AWS Well-Architected Framework

Julho 2020

Este documento descreve o AWS Well-Architected Framework, que permite analisar e aprimorar as arquiteturas baseadas em nuvem e entender melhor o impacto comercial de suas decisões de projeto. Abordamos princípios gerais de design, bem como melhores práticas e orientações específicas em cinco áreas conceituais que definimos como *pilares* do Well-Architected Framework.

Avisos

Os clientes são responsáveis por fazer sua própria avaliação independente das informações neste documento. Este documento (a) é fornecido 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 cria nenhum compromisso ou garantia da AWS e suas afiliadas, fornecedores ou licenciadores. Os produtos ou serviços da AWS são fornecidos no “estado em que se encontram”, sem qualquer garantia, declaração ou condição de qualquer tipo, explícita ou implícita. As responsabilidades e obrigações da AWS para com seus clientes são regidas por contratos da AWS, e este documento não modifica nem faz parte de nenhum contrato entre a AWS e seus clientes.

Introdução

O AWS Well-Architected Framework ajuda a entender os prós e os contras das decisões que você toma ao criar sistemas na AWS. Ao usar o Framework, você aprenderá as melhores práticas de arquitetura para projetar e operar sistemas confiáveis, seguros, eficientes e econômicos na nuvem. Ele fornece uma maneira de você avaliar consistentemente suas arquiteturas em relação às melhores práticas e identificar áreas de melhoria. O processo para revisar uma arquitetura é uma conversa construtiva sobre decisões de arquitetura e não é um mecanismo de auditoria. Acreditamos que ter os sistemas Well-Architected aumenta muito a probabilidade de sucesso nos negócios.

Os arquitetos de soluções da AWS têm anos de experiência na arquitetura de soluções em uma ampla variedade de verticais de negócios e casos de uso. Ajudamos a projetar e analisar as arquiteturas de milhares de clientes na AWS. Por meio dessa experiência, identificamos as melhores práticas e principais estratégias para a arquitetura de sistemas na nuvem.

O AWS Well-Architected Framework documenta um conjunto de perguntas fundamentais que permitem compreender se uma arquitetura específica se alinha bem às melhores práticas da nuvem. A estrutura fornece uma abordagem consistente para avaliar os sistemas em relação às qualidades que você espera dos sistemas modernos baseados em nuvem e a correção necessária para alcançar essas qualidades. Conforme a AWS continua evoluindo, e continuamos a saber mais sobre o trabalho com nossos clientes, continuaremos refinando a definição do Well-Architected.

Este Framework é destinado a pessoas que ocupam cargos de tecnologia, como diretores de tecnologia (CTOs), arquitetos, desenvolvedores e membros da equipe de operações. Ele descreve as melhores práticas e estratégias da AWS a serem usadas ao projetar e operar uma carga de trabalho na nuvem e fornece links para mais detalhes de implementação e padrões de arquitetura. Para obter mais informações, consulte a página inicial do AWS Well-Architected.

A AWS também fornece um serviço para analisar suas cargas de trabalho gratuitamente. O AWS Well-Architected Tool (AWS WA Tool) é um serviço na nuvem que fornece um processo consistente para analisar e medir a arquitetura usando o AWS Well-Architected Framework. O AWS WA Tool fornece recomendações para tornar suas cargas de trabalho mais confiáveis, seguras, eficientes e econômicas.

Para ajudá-lo a aplicar as melhores práticas, criamos o AWS Well-Architected Labs, que oferece um repositório de código e documentação para que você tenha uma experiência prática na implementação das melhores práticas. Também nos juntamos a parceiros selecionados da rede de parceiros da AWS (APN), membros do programa de parceiros do AWS Well-Architected. Esses parceiros do APN têm um profundo conhecimento da AWS e podem ajudá-lo a analisar e melhorar suas cargas de trabalho.

Definições

Todos os dias, os especialistas da AWS ajudam os clientes a arquitetar sistemas para aproveitar as melhores práticas na nuvem. Trabalhamos com você para oferecer vantagens e desvantagens arquitetônicas à medida que seus projetos evoluem. Conforme você implanta esses sistemas em ambientes dinâmicos, aprendemos como esses sistemas se desempenham e as consequências dessas vantagens e desvantagens.

