a aplicação para ler primeiro do cache e voltar para o banco de dados caso o item solicitad o não esteja presente. • Todos os bancos de dados relacionais, como Amazon RDS e Aurora, e bancos de dados NoSQL provisionados, como Neptune e Amazon DocumentDB, permitem adicionar réplicas de leitura para descarreg ar as partes de leitura da workload. • Os bancos de dados de tecnologia sem servidor, como o DynamoDB, ajustarão

Opção de configuração	Exemplos
	a escala automaticamente. Verifique se você tem unidades de capacidade de leitura (RCU) suficientes provisionadas para processar a workload.

Opção de configuração	Exemplos
Escalar gravações (como a fragmentação da chave da partição ou a introdução de uma fila)	• No caso de bancos de dados relaciona is, é possível aumentar o tamanho da instância para acomodar uma workload maior, ou aumentar as IOPs provisionadas para permitir um throughput mais alto no armazenamento subjacente. • Também é possível introduzir uma fila na frente do banco de dados, em vez de gravar diretamente nele. Esse padrão permite desacoplar a ingestão do banco de dados e controlar a taxa de fluxo para que o banco de dados não fique sobrecarregado. • Usar solicitações de gravação em lote em vez de criar muitas transações de curta duração pode ajudar a melhorar o throughput em bancos de dados relacionais de alto volume de gravação. • Os bancos de dados com tecnologia sem servidor, como o DynamoDB, podem ajustar a escala do throughput de gravação automaticamente ou ajustar as unidades da capacidade de gravação (WCU) provision adas, dependendo do modo da capacidad e. • Você ainda pode ter problemas com partições ativas ao atingir os limites de throughput de determinada chave de partição. Isso pode ser mitigado com a escolha de uma chave de partição mais uniformemente distribuída ou por meio da fragmentação da gravação da chave de partição.

Opção de configuração	Exemplos
Políticas para gerenciar o ciclo de vida dos seus conjuntos de dados	• O Amazon S3 Lifecycle pode ser usado para gerenciar seus objetos durante todo o ciclo de vida de cada um. Se os padrões de acesso forem desconhecidos, variáveis ou imprevisíveis, você poderá usar o Amazon S3 Intelligent-Tiering para monitorar os padrões de acesso e mover automaticamente os objetos que não foram acessados para níveis de acesso de baixo custo. Você pode aproveitar as métricas da Lente de Armazenamento do Amazon S3 para identificar oportunidades de otimizaçã o e lacunas no gerenciamento do ciclo de vida. • O gerenciamento do ciclo de vida do Amazon EFS gerencia automaticamente o armazenamento de arquivos para seus sistemas de arquivos.
Gerenciamento e agrupamento de conexões	• O Amazon RDS Proxy pode ser usado com o Amazon RDS e o Aurora para gerenciar as conexões com o banco de dados. • Os bancos de dados com tecnologia sem servidor, como o DynamoDB, não têm conexões associadas a eles, mas considere a capacidade provisionada e as políticas de ajuste de escala automático para lidar com picos na carga.

- Realize experimentos e testes comparativos em um ambiente de não produção para identificar qual opção de configuração pode atender aos requisitos da workload.
- Depois de experimentar, planeje a migração e valide as métricas de performance.

- Use ferramentas de monitoramento da AWS (como o Amazon CloudWatch) e de otimização (como a Lente de Armazenamento do Amazon S3) para otimizar constantemente o datastore usando um padrão de uso real.

Recursos

Documentos relacionados:

- Armazenamento na nuvem com a AWS
- Tipos de volume do Amazon EBS
- Armazenamento do Amazon EC2
- Amazon EFS: performance do Amazon EFS
- Performance do Amazon FSx para Lustre
- Performance do Amazon FSx para Windows File Server
- Amazon S3 Glacier: documentação do S3 Glacier
- Amazon S3: considerações sobre performance e taxa de solicitações
- Características de E/S do Amazon EBS
- Bancos de dados na nuvem com a AWS
- Cache de banco de dados da AWS
- DynamoDB Accelerator
- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Redshift Spectrum
- Práticas recomendadas do Amazon DynamoDB

Vídeos relacionados:

- AWS re:Invent 2023: Melhorar a eficiência do Amazon Elastic Block Store e ser mais econômico
- AWS re:Invent 2023: Otimizar o preço e a performance do armazenamento com o Amazon Simple Storage Service
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon Simple Storage Service
- AWS re:Invent 2023: Novidades do armazenamento de arquivos da AWS

- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS
- Bancos de dados para desenvolvedores
- Dia de imersão na arquitetura de dados moderna na AWS
- Ajuste de escala automático do Amazon EBS
- Exemplos do Amazon S3
- Exemplos do Amazon DynamoDB
- Exemplos de migração de banco de dados da AWS
- Workshop de modernização de bancos de dados
- Trabalhar com parâmetros no Amazon RDS para Postgress DB

PERF03-BP03 Coletar e registrar métricas de performance do datastore

Acompanhe e registre métricas de performance relevantes para o datastore a fim de entender a performance das suas soluções de gerenciamento de dados. Essas métricas podem ajudar você a otimizar o datastore, verificar se os requisitos da workload foram atendidos e fornecer uma visão geral clara da performance da workload.

Práticas comuns que devem ser evitadas:

- Você só usa a pesquisa manual de arquivos de log para métricas.
- Você só publica métricas em ferramentas internas usadas pela equipe e não tem uma imagem abrangente da workload.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.
- Você só monitora as métricas no sistema e não captura as métricas de uso e acesso aos dados.

Benefícios de implementar esta prática recomendada: o estabelecimento de uma linha de base de performance ajuda a compreender o comportamento normal e os requisitos das workloads. Padrões

anormais podem ser identificados e depurados mais rapidamente, melhorando a performance e a confiabilidade do datastore.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Para monitorar a performance dos datastores, é necessário registrar várias métricas de performance ao longo de um período. Isso permite detectar anomalias e avaliar a performance em relação às métricas de negócios para verificar se as necessidades da workload estão sendo atendidas.

As métricas devem incluir as do sistema subjacente que oferece suporte ao datastore e as do banco de dados. As métricas do sistema subjacente podem incluir métricas de utilização de CPU, memória, armazenamento em disco disponível, E/S de disco, taxa de acertos do cache e entrada e saída da rede, enquanto as métricas do datastore devem incluir transações por segundo, tempos de resposta, uso de índice, bloqueios de tabela, tempos limite de consultas e número de conexões abertas. Esses dados são essenciais para compreender a performance da workload e como a solução de gerenciamento de dados é usada. Use essas métricas como parte de uma abordagem orientada por dados para ajustar e otimizar os recursos da workload.

Use ferramentas, bibliotecas e sistemas que registram as medidas de performance relacionadas ao banco de dados.

Etapas de implementação

- Identifique as principais métricas de performance que o datastore deve monitorar.
  - Métricas e dimensões do Amazon S3
  - Monitorar métricas em uma instância do Amazon RDS
  - Monitorar a workload de banco de dados com o Performance Insights no Amazon RDS
  - Visão geral do monitoramento avançado
  - Métricas e dimensões do DynamoDB
  - Monitorar o DynamoDB Accelerator
  - Monitorar o Amazon MemoryDB com o Amazon CloudWatch
  - Que métricas devo monitorar?
  - Monitorar a performance do cluster do Amazon Redshift
  - Métricas e dimensões do Timestream

  - Métricas do Amazon CloudWatch para o Amazon Aurora
  - Registrar em log e monitorar o Amazon Keyspaces (para Apache Cassandra)
  - Monitorar recursos do Amazon Neptune
- Use uma solução aprovada de registro em log e monitoramento para coletar essas métricas. O Amazon CloudWatch pode coletar métricas nos recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas para descobrir métricas de negócio ou derivadas. Use o CloudWatch ou soluções de terceiros para definir alarmes que indicam quando os limites são violados.
- Confira se o monitoramento do datastore pode se beneficiar de uma solução de machine learning que detecta anomalias de performance.
  - O Amazon DevOps Guru para Amazon RDS fornece visibilidade dos problemas de performance e faz recomendações de ações corretivas.
- Configure a retenção de dados em sua solução de monitoramento e de log para corresponder às suas metas operacionais e de segurança.
  - Retenção de dados padrão para métricas do CloudWatch
  - Retenção de dados padrão para CloudWatch Logs

Recursos

Documentos relacionados:

- Cache de banco de dados da AWS
- As 10 melhores dicas de performance para Amazon Athena
- Práticas recomendadas do Amazon Aurora
- DynamoDB Accelerator
- Práticas recomendadas do Amazon DynamoDB
- Práticas recomendadas do Amazon Redshift Spectrum
- Performance do Amazon Redshift
- Bancos de dados na nuvem com a AWS
- Insights de Performance do Amazon RDS

Vídeos relacionados:

- AWS re:Invent 2022: Monitoramento de performance com o Amazon RDS e o Aurora, com destaque para Autodesk
- Monitoramento e ajuste de performance de banco de dados com o Amazon DevOps Guru para Amazon RDS
- AWS re:Invent 2023: Novidades do armazenamento de arquivos na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB
- AWS re:Invent 2023: Criar e otimizar data lakes no Amazon S3
- AWS re:Invent 2023: Novidades do armazenamento de arquivos na AWS
- AWS re:Invent 2023: Mergulho profundo no Amazon DynamoDB
- Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache

Exemplos relacionados:

- Framework de coleta de métricas de ingestão de conjunto de dados na AWS
- Workshop de monitoramento do Amazon RDS
- Workshop de bancos de dados com propósito específico na AWS

PERF03-BP04 Implementar estratégias para melhorar a performance da consulta no datastore

Implemente estratégias para otimizar os dados e melhorar a consulta de dados a fim de permitir mais escalabilidade e performance eficiente para a workload.

Práticas comuns que devem ser evitadas:

- Você não particiona dados no datastore.
- Você armazena dados em apenas um formato de arquivo no datastore.
- Você não usa índices no datastore.

Benefícios de implementar esta prática recomendada: a otimização da performance dos dados e das consultas ocasiona mais eficiência, menor custo e melhor experiência do usuário.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A otimização de dados e o ajuste de consultas são aspectos essenciais da eficiência de performance em um datastore, pois afetam não só a performance, mas também a capacidade de resposta de toda a workload na nuvem. Consultas não otimizadas podem ocasionar maior uso de recursos e gargalos, o que reduz a eficiência geral de um datastore.

A otimização de dados inclui várias técnicas para garantir o armazenamento e o acesso eficientes aos dados. Esse processo também ajuda a melhorar a performance da consulta em um datastore. As principais estratégias incluem particionamento, compactação e desnormalização de dados, o que ajuda a otimizá-los para armazenamento e acesso.

Etapas de implementação

- Entenda e analise as consultas de dados críticos que são realizadas no datastore.
- Identifique as consultas com execução lenta no datastore e use planos de consulta para entender o estado atual delas.
  - Analisar o plano de consulta no Amazon Redshift
  - Usar EXPLAIN e EXPLAIN ANALYZE no Athena
- Implemente estratégias para melhorar a performance da consulta. Algumas das principais estratégias incluem:
  - Usar um formato de arquivo colunar (como Parquet ou ORC).
  - Compactar os dados no datastore para reduzir o espaço de armazenamento e as operações de E/S.
  - Particionar os dados para dividi-los em partes menores e reduzir o tempo de verificação dos dados.
    - Particionamento de dados no Athena
    - Partições e distribuição de dados
  - Indexação de dados nas colunas comuns na consulta.
  - Use visões materializadas para consultas frequentes.
    - Entender as visões materializadas
    - Criar visões materializadas no Amazon Redshift
  - Escolha a operação de junção correta para consulta. Ao unir duas tabelas, especifique a tabela maior no lado esquerdo da junção e a tabela menor no lado direito.

  - Solução de cache distribuído para melhorar a latência e reduzir o número de operações de E/S do banco de dados.
  - Manutenção regular, como aspiração, reindexação e estatísticas de execução.
- Experimente e teste estratégias em um ambiente de não produção.

Recursos

Documentos relacionados:

- Práticas recomendadas do Amazon Aurora
- Performance do Amazon Redshift
- As 10 melhores dicas de performance para Amazon Athena
- Cache de banco de dados da AWS
- Práticas recomendadas para implementar o Amazon ElastiCache
- Particionamento de dados no Athena

Vídeos relacionados:

- AWS re:Invent 2023: Práticas recomendadas de otimização de custos de armazenamento na AWS
- AWS re:Invent 2022: Monitoramento de performance com o Amazon RDS e o Aurora, com destaque para Autodesk
- Otimizar consultas do Amazon Athena com novas ferramentas de análise de consultas

Exemplos relacionados:

- Workshop de bancos de dados com propósito específico na AWS

PERF03-BP05 Implementar padrões de acesso a dados que utilizam cache

Implemente padrões de acesso que possam se beneficiar do armazenamento em cache de dados para recuperação rápida de dados acessados com frequência.

Práticas comuns que devem ser evitadas:

- Armazenar em cache dados que mudam com frequência.
- Depender dos dados em cache como se estivessem armazenados de forma durável e sempre disponíveis.
- Não levar em conta a consistência dos seus dados em cache.
- Não monitorar a eficiência da sua implementação de cache.

Benefícios de implementar esta prática recomendada: armazenar dados em um cache pode melhorar a latência de leitura, o throughput de leitura, a experiência do usuário e a eficiência geral, além de reduzir custos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Um cache é um componente de software ou hardware destinado a armazenar dados para que futuras solicitações dos mesmos dados possam ser atendidas com maior rapidez e eficiência. Os dados armazenados em um cache podem ser reconstruídos se perdidos, repetindo um cálculo anterior ou obtendo-os de outro datastore.

O armazenamento de dados em cache pode ser uma das estratégias mais eficazes para melhorar a performance geral da aplicação e reduzir a carga sobre as fontes de dados primárias subjacentes. Os dados podem ser armazenados em vários níveis na aplicação, como dentro da aplicação e fazendo chamadas remotas, o que é conhecido como cache do lado do cliente, ou usando um serviço secundário rápido para armazenar os dados, conhecido como cache remoto.

Armazenamento em cache no lado do cliente

Com o armazenamento em cache no lado do cliente, cada cliente (uma aplicação ou serviço que consulta o datastore de backend) pode armazenar os resultados de suas consultas exclusivas localmente por um período especificado. Isso pode reduzir o número de solicitações na rede para um datastore ao verificar primeiro o cache do cliente local. Se os resultados não estiverem presentes, a aplicação poderá então consultar o datastore e armazenar esses resultados localmente. Esse padrão permite que cada cliente armazene dados no local mais próximo (o próprio cliente), resultando na menor latência possível. Os clientes também podem continuar a atender algumas consultas quando o datastore de backend não está disponível, aumentando a disponibilidade geral do sistema.

Uma desvantagem dessa abordagem é que, quando vários clientes estão envolvidos, eles podem armazenar os mesmos dados em cache localmente. Isso resulta no uso de armazenamento

duplicado e na inconsistência de dados entre esses clientes. Um cliente pode armazenar em cache os resultados de uma consulta e, um minuto depois, outro cliente pode executar a mesma consulta e obter um resultado diferente.

Armazenamento em cache remoto

Para resolver o problema de dados duplicados entre clientes, um serviço externo rápido ou cache remoto pode ser usado para armazenar os dados consultados. Em vez de verificar um datastore local, cada cliente verificará o cache remoto antes de consultar o datastore de backend. Essa estratégia permite respostas mais consistentes entre clientes, melhor eficiência nos dados armazenados e um volume maior de dados em cache, pois o espaço de armazenamento é dimensionado independentemente dos clientes.

A desvantagem de um cache remoto é que o sistema geral pode ter uma latência maior, pois é necessário um salto de rede adicional para verificar o cache remoto. O cache do lado do cliente pode ser usado junto com o armazenamento em cache remoto para o armazenamento em vários níveis para melhorar a latência.

Etapas de implementação

- Identifique bancos de dados, APIs e serviços de rede que poderiam se beneficiar do armazenamento em cache. Serviços que têm workloads de leitura pesadas, uma alta taxa de leitura e gravação ou que são caros para escalar são candidatos ao armazenamento em cache.
  - Armazenamento em cache de banco de dados
  - Habilitar o armazenamento em cache de APIs para melhorar a capacidade de resposta
- Identifique o tipo apropriado de estratégia de armazenamento em cache que melhor se adapte ao seu padrão de acesso.
  - Estratégias de armazenamento em cache
  - Soluções de armazenamento em cache da AWS
- Siga as práticas recomendadas de armazenamento em cache para seu datastore.
- Configure uma estratégia de invalidação de cache, como um time-to-live (TTL), para todos os dados que equilibre a atualização dos dados e reduza a pressão sobre o datastore de backend.
- Habilite recursos como novas tentativas automáticas de conexão, recuo exponencial, tempos limite no lado do cliente e pool de conexões no cliente, se disponíveis, pois eles podem melhorar a performance e a confiabilidade.
  - Práticas recomendadas: clientes Redis e Amazon ElastiCache (Redis OSS)

- Monitore a taxa de acertos de cache com uma meta de 80% ou mais. Valores mais baixos podem indicar tamanho insuficiente do cache ou um padrão de acesso que não se beneficia do armazenamento em cache.
  - Que métricas devo monitorar?
  - Práticas recomendadas para monitorar workloads do Redis no Amazon ElastiCache
  - Monitorar as práticas recomendadas com o Amazon ElastiCache (Redis OSS) usando o Amazon CloudWatch
- Implemente a replicação de dados para descarregar as leituras em várias instâncias e melhorar a performance e a disponibilidade da leitura de dados.

Recursos

Documentos relacionados:

- Usar a Lente do Well-Architected para o Amazon ElastiCache
- Monitorar as práticas recomendadas com o Amazon ElastiCache (Redis OSS) usando o Amazon CloudWatch
- Que métricas devo monitorar?
- Whitepaper Performance em grande escala com o Amazon ElastiCache
- Desafios e estratégias de armazenamento em cache

Vídeos relacionados:

- Plano de aprendizado do Amazon ElastiCache
- Design para o sucesso com as práticas recomendadas do Amazon ElastiCache
- AWS re:Invent 2020: Design para o sucesso com as práticas recomendadas do Amazon ElastiCache
- AWS re:Invent 2023 [LANÇAMENTO]: Introdução ao Amazon ElastiCache sem servidor
- AWS re:Invent 2022: Cinco excelentes formas de reimaginar sua camada de dados com o Redis
- AWS re:Invent 2021: Mergulho profundo no Amazon ElastiCache (Redis OSS)

Exemplos relacionados:

- Como aumentar a performance de bancos de dados MySQL com o Amazon ElastiCache (Redis OSS)

Rede e entrega de conteúdo

A solução de rede ideal para uma workload varia com base em latência, requisitos de throughput, jitter e largura de banda. Restrições físicas, como recursos de usuário ou on-premises, determinam as opções de localização. Essas restrições podem ser compensadas com locais de borda ou posicionamento de recursos.

Na AWS, as redes são virtualizadas e estão disponíveis em vários tipos e configurações diferentes. Desse modo, é mais fácil atender às suas necessidades de rede. A AWS oferece recursos de produtos (por exemplo, redes avançada, instâncias otimizadas de rede do Amazon EC2, aceleração de transferências do Amazon S3 e Amazon CloudFront dinâmico) para otimizar o tráfego da rede. A AWS também oferece recursos de rede (por exemplo, roteamento de latência do Amazon Route 53, endpoints da Amazon VPC, AWS Direct Connect e AWS Global Accelerator) para reduzir a distância ou o jitter da rede.

Essa área de foco compartilha orientações e práticas recomendadas para projetar, configurar e operar soluções eficientes de rede e entrega de conteúdo na nuvem.

Práticas recomendadas

- PERF04-BP01 Compreender como as redes afetam a performance
- PERF04-BP02 Avaliar os recursos de rede disponíveis
- PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload
- PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos
- PERF04-BP05 Escolher protocolos de rede para melhorar a performance
- PERF04-BP06 Escolher o local da workload com base nos requisitos de rede
- PERF04-BP07 Otimizar a configuração da rede com base em métricas

PERF04-BP01 Compreender como as redes afetam a performance

Analise e entenda como as decisões relacionadas à rede afetam sua workload para fornecer performance eficiente e uma melhor experiência do usuário.

Práticas comuns que devem ser evitadas:

- Todo o tráfego flui por meio dos data centers existentes.

- Você direciona todo o tráfego por meio de firewalls centrais em vez de usar ferramentas de segurança de rede nativas da nuvem.
- Você provisiona conexões do AWS Direct Connect sem entender os requisitos reais de uso.
- Você não considera as características da workload e a sobrecarga da criptografia ao definir suas soluções de redes.
- Você usa conceitos e estratégias de on-premises para soluções de redes na nuvem.

Benefícios de implementar esta prática recomendada: a compreensão de como as redes afetam a performance da workload ajuda a identificar gargalos potenciais, a melhorar a experiência dos usuários, a aumentar a confiabilidade e a reduzir a manutenção operacional à medida que a workload muda.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A rede é responsável pela conectividade entre os componentes da aplicação, os serviços de nuvem, as redes de borda e os dados on-premises. Portanto, ela pode afetar significativamente a performance da workload. Além da performance da workload, a experiência dos usuários também é afetada por latência de rede, largura de banda, protocolos, localização, congestão de rede, jitter, throughput e regras de roteamento.

Ter uma lista documentada dos requisitos de rede da workload, incluindo latência, tamanho de pacotes, regras de roteamento, protocolos e padrões de tráfego compatíveis. Analise as soluções de redes disponíveis e identifique os serviços que atendem às características de rede da sua workload. É possível recriar as redes baseadas na nuvem rapidamente. Portanto, é necessário evoluir sua arquitetura de rede ao longo do tempo para melhorar a eficiência da performance.

Etapas de implementação:

- Defina e documente os requisitos de performance da rede, incluindo métricas como latência da rede, largura de banda, protocolos, locais, padrões de tráfego (picos e frequência), throughput, criptografia, inspeção e regras de roteamento.
- Saiba mais sobre os principais serviços de rede da AWS, como VPCs, AWS Direct Connect, Elastic Load Balancing (ELB) e Amazon Route 53.
- Capture as seguintes características principais da rede:

Características	Ferramentas e métricas
Características básicas de rede	• Logs de fluxo da VPC • Logs de fluxo do AWS Transit Gateway • Métricas de AWS Transit Gateway do • Métricas de AWS PrivateLink do
Características da rede da aplicação	• Adaptador de malha elástica • Métricas de AWS App Mesh do • Métricas do Amazon API Gateway
Características da rede da borda	• Métricas do Amazon CloudFront • Métricas do Amazon Route 53 • Métricas de AWS Global Accelerator do
Características da rede híbrida	• Métricas de AWS Direct Connect do • Métricas de AWS Site-to-Site VPN do • Métricas de AWS Client VPN do • Métricas da WAN da Nuvem AWS
Características da rede de segurança	• Métricas do AWS Shield, AWS WAF e AWS Network Firewall
Características de rastreamento	• AWS X-Ray • VPC Reachability Analyzer • Analisador de Acesso à Rede • Amazon Inspector • Amazon CloudWatch RUM

- Teste comparativo e de performance da rede:
  - Faça o teste comparativo do throughput da rede, pois alguns fatores podem afetar a performance da rede do Amazon EC2 quando as instâncias estão na mesma VPC. Meça a largura de banda da rede entre as instâncias Linux do Amazon EC2 na mesma VPC.
  - Faça testes de carga para experimentar soluções e opções de redes.

Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Gateway de trânsito
- Passar para o roteamento baseado em latência no Amazon Route 53
- VPC Endpoints

Vídeos relacionados:

- AWS re:Invent 2023: Fundamentos de rede na AWS
- AWS re:Invent 2023: O que rede pode fazer por sua aplicação?
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2019: Otimizar a performance da rede para instâncias do Amazon EC2
- AWS Summit Online: Melhorar a performance de rede global para aplicações
- AWS re:Invent 2020: Dicas e práticas recomendadas de rede com o Well-Architected Framework
- AWS re:Invent 2020: Práticas recomendadas de rede na AWS em migrações de grande escala

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS
- Workshop prático de firewall de rede
- Observar e diagnosticar sua rede na AWS

- Como encontrar e lidar com configurações de rede incorretas na AWS

PERF04-BP02 Avaliar os recursos de rede disponíveis

Avalie recursos de rede na nuvem que possam melhorar a performance. Meça o impacto desses recursos por meio de testes, métricas e análises. Por exemplo, utilize os recursos de rede disponíveis para reduzir a latência, a distância ou o jitter da rede.

Práticas comuns que devem ser evitadas:

- Você permanece em uma região, pois é onde a sede da sua empresa ou organização está fisicamente localizada.
- Você usa firewalls em vez de grupos de segurança para filtrar o tráfego.
- Você quebra o TLS para inspeção de tráfego em vez de confiar em grupos de segurança, políticas de endpoint e outras funcionalidades nativas da nuvem.
- Você só usa segmentação baseada em sub-rede em vez de grupos de segurança.

Benefícios de implementar esta prática recomendada: avaliar todos os recursos e opções de serviços pode aumentar a performance da workload, reduzir o custo da infraestrutura, diminuir o esforço necessário para manter sua workload e aumentar sua postura geral de segurança. É possível utilizar o backbone global da AWS para garantir a experiência ideal de rede para os clientes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

A AWS oferece serviços como o AWS Global Accelerator e o Amazon CloudFront que podem ajudar a melhorar a performance da rede, enquanto a maioria dos serviços da AWS tem recursos de produto (como o recurso Amazon S3 Transfer Acceleration) para otimizar o tráfego da rede.