Com base no que aprendemos, criamos o AWS Well-Architected Framework, que fornece um conjunto consistente de melhores práticas para clientes e parceiros avaliarem arquiteturas e um conjunto de perguntas que você pode usar para avaliar o alinhamento de uma arquitetura com as melhores práticas da AWS.

O AWS Well-Architected Framework é baseado em cinco pilares: excelência operacional, segurança, confiabilidade, eficiência de performance e otimização de custos.

Tabela 1. Os pilares do AWS Well-Architected Framework

Nome	Descrição
Excelência operacional	A capacidade de apoiar o desenvolvimento e executar cargas de trabalho com eficácia, obter insights sobre as operações e melhorar continuamente processos e procedimentos de suporte para oferecer valor empresarial.
Segurança	O pilar Segurança refere-se à capacidade de proteger dados, sistemas e ativos para utilizar as tecnologias de nuvem para melhorar sua segurança.
Confiabilidade	É a capacidade de uma carga de trabalho executar a função pretendida de forma correta e consistente quando o esperado. Isso inclui a capacidade de operar e testar a carga de trabalho durante todo o ciclo de vida.
Eficiência de performance	a capacidade de usar recursos de computação com eficiência para atender aos requisitos do sistema e manter essa eficiência à medida que a demanda muda e as tecnologias evoluem.
Otimização de custos	A capacidade de executar sistemas para entregar o valor empresarial ao menor preço

No AWS Well-Architected Framework, usamos esses termos

- Um **componente** é o código, a configuração e os recursos da AWS que juntos atendem a um requisito. Um componente geralmente é a unidade de propriedade técnica e é dissociado de outros componentes.

- Usamos o termo **carga de trabalho** para identificar um conjunto de componentes que juntos fornecem valor empresarial. A carga de trabalho é normalmente o nível de detalhes sobre o qual os líderes de negócios e tecnologia se comunicam.
- **Marcos** assinalam as principais alterações na arquitetura, à medida que passa pelo ciclo de vida do produto (design, teste, ativação e produção).
- Consideramos a **arquitetura** a forma como os componentes funcionam juntos em uma carga de trabalho. Como os componentes se comunicam e interagem é, com frequência, o foco dos diagramas de arquitetura.
- Dentro de uma organização, o **portfólio de tecnologia** é a coleção de cargas de trabalho necessárias para o negócio operar.

Ao arquitetar cargas de trabalho, você obtém vantagens e desvantagens entre pilares com base no contexto da sua empresa. Essas decisões de negócios podem conduzir suas prioridades de engenharia. Você pode otimizar para reduzir custos e assim diminuir a confiabilidade em ambientes de desenvolvimento ou otimizar a confiabilidade e aumentar os custos para soluções importantes. Em soluções de comércio eletrônico, a performance pode afetar a receita e a propensão do cliente a comprar. Segurança e excelência operacional geralmente não têm vantagens e desvantagens em relação aos outros pilares.

Sobre arquitetura

Em ambientes locais, os clientes geralmente têm uma equipe central de arquitetura de tecnologia que atua como uma sobreposição para outras equipes de produtos ou recursos para garantir que estejam seguindo as melhores práticas. As equipes de arquitetura de tecnologia geralmente são compostas por um conjunto de funções, como arquiteto técnico (infraestrutura), arquiteto de soluções (software), arquiteto de dados, arquiteto de redes e arquiteto de segurança. Muitas vezes, essas equipes usam o TOGAF ou o Zachman Framework como parte de um recurso de arquitetura empresarial.