Analise quais opções de configuração de rede estão disponíveis e como elas poderiam afetar a workload. A otimização da performance depende da compreensão de como essas opções interagem com sua arquitetura e do impacto que elas terão na performance medida e na experiência do usuário.

Etapas de implementação

- Crie uma lista de componentes da workload.

  - Considere usar a WAN da Nuvem AWS para criar, gerenciar e monitorar a rede da sua organização ao criar uma rede global unificada.
  - Monitore suas redes globais e centrais com as métricas do Amazon CloudWatch Logs. Utilize o Amazon CloudWatch RUM, que fornece insights para ajudar a identificar, entender e aprimorar a experiência digital dos usuários.
  - Visualize a latência agregada da rede entre as Regiões da AWS e as zonas de disponibilidade, bem como dentro de cada zona de disponibilidade, usando o AWS Network Manager para obter informações sobre como a performance da aplicação se relaciona com a performance da rede da AWS subjacente.
  - Use uma ferramenta existente de banco de dados de gerenciamento de configuração (CMDB) ou um serviço como o AWS Config para criar um inventário da sua workload e de como ela está configurada.
- Se for uma workload existente, identifique e documente a referência para suas métricas de performance, focando os gargalos e nas áreas de melhoria. As métricas de rede associadas à performance irão variar de acordo com a workload com base nos requisitos comerciais e nas características da workload. Como ponto de partida, a análise dessas métricas pode ser importante para sua workload: largura de banda, latência, perda de pacotes, jitter e retransmissões.
- Se essa for uma nova workload, realize testes de carga para identificar gargalos de performance.
- Para os gargalos de performance que identificar, revise as opções de configuração para suas soluções a fim de identificar oportunidades de melhoria da performance. Confira os seguintes recursos de rede e opções importantes:

Oportunidade de melhoria	Solução
Caminho ou rotas de rede	Use o Analisador de Acesso à Rede para identificar caminhos ou rotas.
Protocolos de rede	Consulte PERF04-BP05 Escolher protocolos de rede para melhorar a performance
Topologia de rede	Avalie seus compromissos operacionais e de performance entre o emparelhamento de VPC e o AWS Transit Gateway ao conectar várias contas. O AWS Transit Gateway simplific a a forma como você interconecta todas as

Oportunidade de melhoria	Solução
	suas VPCs, as quais podem se estender por milhares de Contas da AWS e até redes on-premises. Compartilhe seu AWS Transit Gateway entre várias contas usando o AWS Resource Access Manager. Consulte PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload
Serviços de rede	O AWS Global Accelerator é um serviço de rede que melhora a performance do tráfego dos usuários em até 60% usando a infraestrutura de rede global da AWS. O Amazon CloudFront pode melhorar a performance da entrega e da latência de conteúdo da workload globalmente. Use o Lambda@Edge para executar funções que personalizam o conteúdo que o CloudFront entrega mais perto dos usuários, reduzir a latência e melhorar a performance. O Amazon Route 53 oferece opções de roteamento baseado em latência, roteamento por geolocalização, roteamento por geoproximidade e roteamento baseado em IP para ajudar você a melhorar a performance da sua workload para um público global. Identifique qual opção de roteamento otimizaria a performance da workload analisando o respectivo tráfego e a localização do usuário quando ela for distribuída globalmente.

Oportunidade de melhoria	Solução
Recursos do atributo de armazenamento	O Amazon S3 Transfer Acceleration é um recurso que permite aos usuários externos beneficiarem-se de otimizações de rede do CloudFront para o upload de dados para o Amazon S3. Isso melhora a capacidade de transferir grandes quantidades de dados com origem em locais remotos que não têm conectividade dedicada com a Nuvem AWS. O Amazon S3 Multi-Region Access Points replica conteúdo para várias regiões e simplifica a workload ao fornecer um ponto de acesso. Quando um ponto de acesso multirregiões é usado, você pode solicitar ou gravar dados no Amazon S3 com o serviço identificando o bucket de menor latência.

Oportunidade de melhoria	Solução
Atributos dos recursos computacionais	As interfaces de rede elásticas (ENI) usadas por instâncias do Amazon EC2, contêineres e funções do Lambda são limitadas por fluxo. Revise seus grupos de posicionamento para otimizar seu throughput de rede do EC2. Para evitar gargalos em uma abordagem por fluxo, projete sua aplicação para usar vários fluxos. Para monitorar e obter visibilidade de suas métricas de rede relacionadas à computaçã o, use o CloudWatch Metrics e a ethtool. O comando `ethtool` está incluído no driver da ENA e expõe métricas adicionais relacionadas à rede que podem ser publicadas como uma métrica personalizada no CloudWatch. Os adaptadores de rede elástica (ENA) da Amazon aumentam ainda mais a otimizaçã o oferecendo throughput melhor para suas instâncias em um grupo de posicionamento de cluster. O Elastic Fabric Adapter (EFA) é uma interface de rede para instâncias do Amazon EC2 que permite executar workloads que exigem altos níveis de comunicação entre nós em grande escala na AWS. As instâncias otimizadas para Amazon EBS usam uma pilha de configuração otimizada e fornecem capacidade adicional e dedicada para aumentar a E/S do Amazon EBS.

Recursos

Documentos relacionados:

- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Passar para o encaminhamento por latência no Amazon Route 53
- VPC Endpoints
- VPC Flow Logs

Vídeos relacionados:

- AWS re:Invent 2023: Pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2022: Mergulho profundo na infraestrutura de rede da AWS
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2018: Otimizar a performance da rede para instâncias do Amazon EC2
- AWS Global Accelerator

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS
- Observar e diagnosticar sua rede
- Como encontrar e lidar com configurações de rede incorretas na AWS

PERF04-BP03 Escolher a conectividade dedicada ou VPN apropriada para a workload

Quando a conectividade híbrida é necessária para conectar recursos on-premises e na nuvem, provisione a largura de banda adequada para atender aos requisitos de performance. Estime os requisitos de largura de banda e de latência para a workload híbrida. Esses números determinarão seus requisitos de dimensionamento.

Práticas comuns que devem ser evitadas:

- Avaliar somente as soluções de VPN para seus requisitos de criptografia de rede.
- Não avaliar as opções de backup ou de conectividade redundante.
- Não identificar todos os requisitos da workload (necessidades de criptografia, protocolo, largura de banda e tráfego).

Benefícios de implementar esta prática recomendada: selecionar e configurar soluções de conectividade apropriadas aumentará a confiabilidade da workload e maximizará a performance. A identificação dos requisitos da workload, o planejamento antecipado e a avaliação das soluções híbridas podem minimizar alterações dispendiosas da rede física e despesas operacionais, e aumentará seu tempo para geração de valor.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Desenvolva uma arquitetura de rede híbrida com base em seus requisitos de largura de banda. O AWS Direct Connect permite a você conectar sua rede on-premises de forma privada com a AWS. Isso é conveniente quando você precisa de largura de banda alta e baixa latência com performance consistente. Uma conexão VPN estabelece uma conexão segura via Internet. Ela é usada quando apenas uma conexão temporária é necessária, quando o custo é um fator ou como uma contingência enquanto se espera que uma conectividade de rede física resiliente seja estabelecida durante o uso do AWS Direct Connect.

Se seus requisitos de largura de banda forem altos, considere vários serviços do AWS Direct Connect ou de VPN. O tráfego pode ser balanceado entre os serviços, embora o balanceamento de carga entre o AWS Direct Connect e a VPN não seja recomendado devido às diferenças de latência e largura de banda.

Etapas de implementação

- Calcule os requisitos de largura de banda e latência de suas aplicações existentes.
  - Para workloads existentes que estão sendo migradas para a AWS, utilize os dados de seus sistemas de monitoramento de rede internos.
  - Para workloads novas ou existentes para as quais não há dados de monitoramento, consulte os proprietários do produto para determinar métricas de performance adequadas e fornecer uma experiência do usuário satisfatória.
- Escolha uma conexão dedicada ou VPN como sua opção de conectividade. Com base em todos os requisitos da workload (necessidades de criptografia, largura de banda e tráfego), é possível escolher o AWS Direct Connect ou o AWS VPN (ou ambos). O diagrama a seguir ajudará você a escolher o tipo de conexão apropriada.
  - O AWS Direct Connect fornece conectividade dedicada ao ambiente da AWS, de 50 Mbps a 100 Gbps, usando conexões dedicadas ou conexões hospedadas. Isso permite que você tenha latência gerenciada e controlada, além de largura de banda provisionada para que a workload possa se conectar de forma eficiente com outros ambientes. Com os parceiros do AWS Direct Connect, é possível ter conectividade completa para vários ambientes, fornecendo uma rede estendida com performance consistente. A AWS oferece ajuste de escala da largura de banda da conexão direta usando o grupo de agregação nativo (LAG) de 100 Gbps ou o BGP equal-cost multipath (ECMP).
  - A AWS Site-to-Site VPN fornece um serviço de VPN gerenciada compatível com o protocolo de segurança da internet (IPsec). Quando uma conexão VPN é criada, cada conexão VPN inclui dois túneis para alta disponibilidade.
- Siga a documentação da AWS para escolher uma opção de conectividade apropriada:
  - Se você decidir usar o AWS Direct Connect, selecione a largura de banda apropriada para sua conectividade.
  - Se você estiver usando um AWS Site-to-Site VPN em vários locais para se conectar a uma Região da AWS, use uma conexão do Site-to-Site VPN acelerada para ter a oportunidade de melhorar a performance da rede.
  - Se o design da sua rede consistir em uma conexão VPN IPsec via AWS Direct Connect, considere usar uma VPN IP privada para melhorar a segurança e obter segmentação. AWS Uma VPN IP privada site a site é implantada por meio da interface virtual de trânsito (VIF).
  - O AWS Direct Connect SiteLink permite criar conexões redundantes e de baixa latência entre seus datacenters em todo o mundo, enviando dados pelo caminho mais rápido entre locais do AWS Direct Connect, ignorando as Regiões da AWS.

- Valide sua configuração de conectividade antes de implantá-la na produção. Execute testes de segurança e performance para garantir que ela atenda aos requisitos de largura de banda, confiabilidade, latência e conformidade.
- Monitore regularmente a performance e o uso da conectividade e otimize, se necessário.

Fluxograma de performance determinística

Recursos

Documentos relacionados:

- Produtos de rede com a AWS
- AWS Transit Gateway
- Endpoints da VPC
- Construção de uma infraestrutura de rede da AWS Multi-VPC escalável e segura
- VPN do cliente

Vídeos relacionados:

- AWS re:Invent 2023: Criar conectividade de rede híbrida com a AWS
- AWS re:Invent 2023: Proteger a conectividade remota com a AWS
- AWS re:Invent 2022: Otimizar a performance com o Amazon CloudFront
- AWS re:Invent 2019: Conectividade à AWS e arquiteturas de rede na AWS híbridas
- AWS re:Invent 2020: Conectar ao AWS Transit Gateway

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

PERF04-BP04 Usar o balanceamento de carga para distribuir o tráfego em vários recursos

Distribua o tráfego entre vários recursos e serviços para permitir que sua workload aproveite a elasticidade oferecida pela nuvem. Também é possível usar o balanceamento de carga para descarregar a terminação de criptografia a fim de melhorar a performance, a confiabilidade e gerenciar e rotear o tráfego de maneira eficaz.

Práticas comuns que devem ser evitadas:

- Você não considera os requisitos da workload ao escolher o tipo de balanceador de carga.
- Você não utiliza os recursos do balanceador de carga para otimização da performance.
- A workload é exposta diretamente à internet sem um balanceador de carga.
- Você roteia todo o tráfego da Internet por meio de balanceadores de carga existentes.
- Você usa o balanceamento de carga TCP genérico e faz com que cada nó de computação lide com a criptografia SSL.

Benefícios de implementar esta prática recomendada: um balanceador de carga lida com a carga variável do tráfego da sua aplicação em uma única zona de disponibilidade ou em várias zonas de disponibilidade e permite alta disponibilidade, ajuste de escala automático e melhor utilização da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os balanceadores de carga atuam como o ponto de entrada para sua workload, ponto a partir do qual distribuem o tráfego para seus destinos de backend, como instâncias de computação ou contêineres, para melhorar a utilização.

Escolher o tipo certo de balanceador de carga é a primeira etapa para otimizar sua arquitetura. Comece listando as características da workload, como protocolo (como TCP, HTTP, TLS ou WebSockets), o tipo de destino (como instâncias, contêineres ou tecnologia sem servidor), requisitos da aplicação (como conexões de execução longa, autenticação de usuários ou adesão) e posicionamento (como região, zona local, Outpost ou isolamento por zona).

A AWS fornece vários modelos para que suas aplicações usem o balanceamento de carga. O Application Load Balancer é mais adequado para balanceamento de carga de tráfego HTTP e

HTTPS e oferece roteamento de solicitação avançado direcionado para a entrega de arquiteturas de aplicações modernas, incluindo microsserviços e contêineres.

O Network Load Balancer é mais adequado para o balanceamento de carga de tráfego TCP que exija performance extrema. Ele é capaz de processar milhões de solicitações por segundo enquanto mantém latências ultrabaixas, e também é otimizado para lidar com padrões de tráfego súbitos e voláteis.

O Elastic Load Balancing oferece gerenciamento integrado de certificados e descriptografia SSL/TLS, o que proporciona a flexibilidade de gerenciar centralmente as configurações SSL do load balancer e descarregar de sua workload as interações com uso intenso de CPU.

Após escolher o balanceador de carga certo, você poderá começar a utilizar seus recursos para reduzir a quantidade de esforço que seu backend precisa fazer para atender o tráfego.

Por exemplo, ao usar tanto o Application Load Balancer (ALB) quanto o Network Load Balancer (NLB), é possível realizar o descarregamento de criptografia SSL/TLS, que é uma oportunidade de evitar que o handshake TLS com uso intenso da CPU seja concluído pelos destinos e também melhorar o gerenciamento de certificados.

Ao configurar o descarregamento de SSL/TLS no balanceador de carga, ele se torna responsável pela criptografia do tráfego de e para os clientes enquanto entrega o tráfego não criptografado aos backends, liberando os recursos de backend e melhorando o tempo de resposta para os clientes.

O Application Load Balancer também pode fornecer tráfego HTTP/2 sem precisar acomodá-lo em seus destinos. Essa simples decisão pode melhorar o tempo de resposta da aplicação, já que o HTTP/2 usa conexões TCP de forma mais eficiente.

Os requisitos de latência da workload devem ser considerados ao definir a arquitetura. Como exemplo, se você tiver uma aplicação sensível à latência, poderá decidir usar o Network Load Balancer, que oferece latências extremamente baixas. Como alternativa, você pode decidir aproximar a workload dos clientes utilizando o Application Load Balancer em zonas locais da AWS ou mesmo no AWS Outposts.

Outra consideração para workloads sensíveis à latência é o balanceamento de carga entre zonas. Com o balanceamento de carga entre zonas, cada nó do balanceador de carga distribui o tráfego entre os destinos registrados em todas as Zonas de Disponibilidade habilitadas.

Use o Auto Scaling integrado ao balanceador de carga. Um dos principais aspectos de um sistema com performance eficiente está relacionado ao dimensionamento correto dos recursos de backend.

Para fazer isso, é possível utilizar as integrações do balanceador de carga para os recursos de destino de backend. Ao usar a integração do balanceador de carga com os grupos do Auto Scaling, os destinos serão adicionados ou removidos do balanceador de carga conforme exigido em resposta ao tráfego recebido. Os balanceadores de carga também podem ser integrados ao Amazon ECS e ao Amazon EKS para workloads em contêineres.

- Amazon ECS: balanceamento de carga do serviço
- Application Load Balancer no Amazon EKS
- Network Load Balancer no Amazon EKS

Etapas de implementação

- Defina seus requisitos de balanceamento de carga, incluindo volume de tráfego, disponibilidade e escalabilidade de aplicações.
- Escolha o tipo certo de balanceador de carga para sua aplicação.
  - Use o Application Load Balancer para workloads HTTP/HTTPS.
  - Use o Network Load Balancer para workloads não HTTP executadas em TCP ou UDP.
  - Use uma combinação de ambos (ALB como destino do NLB) se quiser aproveitar os recursos de ambos os produtos. Por exemplo, é possível fazer isso se você quiser usar os IPs estáticos do NLB junto com o roteamento baseado em cabeçalho HTTP do ALB, ou se quiser expor a workload HTTP em um AWS PrivateLink.
  - Para ver uma comparação completa dos balanceadores de carga, consulte a comparação de produtos do ELB.
- Use o descarregamento de SSL/TLS, se possível.
  - Configure receptores HTTPS/TLS com o Application Load Balancer e o Network Load Balancer integrados ao AWS Certificate Manager.
  - Observe que algumas workloads podem exigir criptografia completa por motivos de conformidade. Nesse caso, é um requisito para permitir a criptografia nos destinos.
  - Para conhecer as práticas recomendadas de segurança, consulte SEC09-BP02 Aplicar criptografia em trânsito.
- Escolha o algoritmo de roteamento certo (apenas ALB).
  - O algoritmo de roteamento pode fazer a diferença em como os destinos de backend são bem utilizados e, portanto, na forma como afetam a performance. Por exemplo, o ALB fornece duas opções para algoritmos de roteamento:

  - Solicitações menos pendentes: use para obter uma melhor distribuição de carga para seus destinos de backend em casos nos quais as solicitações para a aplicação variam em complexidade ou os destinos variam na capacidade de processamento.
  - Round robin: use quando as solicitações e os destinos forem semelhantes, ou se você precisar distribuir as solicitações igualmente entre os destinos.
- Considere isolamento por zona ou entre zonas.
  - Desative a opção entre zonas (isolamento por zona) para melhorias de latência e domínios com falha de zona. Ela é desativada por padrão no NLB e é possível desativá-la por grupo-alvo no ALB.
  - Ative a opção entre zonas para maior disponibilidade e flexibilidade. Ela é ativada por padrão no ALB e é possível ativá-la por grupo-alvo no NLB.
- Ative as manutenções de funcionamento de HTTP para as workloads HTTP (apenas ALB). Com esse recurso, o balanceador de carga pode reutilizar as conexões de backend até expirar o tempo limite da manutenção de funcionamento, melhorando a solicitação HTTP e o tempo de resposta, além de reduzir a utilização de recursos nos destinos de backend. Para obter detalhes sobre como fazer isso para o Apache e o Nginx, consulte Quais são as configurações ideais para usar o Apache ou o NGINX como servidor de backend para o ELB?
- Ative o monitoramento do balanceador de carga.
  - Ative os logs de acesso para seu Application Load Balancer e Network Load Balancer.
  - Os principais campos a considerar para o ALB são `request_processing_time`, `request_processing_time` e `response_processing_time`.
  - Os principais campos a considerar para o NLB são `connection_time` e `tls_handshake_time`.
  - Esteja pronto para consultar os logs quando precisar deles. É possível usar o Amazon Athena para consultar tanto os logs do ALB quanto os logs do NLB.
  - Crie alarmes para métricas relacionadas à performance, como `TargetResponseTime` para ALB.

Recursos

Documentos relacionados:

- Comparação de produtos de ELB
- Infraestrutura global da AWS

- Melhorar a performance e reduzir os custos usando a afinidade de zona de disponibilidade
- Passo a passo para a análise de logs com o Amazon Athena
- Consultar logs do Application Load Balancer
- Monitorar seus Application Load Balancers
- Monitorar os Network Load Balancers
- Usar o Elastic Load Balancing para distribuir tráfego entre as instâncias no grupo do Auto Scaling

Vídeos relacionados:

- AWS re:Invent 2023: O que rede pode fazer por sua aplicação?
- AWS re:Inforce 20: Como usar o Elastic Load Balancing para melhorar seu procedimento de segurança em escala
- AWS re:Invent 2018: Mergulho profundo e práticas recomendadas do Elastic Load Balancing
- AWS re:Invent 2021: Como escolher o balanceador de carga certo para suas workloads da AWS
- AWS re:Invent 2019: Como aproveitar ao máximo o Elastic Load Balancing para diferentes workloads

Exemplos relacionados:

- Gateway Load Balancer
- CDK e exemplos do AWS CloudFormation para análise de logs com o Amazon Athena

PERF04-BP05 Escolher protocolos de rede para melhorar a performance

Tome decisões sobre protocolos de comunicação entre sistemas e redes com base no impacto na performance da workload.

Há uma relação entre latência e largura de banda para alcançar o throughput. Por exemplo, se a transferência de arquivos estiver usando TCP, latências mais altas provavelmente reduzirão o throughput geral. Existem abordagens para corrigir isso com ajuste de TCP e protocolos de transferência otimizados, mas uma solução é usar o protocolo UDP.

Práticas comuns que devem ser evitadas:

- Você usa TCP para todas as workloads, independentemente dos requisitos de performance.

Benefícios de implementar esta prática recomendada: verificar se um protocolo apropriado é usado para comunicação entre usuários e componentes da workload ajuda a melhorar a experiência geral do usuário para as aplicações. Por exemplo, o UDP sem conexão permite alta velocidade, mas não oferece retransmissão ou alta confiabilidade. O TCP é um protocolo completo, mas requer maior sobrecarga para processar os pacotes.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Se você puder escolher protocolos diferentes para sua aplicação e tiver experiência nessa área, otimize sua aplicação e a experiência do usuário final usando um protocolo diferente. Observe que essa abordagem apresenta dificuldades significativas e só deve ser experimentada se você tiver otimizado sua aplicação de outras maneiras primeiro.

Uma consideração primária para melhorar a performance da workload é entender os requisitos de latência e throughput e escolher os protocolos de rede que otimizam essa performance.

Quando considerar o uso do TCP

O TCP oferece entrega de dados confiável e pode ser usado para comunicação entre componentes da workload em que a confiabilidade e a entrega garantida de dados é importante. Muitas aplicações baseadas na Web dependem de protocolos baseados em TCP, como HTTP e HTTPS, para abrir soquetes TCP para comunicação entre componentes da aplicação. As transferências de dados por e-mail e arquivo são aplicações comuns que também usam o TCP, pois é um mecanismo de transferência simples e confiável entre componentes de aplicações. Usar o TLS com TCP pode adicionar sobrecarga à comunicação, o que pode resultar em maior latência e redução de throughput, mas traz a vantagem da segurança. A sobrecarga vem principalmente da sobrecarga adicionada do processo de handshake, que pode exigir várias idas e voltas para ser concluído. Quando o handshake for concluído, a sobrecarga da criptografia e descriptografia de dados será relativamente pequena.

Quando considerar o uso do UDP

O UDP é um protocolo sem conexão e, portanto, é adequado para aplicações que precisam de uma transmissão rápida e eficiente, como log, monitoramento e dados de VoIP. Além disso, considere usar o UDP se você tiver componentes da workload que respondam a pequenas consultas de grandes números de clientes para garantir a performance ideal da workload. O Datagram Transport

Layer Security (DTLS) é o equivalente UDP do Transport Layer Security (TLS). Ao usar DTLS com UDP, a sobrecarga vem da criptografia e descriptografia de dados, já que o processo de handshake é simplificado. O DTLS também adiciona uma pequena quantidade de sobrecarga aos pacotes de UDP, já que inclui campos adicionais para indicar os parâmetros de segurança e detectar violações.

Quando considerar o uso do SRD

O SRD (datagrama confiável escalável) é um protocolo de transporte de rede otimizado para workloads de alto throughput devido à sua capacidade de fazer o balanceamento de carga do tráfego em vários caminhos e de se recuperar rapidamente de quedas de pacote ou falhas no link. Assim, o SRD é melhor nos casos de workloads de computação de alta performance (HPC) que exigem comunicação de alto throughput e baixa latência entre os nós de computação. Isso pode incluir tarefas de processamento paralelas, como simulação, modelagem e análise de dados que envolvem uma grande quantidade de transferência de dados entre os nós.

Etapas de implementação

- Use os serviços AWS Global Accelerator e AWS Transfer Family para melhorar o throughput de suas aplicações de transferência de arquivos online. O serviço AWS Global Accelerator ajuda você a obter baixa latência entre os dispositivos cliente e a workload na AWS. Com o AWS Transfer Family, é possível usar protocolos baseados em TCP, como SFTP e FTPS, para escalar e gerenciar com segurança as transferências de arquivos para os serviços de armazenamento da AWS.
- Use a latência de rede para determinar se o TCP é adequado para comunicação entre os componentes da workload. Se a latência de rede entre a aplicação cliente e o servidor for alta, o handshake de três vias do TCP pode levar um tempo, afetando, assim, a capacidade de resposta da aplicação. Métricas como tempo até o primeiro byte (TTFB) e tempo de ida e volta (RTT) podem ser usadas para medir a latência da rede. Se sua workload serve conteúdo dinâmico para os usuários, considere usar o Amazon CloudFront, que estabelece uma conexão persistente com cada origem de conteúdo dinâmico para remover o tempo de configuração da conexão que, de outra forma, diminuiria a velocidade de cada solicitação do cliente.
- Usar TLS com TCP ou UDP pode resultar em maior latência e menor throughput para a workload devido ao impacto da criptografia e descriptografia. Para workloads desse tipo, considere usar o descarregamento de SSL/TLS no Elastic Load Balancing para melhorar a performance da workload, permitindo que o balanceador de carga lide com o processo de criptografia e descriptografia de SSL/TLS em vez de deixar que as instâncias de backend façam isso. Isso pode ajudar a reduzir a utilização da CPU nas instâncias de backend, o que pode melhorar a performance e aumentar a capacidade.