Na AWS, preferimos distribuir os recursos para as equipes, em vez de termos uma equipe centralizada com esse recurso. Existem riscos na escolha de distribuir autoridade para tomada de decisões como, por exemplo, garantir que as equipes atendam aos padrões internos. Atenuamos esses riscos de duas formas. Primeiro, temos *práticas* [1] que se concentram em permitir que cada equipe tenha essa capacidade, e colocamos em prática especialistas que garantem que as equipes elevem o nível dos padrões que precisam cumprir. Segundo, implementamos *mecanismos* [2] que realizam verificações automatizadas para garantir que os padrões estejam sendo atendidos. Essa aborda-

[1] Formas de fazer as coisas, processos, padrões e normas aceitas.
[2] "Boas intenções nunca funcionam, você precisa de bons mecanismos para fazer qualquer coisa acontecer com" Jeff Bezos. Isso significa substituir os melhores esforços humanos por mecanismos (muitas vezes automatizados) que verificam a conformidade com regras ou processos.

gem distribuída é apoiada pelos princípios de liderança da Amazon e estabelece uma cultura em todas as funções que *funciona retroativamente* [3] do cliente. As equipes dedicadas ao cliente criam produtos em resposta a uma necessidade do cliente.

Na arquitetura, isso significa que esperamos que todas as equipes tenham a capacidade de criar arquiteturas e seguir as melhores práticas. Para ajudar as novas equipes a chegar nessa capacidade ou as equipes existentes a elevar seus padrões, viabilizamos o acesso a uma comunidade virtual de engenheiros principais que podem analisar os projetos delas e ajudá-las a entender quais são as melhores práticas da AWS. A comunidade de engenharia principal trabalha para tornar as melhores práticas visíveis e acessíveis. Uma forma de fazer isso, por exemplo, é por meio de palestras na hora do almoço, focadas na aplicação das melhores práticas a exemplos reais. Essas conversas são gravadas e podem ser usadas como parte dos materiais de integração para novos membros da equipe.

As melhores práticas da AWS surgem de nossa experiência na execução de milhares de sistemas em escala da internet. Preferimos usar dados para definir as melhores práticas, mas também usamos especialistas no assunto, como os engenheiros principais, para defini-los. À medida que os engenheiros principais veem surgir novas melhores práticas, eles trabalham como uma comunidade para garantir que as equipes as sigam. Com o tempo, essas melhores práticas são formalizadas em nossos processos internos de análise, bem como em mecanismos que reforçam a conformidade. O Well-Architected é a implementação voltada para o cliente do nosso processo de análise interna, em que codificamos nosso pensamento de engenharia principal em funções de campo, como a arquitetura de soluções e equipes de engenharia internas. O Well-Architected é um mecanismo escalável que permite que você aproveite esses aprendizados.

Seguindo a abordagem de uma comunidade de engenharia principal com propriedade distribuída da arquitetura, acreditamos que uma arquitetura corporativa do Well-Architected pode emergir, impulsionada pela necessidade do cliente. Líderes de tecnologia (como CTOs ou gerentes de desenvolvimento), realizando análises do Well-Architected em todas as suas cargas de trabalho, permitirão uma melhor compreensão dos riscos em seu portfólio de tecnologia. Usando essa abordagem, você pode identificar temas entre as equipes que sua organização poderia abordar por mecanismos, treinamentos ou palestras na hora do almoço, em que seus engenheiros principais possam compartilhar seus pensamentos sobre áreas específicas com várias equipes.

Princípios gerais do projeto

O Well-Architected Framework identifica um conjunto de princípios gerais do projeto para facilitar um bom projeto na nuvem:

[3] O funcionamento retroativo é uma parte fundamental do nosso processo de inovação. Começamos com o cliente e o que ele quer, e deixamos que isso defina e oriente os nossos esforços.