- Use o Network Load Balancer (NLB) para implantar serviços que dependem do protocolo UDP, como autenticação e autorização, registro em log, DNS, IoT e mídia de streaming, visando melhorar a performance e a confiabilidade da workload. O NLB distribui o tráfego de UDP de entrada em vários destinos, permitindo escalar a workload horizontalmente, aumentar a capacidade e reduzir a sobrecarga de um único destino.
- Para suas workloads de computação de alta performance (HPC), considere usar a funcionalidade de Adaptador de Rede Elástica (ENA) Express, que usa o protocolo SRD para melhorar a performance da rede, fornecendo uma maior largura de banda de fluxo único (25 Gbps) e menor latência final (99,9 percentil) para tráfego de rede entre instâncias do EC2.
- Use o Application Load Balancer (ALB) para rotear e balancear a carga do tráfego de gRPC (Chamadas de procedimento remoto) entre os componentes da workload ou entre os serviços e clientes com gRPC habilitadas. As gRPC usam o protocolo HTTP/2 baseado em TCP para transporte e oferece benefícios de performance, como pegada de rede mais leve, compactação, serialização binária eficiente, suporte para várias linguagens e streaming bidirecional.

Recursos

Documentos relacionados:

- Como rotear tráfego UDP para o Kubernetes
- Application Load Balancer
- Rede avançada do EC2 no Linux
- Rede avançada do EC2 no Windows
- Grupos de posicionamento do EC2
- Como habilitar a rede avançada com o Adaptador de Rede Elástica (ENA) em instâncias Linux
- Network Load Balancer
- Produtos de redes com a AWS
- Passar para o encaminhamento por latência no Amazon Route 53
- VPC Endpoints

Vídeos relacionados:

- AWS re:Invent 2022: Escalar a performance da rede em instâncias do Amazon Elastic Compute Cloud de última geração

- AWS re:Invent 2022: Fundamentos de rede das aplicações

Exemplos relacionados:

- AWS Transit Gateway e soluções de segurança escaláveis
- Workshops de redes da AWS

PERF04-BP06 Escolher o local da workload com base nos requisitos de rede

Avalie as opções para o posicionamento de recursos visando reduzir a latência da rede e melhorar o throughput, proporcionando uma ótima experiência do usuário ao reduzir os tempos de carregamento da página e de transferência de dados.

Práticas comuns que devem ser evitadas:

- Consolidar todos os recursos da workload em uma única localização geográfica.
- Escolher a região mais próxima ao seu local, mas não ao usuário final da workload.

Benefícios de implementar esta prática recomendada: a experiência do usuário é muito afetada pela latência entre o usuário e sua aplicação. Ao usar Regiões da AWS adequadas e a rede global privada da AWS, é possível reduzir a latência e oferecer uma melhor experiência aos usuários remotos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Recursos, como instâncias do Amazon EC2, são colocados em zonas de disponibilidade em Regiões da AWS, zonas locais da AWS, AWS Outposts ou zonas do AWS Wavelength. A escolha desse local influencia o throughput e a latência da rede de determinado local do usuário. Serviços de borda como o Amazon CloudFront e o AWS Global Accelerator também podem ser usados para melhorar a performance da rede, seja armazenando o conteúdo em cache nos locais da borda ou oferecendo aos usuários um ótimo caminho para a workload por meio da rede global da AWS.

O Amazon EC2 oferece grupos de posicionamento para redes. Um grupo de posicionamento é um agrupamento lógico de instâncias para diminuir a latência. O uso de grupos de posicionamento com

tipos de instância compatíveis e um Adaptador de Rede Elástica (ENA) permite que as workloads participem de uma rede de baixa latência e com jitter reduzido e de 25 Gbps. Recomenda-se o uso de grupos de posicionamento para workloads que se beneficiam de baixa latência de rede, alto throughput de rede ou ambos.

Serviços sensíveis à latência são fornecidos em locais de borda usando uma rede global da AWS, como o Amazon CloudFront. Esses locais de borda costumam oferecer serviços, como rede de entrega de conteúdo (CDN) e sistema de nomes de domínio (DNS). Ao ter esses serviços na borda, as workloads podem responder com baixa latência a solicitações de conteúdo ou resolução de DNS. Esses serviços também fornecem serviços geográficos, como direcionamento geográfico de conteúdo (fornecendo conteúdo diferente conforme o local do usuário final) ou encaminhamento por latência para direcionar os usuários finais à região mais próxima (latência mínima).

Use serviços de borda para reduzir a latência e possibilitar o armazenamento do conteúdo em cache. Configure corretamente o controle de cache para DNS e HTTP/HTTPS a fim de aproveitar ao máximo essas abordagens.

Etapas de implementação

- Capture informações sobre o tráfego IP que entra e sai das interfaces de rede.
  - Como registrar tráfego IP em log com Logs de fluxo da VPC
  - Como o endereço IP do cliente é preservado no AWS Global Accelerator
- Analise os padrões de acesso à rede em sua workload para identificar como os usuários utilizam sua aplicação.
  - Use ferramentas de monitoramento, como o Amazon CloudWatch e o AWS CloudTrail, para coletar dados sobre atividades de rede.
  - Analise os dados para identificar o padrão de acesso à rede.
- Selecione as Regiões para implantação da workload com base nos seguintes elementos fundamentais:
  - Onde seus dados estão localizados: para aplicações com uso intenso de dados (como big data e machine learning), o código da aplicação deve ser executado o mais perto possível dos dados.
  - Onde seus usuários estão localizados: para aplicações voltadas ao usuário, escolha uma ou mais regiões perto dos clientes da workload.
  - Outras restrições: considere restrições como custo e conformidade, conforme explicado em O que considerar ao selecionar uma região para suas workloads.

- Use zonas locais da AWS para executar workloads como renderização de vídeo. As zonas locais permitem que você se beneficie de ter recursos de computação e armazenamento mais próximos dos usuários finais.
- Use o AWS Outposts para workloads que precisam permanecer on-premises e onde você deseja que essa workload seja executada ininterruptamente com o restante de suas workloads na AWS.
- Aplicações como streaming de vídeo ao vivo em alta resolução, áudio de alta fidelidade ou realidade aumentada/realidade virtual (RA/RV) exigem latência ultrabaixa para dispositivos 5G. Para aplicações desse tipo, considere o AWS Wavelength. O AWS Wavelength incorpora serviços de armazenamento e computação da AWS em redes 5G, fornecendo a infraestrutura móvel de computação de borda para desenvolver, implantar e escalar aplicações de latência ultrabaixa.
- Use o armazenamento em cache local ou soluções de armazenamento em cache da AWS para dados usados com frequência a fim de aumentar a performance, reduzir a movimentação de dados e diminuir o impacto ambiental.

Serviço	Quando usar
Amazon CloudFront	Use para armazenar conteúdo estático em cache, como imagens, scripts e vídeos, além de conteúdo dinâmico como respostas de API ou aplicações Web.
Amazon ElastiCache	Use para armazenar conteúdo em cache para aplicações Web.
DynamoDB Accelerator	Use para adicionar aceleração na memória às suas tabelas do DynamoDB.

- Use serviços que podem ajudar você a executar código mais perto dos usuários da workload, como a seguir:

Serviço	Quando usar
Lambda@Edge	Use para operações com uso intenso de computação que são iniciadas quando objetos não estão no cache.

Serviço	Quando usar
Amazon CloudFront Functions	Use para casos de uso simples, como solicitações HTTP(s) ou manipulações de resposta que podem ser iniciadas por funções de curta duração.
AWS IoT Greengrass	Use para executar computação local, mensagens e armazenamento de dados em cache para dispositivos conectados.

- Algumas aplicações exigem pontos de entrada fixos ou maior performance ao reduzir o jitter e a latência de primeiro byte, além de aumentar o throughput. Essas aplicações podem se beneficiar de serviços de rede que fornecem endereços IP anycast estáticos e terminação TCP em locais da borda. O AWS Global Accelerator pode melhorar a performance das suas aplicações em até 60% e fornecer failover rápido para arquiteturas multirregiões. O AWS Global Accelerator fornece endereços IP anycast estáticos que servem como um ponto de entrada fixo para suas aplicações hospedadas em uma ou mais Regiões da AWS. Esses endereços IP permitem que o tráfego entre na rede global da AWS o mais próximo possível dos usuários. O AWS Global Accelerator reduz o tempo de configuração da conexão inicial ao estabelecer uma conexão TCP entre o cliente e o local da borda da AWS mais próximo ao cliente. Analise o uso do AWS Global Accelerator para melhorar a performance das workloads de TCP/UDP e forneça failover rápido para arquiteturas de várias Regiões.

Recursos

Práticas recomendadas relacionadas:

- COST07-BP02 Implementar regiões com base nos custos
- COST08-BP03 Implementar serviços para reduzir custos de transferência de dados
- REL10-BP01 Implantar a workload em vários locais
- REL10-BP02 Escolher os locais apropriados para sua implantação de vários locais
- SUS01-BP01 Escolher a região com base nos requisitos empresariais e nas metas de sustentabilidade
- SUS02-BP04 Otimizar o posicionamento geográfico das workloads com base nos respectivos requisitos de rede

- SUS04-BP07 Minimizar a movimentação de dados entre redes

Documentos relacionados:

- Infraestrutura global da AWS
- Zonas locais da AWS e AWS Outposts: como escolher a tecnologia certa para sua workload de borda)
- Grupos de posicionamento
- Zonas locais da AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vídeos relacionados:

- Vídeo explicativo de zonas locais da AWS
- Visão geral do AWS Outposts e como ele funciona
- AWS re:Invent 2023: Uma estratégia de migração para workloads periféricas e on-premises
- AWS re:Invent 2021: AWS Outposts: como trazer a experiência da AWS para ambientes on-premises
- AWS re:Invent 2020: AWS Wavelength: executar aplicações com latência ultrabaixa na borda 5G
- AWS re:Invent 2022: Zonas locais da AWS: como criar aplicações para uma borda distribuída
- AWS re:Invent 2021: Criar sites de baixa latência com o Amazon CloudFront
- AWS re:Invent 2022: Aprimorar a performance e a disponibilidade com o AWS Global Accelerator
- AWS re:Invent 2022: Criar sua rede de longa distância usando a AWS
- AWS re:Invent 2020: Gerenciamento de tráfego global com o Amazon Route 53

Exemplos relacionados:

- Workshop de roteamento personalizado no AWS Global Accelerator
- Como lidar com reescritas e redirecionamentos usando funções da borda

PERF04-BP07 Otimizar a configuração da rede com base em métricas

Use dados coletados e analisados para tomar decisões bem informadas sobre a otimização da configuração da rede.

Práticas comuns que devem ser evitadas:

- Pressupor que todos os problemas relacionados à performance são relacionados à aplicação.
- Testar a performance da rede a partir de um local próximo ao local em que a workload foi implantada.
- Usar configurações-padrão para todos os serviços de rede.
- Provisionar em excesso recursos de rede para fornecer capacidade suficiente.

Benefícios de implementar esta prática recomendada: coletar as métricas necessárias da rede da AWS e implementar ferramentas de monitoramento de rede permite entender a performance da rede e otimizar as respectivas configurações.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Monitorar o tráfego de entrada e saída das VPCs, sub-redes ou interfaces de rede é fundamental para entender como utilizar os recursos de rede da AWS e otimizar as configurações da rede. Ao usar as ferramentas de rede da AWS a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs.

Etapas de implementação

- Identifique as principais métricas de performance, como latência ou perda de pacotes. A AWS fornece diversas ferramentas que podem ajudar você a coletar essas métricas. Ao usar as ferramentas a seguir, é possível verificar mais informações sobre o uso do tráfego, o acesso à rede e os logs:

Ferramenta AWS	Onde usar
Gerenciador de endereços IP da Amazon VPC.	Use o IPAM para planejar, rastrear e monitorar endereços IP para workloads da AWS e on-premises. Essa é uma prática recomendada para otimizar o uso e a alocação de endereços IP.
Logs de fluxo da VPC	Use os Logs de fluxo da VPC para obter informações detalhadas sobre o tráfego de entrada e saída das interfaces de rede nas VPCs. Com os logs de fluxo da VPC, é possível diagnosticar regras extremamente restritivas ou permissivas do grupo de segurança e determinar a direção do tráfego de entrada e saída das interfaces de rede.
Logs de fluxo do AWS Transit Gateway	Use os logs de fluxo do AWS Transit Gateway para capturar informações sobre o tráfego IP que entra e sai dos seus gateways de trânsito.
Registro em log de consultas ao DNS	Registre informações sobre consultas ao DNS, públicas ou privadas recebidas pelo Route 53. Com os logs de DNS, é possível otimizar as configurações de DNS entendendo o domínio ou subdomínio solicitado ou os locais da borda do Route 53 que responderam às consultas ao DNS.