- **Pare de adivinhar suas necessidades de capacidade**: Elimine as suposições ao determinar sua necessidade de capacidade de infraestrutura. Ao tomar uma decisão de capacidade antes de implantar um sistema, você pode ficar com recursos ociosos caros ou lidar com as implicações da performance de capacidade limitado. Com a computação em nuvem, esses problemas terminaram. Você pode usar a quantidade de capacidade e aumentar e diminuir a escala automaticamente.
- **Teste sistemas em escala de produção**: Na nuvem, você pode criar um ambiente de teste em escala de produção sob demanda, concluir seus testes e descomissionar os recursos. Como você paga somente pelo ambiente de teste quando está em execução, é possível simular seu ambiente ativo por uma fração do custo dos testes no local.
- **Automatize para facilitar a experimentação arquitetônica**: A automação permite criar e replicar seus sistemas a baixo custo e evitar a despesa do esforço manual. Você pode acompanhar as alterações em sua automação, auditar o impacto e reverter para os parâmetros anteriores, quando necessário.
- **Permita arquiteturas evolutivas**: Permita arquiteturas evolutivas. Em um ambiente tradicional, as decisões de arquitetura são frequentemente implementadas como eventos estáticos e únicos, com algumas versões principais de um sistema durante sua vida útil. À medida que uma empresa e seu contexto continuam a mudar, essas decisões iniciais podem prejudicar a capacidade do sistema de fornecer requisitos de negócios variáveis. Na nuvem, a capacidade de automatizar e testar sob demanda reduz o risco de impacto das alterações no projeto. Isso permite que os sistemas evoluam com o tempo, para que as empresas possam tirar proveito das inovações como prática padrão.
- **Impulsione arquiteturas usando dados**: Na nuvem, você pode coletar dados sobre como suas escolhas arquitetônicas afetam o comportamento da carga de trabalho. Isso permite que você tome decisões baseadas em fatos sobre como melhorar sua carga de trabalho. Sua infraestrutura de nuvem é código, portanto, você pode usar esses dados para informar suas escolhas e melhorias na arquitetura ao longo do tempo.
- **Aprimore por meio dos dias de jogo**: Teste a performance e os processos de sua arquitetura agendando regularmente dias de jogo para simular eventos em produção. Isso ajudará a compreender onde as melhorias podem ser feitas e pode ajudar a desenvolver experiência organizacional ao lidar com eventos.

Os cinco pilares do Framework

Criar um sistema de software é como construir um edifício. Se a fundação não for sólida, os problemas estruturais poderão prejudicar a integridade e a função do edifício. Ao arquitetar soluções tecnológicas, se você negligenciar os cinco pilares (excelência operacional, segurança, confiabilidade, eficiência de performance e otimização de custos), poderá ser um desafio criar um sistema que atenda às suas expectativas e exigências. A incorporação desses pilares em sua arquitetura o ajudará a produzir sistemas estáveis e eficientes. Isso permitirá que você se concentre nos outros aspectos do projeto, como requisitos funcionais.

Excelência operacional

O pilar (pilar) inclui (descrição)

O pilar Excelência operacional apresenta uma visão geral dos princípios de design, melhores práticas e perguntas. Você pode encontrar orientações prescritivas sobre implementação no whitepaper Pilar Excelência operacional.

Princípios de design

Existem (contagem) princípios do projeto para (pilar inferior) na nuvem:

- **Executar operações como código**: Na nuvem, você pode aplicar a mesma disciplina de engenharia usada para o código do aplicativo em todo o ambiente. É possível definir toda a sua carga de trabalho (aplicativos, infraestrutura) como código e atualizá-la com código. Você pode implementar seus procedimentos de operações como código e automatizar sua execução acionando-os em resposta a eventos. Ao executar operações como código, você limita o erro humano e permite respostas consistentes aos eventos.
- **Fazer alterações frequentes, pequenas e reversíveis**: Projetar cargas de trabalho para permitir que os componentes sejam atualizados regularmente. Faça alterações em pequenos incrementos que possam ser revertidas em caso de falha (sem afetar os clientes quando possível).
- **Refinar procedimentos de operações com frequência**: Ao usar os procedimentos de operação, procure oportunidades para melhorá-los. Conforme você evolui sua carga de trabalho, evolua seus procedimentos adequadamente. Organize dias de jogo regularmente para analisar e validar se todos os procedimentos são eficazes e se as equipes estão familiarizadas com eles.
- **Antecipar falhas**: Execute os exercícios “pré-mortem” para identificar as potenciais origens de falhas, para que assim elas possam ser removidas ou mitigadas. Testar

cenários de fala e validar como você compreende o impacto deles. Teste seus procedimentos de resposta para garantir que eles são eficazes e que as equipes estão familiarizadas com a execução deles. Organize dias de jogo regularmente para testar cargas de trabalho e respostas da equipe a eventos simulados.

- **Aprenda com todas as falhas operacionais**: Promova a melhoria através das lições aprendidas em todos os eventos e falhas operacionais. Compartilhe o que foi aprendido com as equipes e a organização inteira.