Ferramenta AWS	Onde usar
Reachability Analyzer	O Reachability Analyzer ajuda a analisar e depurar a acessibilidade da rede. O Reachability Analyzer é uma ferramenta de análise de configuração que permite realizar testes de conectividade entre um recurso de origem e um recurso de destino em suas VPCs. Essa ferramenta ajuda a verificar se a configuração da rede corresponde à conectividade pretendida.
Analisador de Acesso à Rede	O Analisador de Acesso à Rede ajuda a entender o acesso via rede aos seus recursos. O Analisador de Acesso à Rede pode ser usado para especificar os requisitos de acesso à rede e identificar possíveis caminhos de rede que não atendam aos requisitos especificados. Ao otimizar a configuração da rede correspondente, é possível entender e verificar o estado da rede e demonstrar se a rede na AWS atende aos seus requisitos de conformidade.
Amazon CloudWatch	Use o Amazon CloudWatch e ative as métricas apropriadas para as opções de rede. Escolha a métrica de rede certa para sua workload. Por exemplo, é possível habilitar métricas para o uso do endereço de rede da VPC, o gateway NAT da VPC, o AWS Transit Gateway, o túnel da VPN, o AWS Network Firewall, o Elastic Load Balancing e o AWS Direct Connect. Monitorar continuamente as métricas é uma prática recomendada para observar e entender o status e o uso da rede, o que ajuda a otimizar a configuração da rede com base em suas observações.

Ferramenta AWS	Onde usar
AWS Network Manager	Com o AWS Network Manager, é possível monitorar a performance histórica e em tempo real da Rede Global da AWS para fins operacionais e de planejamento. O Gerenciador de Rede fornece latência de rede agregada entre as Regiões da AWS e as zonas de disponibilidade e dentro de cada zona de disponibilidade, permitindo que você entenda melhor como a performance da sua aplicação se relaciona à performance da rede da AWS subjacente.
Amazon CloudWatch RUM	Use o Amazon CloudWatch RUM para coletar as métricas que fornecem os insights que ajudam a identificar, entender e melhorar a experiência do usuário.

- Identifique os principais interlocutores e os padrões de tráfego de aplicações usando VPC e logs de fluxo do AWS Transit Gateway.
- Avalie e otimize sua arquitetura de rede atual, incluindo VPCs, sub-redes e roteamento. Como exemplo, você pode avaliar como diferentes emparelhamentos de VPC ou AWS Transit Gateway podem ajudar a melhorar a rede em sua arquitetura.
- Avalie os caminhos de roteamento em sua rede para verificar se o caminho mais curto entre os destinos é sempre usado. O Analisador de Acesso à Rede pode ajudar a fazer isso.

Recursos

Documentos relacionados:

- Log de consultas ao DNS público
- O que é IPAM?
- O que é o Reachability Analyzer?
- O que é o Analisador de Acesso à Rede?
- Métricas do CloudWatch para suas VPCs

- Otimizar a performance e reduzir os custos de análise de rede com os Logs de fluxo da VPC no formato Apache Parquet
- Monitorar suas redes global e básica com métricas do Amazon CloudWatch
- Monitorar continuamente o tráfego e os recursos da rede

Vídeos relacionados:

- AWS re:Invent 2023: Um guia do desenvolvedor para redes na nuvem
- AWS re:Invent 2023: Pronto para o que vem a seguir? Desenvolver redes para crescimento e flexibilidade
- AWS re:Invent 2023: Designs e novos recursos de VPCs avançadas
- AWS re:Invent 2022: Mergulho profundo na infraestrutura de rede da AWS
- AWS re:Invent 2020: Dicas e práticas recomendadas de rede com o AWS Well-Architected Framework
- AWS re:Invent 2020: Monitorar e solucionar problemas de tráfego de rede

Exemplos relacionados:

- Workshops de redes da AWS
- Monitoramento de rede da AWS
- Observar e diagnosticar sua rede na AWS
- Como encontrar e lidar com configurações de rede incorretas na AWS

Processo e cultura

Ao arquitetar workloads, há princípios e práticas que você pode adotar para ajudar na melhor execução de workloads na nuvem eficientes e de alta performance. Essa área de foco oferece as práticas recomendadas para ajudar a adotar uma cultura que promova a eficiência de performance das workloads na nuvem.

Considere estes princípios fundamentais para construir essa cultura:

- Infraestrutura como código: defina sua infraestrutura como código usando abordagens como modelos do AWS CloudFormation CloudFormation. O uso de modelos permite colocar a infraestrutura no controle de origem junto com o código e as configurações de sua aplicação. Isso permite aplicar à sua infraestrutura as mesmas práticas usadas para desenvolver software, possibilitando uma iteração rápida.
- Pipeline de implantação: use um pipeline de integração e implantação contínuas (CI/CD) (por exemplo, repositório de código-fonte, sistemas de compilação, implantação e automação de teste) para implantar sua infraestrutura. Isso permite a você implantar de maneira repetível, consistente e econômica enquanto itera.
- Métricas bem-definidas: configure e monitore métricas para capturar os indicadores-chave de performance (KPIs). Recomendamos usar tanto de métricas técnicas quanto de negócios. Para aplicações móveis ou sites, métricas importantes são a captura do tempo até o primeiro byte ou renderização. Outras métricas geralmente aplicáveis incluem contagem de threads, taxa de coleta de resíduos e estados de espera. As métricas de negócios, como o custo cumulativo agregado por solicitação, podem alertar sobre maneiras de reduzir os custos. Considere com cuidado como você planeja interpretar as métricas. Por exemplo, você poderia escolher o máximo ou o 99º percentil, em vez da média.
- Teste a performance automaticamente:como parte do processo de implantação, inicie automaticamente os testes de performance após a aprovação bem-sucedida nos testes de execução mais rápida. A automação deve criar um novo ambiente, configurar as condições iniciais, como dados de teste, e então executar uma série de testes comparativos e de carga. Os resultados desses testes então devem ser vinculados de volta à compilação para que você possa rastrear as mudanças de performance ao longo do tempo. Para testes de execução longa, você pode tornar essa parte do pipeline assíncrona em relação ao restante da compilação. Como alternativa, é possível realizar testes de performance durante a noite usando instâncias spot do Amazon EC2.

- Geração de carga: você deve criar uma série de scripts de teste que repliquem jornadas sintéticas ou pré-gravadas do usuário. Esses scripts devem ser idempotentes e não acoplados, e talvez você precise incluir scripts de pré-aquecimento para gerar resultados válidos. Seus scripts de teste devem replicar o máximo possível o comportamento do uso na produção. É possível usar soluções de software ou software como serviço (SaaS) para gerar a carga. Cogite o uso de soluções do AWS Marketplace e de instâncias spot: elas podem ser maneiras econômicas de gerar a carga.
- Visibilidade da performance: as métricas principais devem estar visíveis para a sua equipe, especialmente as métricas relacionadas a cada versão de compilação. Isso permite que você identifique qualquer tendência positiva ou negativa importante ao longo do tempo. Você também deve exibir métricas do número de erros ou exceções para garantir que esteja testando um sistema em funcionamento.
- Visualização: use técnicas de visualização que deixem claro onde os problemas de performance, hot spots, estados de espera ou baixa utilização estão ocorrendo. Sobreponha métricas de performance a diagramas de arquitetura: código ou gráficos de chamada podem ajudar a identificar problemas rapidamente.
- Revise os processos regularmente: arquiteturas com baixa performance geralmente são o resultado de um processo de análise de performance inexistente ou problemático. Se sua arquitetura está funcionando mal, a implementação de um processo de análise de performance permite promover melhorias iterativas.
- Otimização contínua: adote uma cultura para otimizar continuamente a eficiência de performance da workload na nuvem.

Práticas recomendadas

- PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload
- PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica
- PERF05-BP03 Definir um processo para melhorar a performance da workload
- PERF05-BP04 Fazer o teste de carga da workload
- PERF05-BP05 Usar automação para corrigir proativamente problemas relacionados a performance
- PERF05-BP06 Manter a workload e os serviços atualizados
- PERF05-BP07 Revisar as métricas regularmente

PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload

Identifique os KPIs que medem a performance da workload de forma quantitativa e qualitativa. Os KPIs ajudam você a medir a integridade e a performance de uma workload relacionada a uma meta empresarial.

Práticas comuns que devem ser evitadas:

- Monitorar as métricas somente no nível do sistema para obter informações da workload e não compreende aos impactos dessas métricas nos negócios.
- Pressupor que os KPIs já estejam publicados e compartilhados como dados de métricas comuns.
- Não definir um KPI quantitativo e mensurável.
- Não alinhar os KPIs às metas ou estratégias empresariais.

Benefícios de implementar esta prática recomendada: identificar KPIs específicos que representam a integridade e a performance da workload ajuda a alinhar as equipes em suas prioridades e a definir resultados empresariais bem-sucedidos. O compartilhamento dessas métricas com todos os departamentos fornece visibilidade e alinhamento dos limites, das expectativas e do impacto nos negócios.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Os KPIs permitem que as empresas e as equipes de engenharia alinhem a medição das metas e estratégias de como esses fatores são combinados para produzir resultados comerciais. Por exemplo, a workload de um site pode usar o tempo de carregamento da página como uma indicação da performance geral. Essa métrica seria um dos vários pontos de dados que medem a experiência do usuário. Além de identificar os limites do tempo de carregamento da página, documente o resultado esperado ou o risco da empresa se a performance ideal não for atingida. Um longo tempo de carregamento da página afeta diretamente os usuários finais, diminui a classificação da experiência do usuário e pode resultar em perda de clientes. Ao definir os limites dos KPIs, combine os testes comparativos do setor e as expectativas dos usuários finais. Por exemplo, se o teste comparativo do setor aplicável for o carregamento de uma página da Web em dois segundos, mas os usuários finais esperarem que uma página da Web seja carregada em um segundo, você deverá pensar nos dois pontos de dados ao estabelecer o KPI.

Sua equipe deve avaliar os KPIs da workload usando dados detalhados em tempo real e dados históricos para referência e criar painéis que calculem as métricas nos dados de KPI para derivar informações operacionais e de utilização. Os KPIs devem ser documentados e incluir limites que apoiem as metas e estratégias empresariais, bem como mapeados de acordo com as métricas que estão sendo monitoradas. Os KPIs devem ser revisitados quando as metas e as estratégias da empresa ou os requisitos dos usuários finais mudam.

Etapas de implementação

- Identifique as partes interessadas: identifique e documente as principais partes interessadas da empresa, incluindo as equipes de desenvolvimento e operações.
- Defina objetivos: trabalhe com essas partes interessadas para definir e documentar os objetivos da workload. Considere os aspectos críticos de performance das workloads, como throughput, tempo de resposta e custo, bem como as metas de negócios, como a satisfação dos usuários.
- Revise as práticas recomendadas do setor: revise as práticas recomendadas do setor para identificar KPIs relevantes alinhados aos objetivos da workload.
- Identifique métricas: identifique métricas que estejam alinhadas aos objetivos da sua workload e possam ajudar a medir a performance e as metas de negócios. Estabeleça KPIs com base nessas métricas. Exemplos de métricas são tempo médio de resposta, número de usuários simultâneos, entre outras.
- Defina e documente KPIs: use as práticas recomendadas do setor e os objetivos da workload para definir metas de KPI da workload. Use essas informações para definir limites de KPI no nível de gravidade ou de alarme. Identifique e documente o risco e o impacto no caso de um KPI não ser atendido.
- Implemente monitoramento: use ferramentas de monitoramento como o Amazon CloudWatch ou o AWS Config para coletar métricas e medir KPIs.
- Comunique visualmente os KPIs: use ferramentas de painel como o Amazon QuickSight para visualizar e comunicar os KPIs com as partes interessadas.
- Analise e otimize: revise e analise regularmente as métricas para identificar áreas da workload que precisam ser aprimoradas. Trabalhe com as partes interessadas para implementar essas melhorias.
- Revise e refine: revise regularmente as métricas e os KPIs para avaliar sua eficácia, especialmente quando as metas de negócios ou a performance da workload mudam.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- AWS Partners de monitoramento, registro em log e performance
- Ferramentas de observabilidade da AWS
- A importância dos indicadores-chave de performance (KPIs) para migrações para a nuvem em grande escala
- Como rastrear KPIs de otimização de custos com o painel de KPI
- Documentação do X-Ray
- Usar painéis do Amazon CloudWatch
- KPIs do QuickSight

Vídeos relacionados:

- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2023: Gerenciar eventos do ciclo de vida dos recursos em grande escala com o AWS Health
- AWS re:Invent 2023: Performance e eficiência no Pinterest: otimizando as instâncias mais recentes
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2023: Escalar na AWS para seus primeiros 10 milhões de usuários
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- Criar uma estratégia de métricas eficaz para sua empresa | Eventos da AWS

Exemplos relacionados:

- Criar um painel com o QuickSight

PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica

Entenda e identifique áreas em que aumentar a performance de sua workload causará um impacto positivo sobre a eficiência ou a experiência do cliente. Por exemplo, um site que tenha muita interação com o cliente se beneficiaria do uso de serviços de borda para aproximar a entrega de conteúdo dos clientes.

Práticas comuns que devem ser evitadas:

- Você pressupõe que as métricas de computação padrão, como utilização de CPU ou pressão de memória, são suficientes para detectar problemas de performance.
- Você só usa as métricas comuns registradas pelo software de monitoramento selecionado.
- Você só revisa as métricas quando há um problema.

Benefícios de implementar esta prática recomendada: compreender áreas críticas de performance ajuda os proprietários de workloads a monitorar KPIs e priorizar melhorias de alto impacto.

Nível de risco exposto se esta prática recomendada não for estabelecida: Alto

Orientação para implementação

Configure um rastreamento completo para identificar padrões de tráfego, latência e áreas de performance críticas. Monitore os padrões de acesso aos dados para consultas lentas ou dados particionados e fragmentados incorretamente. Identifique as áreas de restrição da workload usando o teste ou monitoramento de carga.

Aumente a eficiência de performance entendendo sua arquitetura, os padrões de tráfego e os padrões de acesso aos dados, além de identificar os tempos de latência e processamento. Identifique possíveis gargalos que possam afetar a experiência do cliente com o crescimento da workload. Depois de investigar essas áreas, veja qual solução você pode implantar para eliminar esses problemas de performance.

Etapas de implementação

- Configure um monitoramento completo para capturar todos os componentes e as métricas da workload. Aqui estão alguns exemplos de soluções de monitoramento na AWS.

Serviço	Onde usar
Amazon CloudWatch Real-User Monitoring (RUM)	Para capturar as métricas de performance da aplicação de sessões de frontend e do lado do cliente de usuários reais.
AWS X-Ray	Para monitorar o tráfego por meio das camadas de aplicação e identificar a latência entre componentes e dependências. Use os mapas do serviço X-Ray para ver os relaciona mentos e a latência entre os componentes da workload.
Insights de performance do Amazon Relationa l Database Service	Para ver as métricas de performance do banco de dados e identificar melhorias de performance.
Monitoramento avançado do Amazon RDS	Para ver métricas de performance do SO do banco de dados.
Amazon DevOps Guru	Para detectar padrões operacionais anormais a fim de identificar problemas operacionais antes que eles afetem os clientes.

- Realize testes para gerar métricas, identificar padrões de tráfego, gargalos e áreas de performance críticas. Aqui estão alguns exemplos de como realizar testes:
  - Configure os CloudWatch Synthetic Canaries para imitar programaticamente as atividades do usuário baseadas no navegador usando trabalhos cron do Linux ou expressões rate para gerar métricas consistentes ao longo do tempo.
  - Use a solução AWS Distributed Load Testing para gerar tráfego de pico ou testar a workload na taxa de crescimento esperada.
- Avalie as métricas e a telemetria para identificar as áreas de performance críticas. Avalie essas áreas com sua equipe para discutir sobre o monitoramento e as soluções visando evitar gargalos.
- Experimente com melhorias de performance e meça essas alterações com dados. Como exemplo, você pode usar o CloudWatch Evidently para testar novas melhorias e impactos de performance em sua workload.

Recursos

Documentos relacionados:

- Novidades no AWS Observability na re:Invent 2023
- Amazon Builders' Library
- Documentação do X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vídeos relacionados:

- AWS re:Invent 2023: [LANÇAMENTO] Monitoramento de aplicações para workloads modernas
- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2022: Amazon Builders' Library: 25 anos de excelência operacional da Amazon
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- Monitoramento visual de aplicações com o Amazon CloudWatch Synthetics

Exemplos relacionados:

- Medir o tempo de carregamento da com o Amazon CloudWatch Synthetics
- Cliente Web do Amazon CloudWatch RUM
- X-Ray SDK para Python
- Teste de carga distribuída na AWS

PERF05-BP03 Definir um processo para melhorar a performance da workload

Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações à medida que eles se tornam disponíveis. Por exemplo, execute testes de performance existentes em novas ofertas de instância para determinar o potencial delas de aprimorar sua workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você apresenta alterações de arquitetura ao longo do tempo sem justificativa de métrica.

Benefícios de implementar esta prática recomendada: ao definir seu processo para fazer alterações de arquitetura, é possível usar os dados coletados para influenciar o projeto da workload ao longo do tempo.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

A performance da sua workload tem algumas restrições importantes. Guarde essas restrições para saber que tipos de inovação podem aumentar a performance da sua workload. Use essas informações enquanto estiver aprendendo sobre novos serviços ou tecnologias à medida que surgem e identificar maneiras de reduzir restrições ou gargalos.

Identifique as principais restrições de performance da workload. Documente suas restrições de performance da workload para que você saiba quais tipos de inovação podem aprimorar a performance da workload.

Etapas de implementação

- Identifique os KPIs: identifique os KPIs de performance da workload conforme descrito em PERF05-BP01 Estabelecer indicadores-chave de performance (KPIs) para medir a integridade e a performance da workload para definir sua workload.
- Implemente monitoramento: use ferramentas de observabilidade da AWS para coletar métricas de performance e medir KPIs.

- Analise: faça uma análise aprofundada para identificar as áreas (como configuração e código da aplicação) na workload que apresentam baixa performance, conforme descrito em PERF05-BP02 Usar soluções de monitoramento para entender as áreas em que a performance é mais crítica. Use suas ferramentas de análise e performance para identificar as estratégias de melhoria de performance.
- Valide as melhorias: use ambientes de sandbox ou de pré-produção para validar a eficácia das estratégias de aperfeiçoamento.
- Implemente mudanças: implemente as mudanças na produção e monitore constantemente a performance da workload. Documente as melhorias e comunique as mudanças às partes interessadas.
- Revise e refine: revise regularmente seu processo de melhoria de performance para identificar áreas a serem aprimoradas.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS
- AWS Skill Builder

Vídeos relacionados:

- AWS re:Invent 2022: Como entregar arquiteturas sustentáveis e de alta performance
- AWS re:Invent 2023: Otimizar custos e performance e rastrear o progresso rumo à mitigação
- AWS re:Invent 2022: Otimização da AWS: etapas acionáveis para resultados imediatos
- AWS re:Invent 2022: Otimize suas workloads da AWS com a orientação de práticas recomendadas

Exemplos relacionados:

- GitHub da AWS

PERF05-BP04 Fazer o teste de carga da workload

Teste sua workload para verificar se ela pode lidar com a carga de produção e identificar qualquer gargalo de performance.

Práticas comuns que devem ser evitadas:

- Você faz um teste de carga de partes individuais da workload, mas não de toda ela.
- Você faz um teste de carga em uma infraestrutura que não é igual ao seu ambiente de produção.
- Você só faz testes de carga para a carga esperada, mas para nada além dela, para ajudar a prever onde pode haver problemas futuros.
- Você faz testes de carga sem consultar a política de testes do Amazon EC2 e enviar um formulário de envio de eventos simulados. Isso faz com que o teste não seja executado, pois parece um evento de negação de serviço.

Benefícios de implementar esta prática recomendada: medir sua performance em um teste de carga mostrará onde você será afetado à medida que a carga aumentar. Com isso você terá a capacidade de antecipar as alterações necessárias antes que elas afetem sua workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

O teste de carga na nuvem é um processo para medir a performance da workload na nuvem em condições realistas com a carga esperada do usuário. Esse processo envolve o provisionamento de um ambiente de nuvem semelhante ao de produção, o uso de ferramentas de teste de carga para gerar carga e a análise de métricas para avaliar a capacidade da workload de lidar com cargas realistas. Execute os testes de carga usando versões sintéticas ou limpas dos dados de produção (remova informações confidenciais ou de identificação). Realize testes de carga automaticamente como parte de seu pipeline de entrega e compare os resultados a KPIs e limites predefinidos. Esse processo ajuda você a continuar alcançando a performance necessária.

Etapas de implementação

- Defina seus objetivos de teste: identifique os aspectos de performance da workload que você deseja avaliar, como throughput e tempo de resposta.
- Selecione uma ferramenta de teste: escolha e configure a ferramenta de teste de carga adequada à workload.

- Configure seu ambiente: configure o ambiente de teste com base no ambiente de produção. É possível usar os serviços da AWS para executar ambientes em escala de produção para testar a arquitetura.
- Implemente o monitoramento: use ferramentas de monitoramento como o Amazon CloudWatch para coletar métricas dos recursos em sua arquitetura. Você também pode coletar e publicar métricas personalizadas.
- Defina cenários: defina os cenários e parâmetros do teste de carga (como duração do teste e número de usuários).
- Faça testes de carga: realize cenários de teste em grande escala. Aproveite a Nuvem AWS para testar a workload e descobrir se há uma falha na escala ou se ela está com a escala reduzida horizontalmente de maneira não linear. Por exemplo, use instâncias spot para gerar cargas a um baixo custo e descobrir gargalos antes que eles ocorram em produção.
- Analise os resultados do teste: analise os resultados para identificar gargalos de performance e áreas para melhorias.
- Documente e compartilhe descobertas: documente e relate as descobertas e recomendações. Compartilhe essas informações com as partes interessadas para ajudá-las a tomar decisões embasadas sobre estratégias de otimização da performance.
- Faça iterações contínuas: o teste de carga deve ser realizado regularmente, especialmente após uma alteração ou atualização do sistema.

Recursos

Documentos relacionados:

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Teste de carga distribuída na AWS

Vídeos relacionados:

- AWS Summit ANZ 2023: Acelere com confiança com o teste de carga distribuída da AWS
- AWS re:Invent 2022: Escalar na AWS para seus primeiros 10 milhões de usuários
- Resolver com soluções da AWS: teste de carga distribuída

- AWS re:Invent 2021: Otimize aplicações com base em insights do usuário final com o Amazon CloudWatch RUM
- Demonstração do Amazon CloudWatch Synthetics

Exemplos relacionados:

- Teste de carga distribuída na AWS

PERF05-BP05 Usar automação para corrigir proativamente problemas relacionados a performance

Use indicadores-chave de performance (KPIs), aliados a sistemas de monitoramento e alerta, para abordar proativamente problemas relacionados à performance.

Práticas comuns que devem ser evitadas:

- Você só permite que a equipe de operações faça alterações operacionais na workload.
- Você permite todos os filtros de alarmes para a equipe de operações, sem correção proativa.

Benefícios de implementar esta prática recomendada: a correção proativa de ações de alarme permite que a equipe de suporte se concentre nos itens que não são acionáveis automaticamente. Isso ajuda a equipe de operações a lidar com todos os alarmes sem ficar sobrecarregada e, em vez disso, se concentrar apenas nos alarmes críticos.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Sempre que possível, use alarmes para desencadear ações automatizadas visando corrigir problemas. Se a resposta automatizada não for possível, encaminhe o alarme para aqueles capazes de responder. Por exemplo, você pode ter um sistema capaz de prever os valores de indicadores-chave de performance (KPI) esperados e emitir um alarme quando eles ultrapassarem determinados limites, ou uma ferramenta capaz de interromper ou reverter automaticamente as implantações caso os KPIs estejam fora dos valores esperados.

Implemente processos que deem visibilidade à performance à medida que a workload estiver sendo executada. Para determinar se a performance da workload é ideal, crie painéis de monitoramento e estabeleça normas de linha de base para as expectativas de performance.

Etapas de implementação

- Identifique o fluxo de trabalho de correção: identifique e compreenda o problema de performance que pode ser corrigido automaticamente. Use soluções de monitoramento da AWS como o Amazon CloudWatch ou o AWS X-Ray para obter ajuda para entender melhor a causa-raiz do problema.
- Defina o processo de automação: crie um plano e um processo de correção detalhados que possam ser usados para corrigir automaticamente o problema.
- Configure o evento de iniciação: configure o evento para iniciar automaticamente o processo de correção. Por exemplo, você pode definir um acionador para reiniciar automaticamente uma instância quando ela atinge determinado limite de utilização da CPU.
- Automatize a correção: use serviços e tecnologias da AWS para automatizar o processo de correção. Por exemplo, o AWS Systems Manager Automation fornece uma maneira segura e escalável de automatizar o processo de correção. Use a lógica de autocorreção para reverter as alterações se elas não conseguirem resolver o problema.
- Teste o fluxo de trabalho: teste o processo de correção automatizado em um ambiente de pré-produção.
- Implemente o fluxo de trabalho: implemente a correção automatizada no ambiente de produção.
- Desenvolva um playbook: desenvolva e documente um playbook que descreva as etapas do plano de correção, incluindo os eventos de iniciação, a lógica de correção e as ações tomadas. Treine as partes interessadas para ajudá-las a responder com eficácia aos eventos de correção automatizada.
- Revise e refine: avalie regularmente a eficácia do fluxo de trabalho automatizado de correção. Ajuste os eventos de iniciação e a lógica de correção, se necessário.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Parceiros de monitoramento, log e performance da AWS Partner Network

- Documentação do X-Ray
- Usar alarmes e ações de alarme no CloudWatch
- Criar uma prática de automação de nuvem para excelência operacional: práticas recomendadas do AWS Managed Services
- Automatizar o ajuste de performance do Amazon Redshift com a otimização automática de tabelas

Vídeos relacionados:

- AWS re:Invent 2023: Estratégias para escalação automatizada, correção e autocorreção inteligente
- AWS re:Invent 2023: [LANÇAMENTO] Monitoramento de aplicações para workloads modernas
- AWS re:Invent 2023: Como implementar a observabilidade de aplicações
- AWS re:Invent 2021: Automatizar de forma inteligente as operações na nuvem
- AWS re:Invent 2022: Configurar controles em escala em seu ambiente da AWS
- AWS re:Invent 2022: Automatizar o gerenciamento e a conformidade de patches usando a AWS
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- AWS re:Invent 2023: Relaxe: diagnostique e resolva problemas de performance com o Amazon RDS
- AWS re:Invent 2021: {Novo lançamento} Detecte e resolva problemas automaticamente com o Amazon DevOps Guru
- AWS re:Invent 2023: Centralize suas operações

Exemplos relacionados:

- O CloudWatch Logs personaliza alarmes

PERF05-BP06 Manter a workload e os serviços atualizados

Fique em dia com os novos serviços e atributos de nuvem para adotar recursos eficientes, remover problemas e melhorar a eficiência geral da performance da workload.

Práticas comuns que devem ser evitadas:

- Você pressupõe que sua arquitetura atual é estática e não será atualizada ao longo do tempo.
- Você não tem nenhum sistema ou ritmo regular para avaliar se software ou pacotes atualizados são compatíveis com sua workload.

Benefícios de implementar esta prática recomendada: ao estabelecer um processo para se atualizar sobre novos serviços e ofertas, você pode adotar novos atributos e recursos, resolver problemas e melhorar a performance da workload.

Nível de risco exposto se esta prática recomendada não for estabelecida: Baixo

Orientação para implementação

Avalie maneiras de melhorar a performance à medida que novos serviços, padrões de design e atributos de produtos são disponibilizados. Determine quais deles poderiam aprimorar a performance ou aumentar a eficiência da workload por meio de avaliações, discussões internas ou análises externas. Defina um processo para avaliar atualizações, novos recursos e serviços relevantes para sua workload. Por exemplo, crie uma prova de conceito que use novas tecnologias ou consulte um grupo interno. Ao testar novas ideias ou serviços, faça testes de performance para medir o impacto causado por eles na performance da workload.

Etapas de implementação

- Faça o inventário da workload: faça o inventário de software e arquitetura da workload e identifique os componentes que precisam ser atualizados.
- Identifique fontes de atualizações: identifique novidades e atualize fontes relacionadas aos componentes da workload. Como exemplo, você pode assinar Novidades no blog da AWS para ver os produtos que correspondem ao componente da sua workload. Você pode assinar o feed RSS ou gerenciar suas assinaturas de e-mail.
- Defina um cronograma de atualizações: defina um cronograma para avaliar novos serviços e atributos para a workload.
  - É possível usar o AWS Systems Manager Inventory para coletar metadados de sistema operacional (SO), aplicação e instância das instâncias do Amazon EC2 e entender rapidamente quais instâncias executam o software e as configurações exigidas pela política de software e quais instâncias precisam ser atualizadas.
- Avalie a nova atualização: entenda como atualizar os componentes da sua workload. Aproveite a agilidade da nuvem para testar rapidamente como novos recursos podem melhorar a workload com o intuito de obter eficiência de performance.

- Use automação: use automação no processo de atualização para reduzir o nível de esforço para implantar novos recursos e limitar erros causados por processos manuais.
  - É possível usar CI/CD para atualizar automaticamente AMIs, imagens de contêiner e outros artefatos relacionados à aplicação de nuvem.
  - Você pode usar ferramentas como o AWS Systems Manager Patch Manager para automatizar o processo de atualizações do sistema e agendar a atividade usando as Janelas de Manutenção do AWS Systems Manager.
- Documente o processo: documente seu processo para avaliar atualizações e novos serviços. Forneça aos proprietários o tempo e o espaço necessários para pesquisar, testar, experimentar e validar atualizações e novos serviços. Consulte novamente os KPIs e requisitos de negócios documentados para ajudar a priorizar qual atualização trará um impacto positivo à empresa.

Recursos

Documentos relacionados:

- Blog da AWS
- Novidades da AWS
- Implementar imagens atualizadas com pipelines automatizados do EC2 Image Builder

Vídeos relacionados:

- AWS re:Inforce 2022: Automatizar o gerenciamento e a conformidade de patches usando a AWS
- All Things Patch: AWS Systems Manager | Eventos da AWS

Exemplos relacionados:

- Gerenciamento de inventário e patches
- Workshop One Observability

PERF05-BP07 Revisar as métricas regularmente

Como parte da manutenção de rotina, ou em resposta a eventos ou incidentes, revise quais métricas são coletadas. Use essas análises para identificar quais métricas foram essenciais para resolver

problemas e quais métricas adicionais poderiam ajudar a identificar, resolver ou prevenir problemas se estivessem sendo acompanhadas.

Práticas comuns que devem ser evitadas:

- Você permite que as métricas permaneçam em um estado de alarme por um período prolongado.
- Você cria alarmes que não são acionáveis por um sistema de automação.

Benefícios de implementar esta prática recomendada: analise continuamente as métricas que estão sendo coletadas para garantir que identifiquem, resolvam ou evitem problemas corretamente. As métricas também podem se tornar obsoletas se você permitir que elas permaneçam em um estado de alarme por um período prolongado.

Nível de risco exposto se esta prática recomendada não for estabelecida: Médio

Orientação para implementação

Melhore constantemente a coleta e o monitoramento de métricas. Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use este método para aprimorar a qualidade das métricas coletadas para prevenir ou resolver incidentes futuros mais rapidamente.

Como parte da resposta a incidentes ou eventos, avalie as métricas que foram úteis para resolver o problema e quais poderiam ter ajudado, mas não estão sendo acompanhadas no momento. Use esses dados para aprimorar a qualidade das métricas coletadas para prevenir ou resolver incidentes futuros mais rapidamente.

Etapas de implementação

- Defina métricas: defina métricas críticas de performance para monitorar que estejam alinhadas aos objetivos da sua workload, incluindo métricas como tempo de resposta e utilização de recursos.
- Estabeleça linhas de base: defina uma linha de base e um valor desejável para cada métrica. A linha de base deve fornecer pontos de referência para a identificação de desvios ou anomalias.
- Defina uma frequência: defina uma frequência (como semanal ou mensal) para revisar as métricas essenciais.
- Identifique problemas de performance: durante cada revisão, avalie as tendências e o desvio dos valores base. Procure gargalos ou anomalias de performance. Para os problemas identificados, realize uma análise aprofundada da causa-raiz para entender o principal motivo do problema.

- Identifique ações corretivas: use sua análise para identificar ações corretivas. Isso pode incluir ajuste de parâmetros, correção de bugs e ajustes na escala dos recursos.
- Documente as descobertas: documente suas descobertas, incluindo problemas identificados, causas-raiz e ações corretivas.
- Itere e aprimore: avalie e melhore constantemente o processo de revisão de métricas. Use a lição aprendida com a análise anterior para aprimorar o processo ao longo do tempo.

Recursos

Documentos relacionados:

- Documentação do CloudWatch
- Coletar métricas e logs de instâncias do Amazon EC2 e servidores on-premises com o CloudWatch Agent
- Consultar métricas com o CloudWatch Metrics Insights
- Parceiros de monitoramento, log e performance da AWS Partner Network
- Documentação do X-Ray

Vídeos relacionados:

- AWS re:Invent 2022: Configurar controles em escala em seu ambiente da AWS
- AWS re:Invent 2022: Como a Amazon usa métricas melhores para aprimorar a performance de sites
- AWS re:Invent 2023: Criar uma estratégia efetiva de observabilidade
- AWS Summit SF 2022: Observabilidade full-stack e monitoramento de aplicações com a AWS
- AWS re:Invent 2023: Relaxe: diagnostique e resolva problemas de performance com o Amazon RDS

Exemplos relacionados:

- Criar um painel com o QuickSight
- Painéis do CloudWatch

Conclusão

Atingir e manter a eficiência de performance requer uma abordagem conduzida por dados. Considere ativamente os padrões de acesso e as concessões que permitirão a você otimizar para uma maior performance. O uso de um processo de análise baseado em benchmarks e testes de carga permite selecionar os tipos de recursos e as configurações adequados. Tratar sua infraestrutura como código permite que você promova avanços em sua arquitetura de modo rápido e seguro, enquanto usa dados para tomar decisões baseadas em fatos sobre sua arquitetura. O estabelecimento de uma combinação de monitoramentos ativo e passivo garante que a performance de sua arquitetura não apresente degradação ao longo do tempo.

A AWS não mede esforços para ajudar você a criar arquiteturas que ofereçam uma performance eficiente enquanto entregam valor empresarial. Use as ferramentas e técnicas abordadas neste artigo para garantir o sucesso.

Colaboradores

Os seguintes indivíduos e organizações contribuíram para este documento:

- Sam Mokhtari, arquiteto líder de soluções de eficiência, Amazon Web Services
- Josh Hart, arquiteto de soluções, Amazon Web Services
- Richard Trabing, arquiteto de soluções, Amazon Web Services
- Brett Looney, arquiteto líder de soluções, Amazon Web Services
- Nina Vogl, arquiteta líder de soluções, Amazon Web Services
- Eric Pullen, arquiteto de soluções, Amazon Web Services
- Julien Lépine, gerente especialista de SA, Amazon Web Services
- Ronnen Slasky, arquiteto de soluções, Amazon Web Services

Outras fontes de leitura

Para obter ajuda adicional, consulte as seguintes fontes:

- Framework Well-Architected da AWS
- Centro de Arquitetura do AWS

Revisões do documento

Para ser notificado sobre atualizações desse whitepaper, inscreva-se no feed RSS.

Alteração	Descrição	Data
Atualizações secundárias em práticas recomendadas	A PERF03-BP04 foi atualizada com novas recomendações de serviço.	6 de novembro de 2024
Orientação atualizada sobre práticas recomendadas	Várias pequenas atualizações em todo o pilar.	27 de junho de 2024
Atualização e reestruturação importantes	O pilar foi reestruturado para ter cinco áreas de práticas recomendadas (antes eram oito). O conteúdo foi consolida do nas cinco áreas e atualizad o. As novas áreas de práticas recomendadas são Seleção de arquitetura, Computação e hardware, Gerenciamento de dados, Redes e entrega de conteúdo e Processo e cultura.	3 de outubro de 2023
Atualização secundária	Remoção de linguagem não inclusiva.	13 de abril de 2023
Atualizações para o novo Framework	Atualizações nas práticas recomendadas com recomendações e adição de novas práticas recomendadas.	10 de abril de 2023
Whitepaper atualizado	Práticas recomendadas atualizadas com novas	15 de dezembro de 2022

	orientações para implement ação.	
Whitepaper atualizado	Práticas recomendadas ampliadas e planos de melhoria adicionados.	20 de outubro de 2022
Atualização secundária	Remoção de linguagem não inclusiva.	22 de abril de 2022
Atualizações menores	Links atualizados.	10 de março de 2021
Atualizações menores	Alteração do tempo limite do AWS Lambda para 900 segundos e correção do nome do Amazon Keyspaces (para Apache Cassandra).	5 de outubro de 2020
Atualização secundária	Link quebrado corrigido.	15 de julho de 2020
Atualizações para o novo Framework	Revisão e atualização importantes no conteúdo	8 de julho de 2020
Whitepaper atualizado	Pequena atualização devido a problemas gramaticais	1º de julho de 2018
Whitepaper atualizado	Whitepaper atualizado para refletir as alterações na AWS	1 de novembro de 2017
Publicação inicial	Publicação do pilar Eficiênci a de performance: AWS Well-Architected Framework.	1º de novembro de 2016

Avisos

Os clientes são responsáveis por fazer a própria avaliação independente das informações contidas neste documento. Este documento: (a) é apenas para fins informativos, (b) representa as ofertas e práticas de produtos atuais da AWS, que estão sujeitas a alterações sem aviso prévio e (c) não criam nenhum compromisso ou garantia da AWS e de suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos “no estado em que se encontram”, sem garantias, representações ou condições de qualquer tipo, expressas ou implícitas. As responsabilidades e as obrigações da AWS com os seus clientes são controladas por contratos da AWS, e este documento não é parte de, nem modifica, qualquer contrato entre a AWS e seus clientes.

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

Glossário da AWS

Para obter a terminologia mais recente da AWS, consulte o glossário da AWS na Referência do Glossário da AWS.