Definição

Existem (contagem) melhores práticas para (pilar inferior) na nuvem:

- **Organização**
- **Preparar**
- **Operar**
- **Evoluir**

A liderança da sua organização define objetivos empresariais. Sua organização deve compreender requisitos e prioridades e usá-los para organizar e conduzir trabalhos para apoiar a obtenção de resultados empresariais. Sua carga de trabalho deve emitir as informações necessárias para apoiá-la. A implementação de serviços para possibilitar a integração, a implantação e a entrega de sua carga de trabalho permitirá um fluxo maior de alterações benéficas na produção por meio da automação de processos repetitivos.

Pode haver riscos inerentes à operação da carga de trabalho. Você deve compreender esses riscos e tomar uma decisão embasada para entrar na produção. Suas equipes devem ser capazes de dar suporte à sua carga de trabalho. As métricas operacionais e de negócios derivadas dos resultados de negócios desejados permitirão que você compreenda a integridade da carga de trabalho e as atividades de operações e responda a incidentes. Suas prioridades mudarão à medida que suas necessidades de negócios e o ambiente de negócios mudarem. Use isso como um ciclo de comentários para promover continuamente melhorias para a sua organização e a operação da sua carga de trabalho.

Melhores práticas

Organização

Suas equipes precisam ter um entendimento compartilhado de toda a sua carga de trabalho, da função que desempenham em tudo isso e dos objetivos de negócios

compartilhados a fim de definir as prioridades que permitirão o êxito dos negócios. Prioridades bem definidas maximizarão os benefícios dos seus esforços. Avalie as necessidades de clientes internos e externos, envolvendo as principais partes interessadas, incluindo equipes corporativas, de desenvolvimento e operacionais, a fim de determinar onde concentrar os esforços. A avaliação das necessidades do cliente garantirá que você tenha um entendimento completo do suporte necessário para obter resultados nos negócios. Esteja ciente das diretrizes ou obrigações definidas pela governança organizacional e de fatores externos, como requisitos de conformidade regulamentar e normas do setor, que podem exigir ou enfatizar um foco específico. Confirme se você tem os mecanismos para identificar alterações na governança interna e nos requisitos de conformidade externos. Se nenhum requisito for identificado, aplique a auditoria devida para essa determinação. Analise suas prioridades regularmente para que elas possam ser atualizadas conforme as necessidades mudam.

Avalie ameaças à empresa (por exemplo, riscos e passivos empresariais e ameaças à segurança da informação) e mantenha essas informações em um registro de risco. Avalie o impacto dos riscos e as compensações entre interesses concorrentes ou abordagens alternativas. Por exemplo, a aceleração da velocidade de entrada no mercado de novos recursos pode ser enfatizada em relação à otimização de custos, ou você pode escolher um banco de dados relacional para dados não relacionais para simplificar o esforço de migração de um sistema. Gerencie benefícios e riscos para tomar decisões informadas ao determinar onde concentrar os esforços. Alguns riscos ou opções podem ser aceitáveis por um tempo. Talvez seja possível mitigar os riscos associados ou talvez seja inaceitável permitir que um risco permaneça; nesse caso você tomará as devidas medidas para resolver o risco.

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes precisam entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas e ter objetivos compartilhados. Entender a responsabilidade, a propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes. As necessidades de uma equipe são modeladas pelo cliente que ela auxilia, pela organização, pela formação da equipe e pelas características da carga de trabalho. Não é sensato esperar que um modelo operacional único seja capaz de dar suporte a todas as equipes e suas respectivas cargas de trabalho em sua organização.

Certifique-se de que haja proprietários identificados para cada componente de aplicativo, carga de trabalho, plataforma e infraestrutura, e que cada processo e procedimento tenha um proprietário identificado responsável pela definição e proprietários responsáveis pela performance. Entender o valor empresarial de cada componente, processo e procedimento, da razão pela qual esses recursos estão em vigor ou de por que as atividades são executadas e por que essa propriedade existe informará as ações dos membros da equipe. Defina claramente as responsabilidades dos membros da equipe para que eles possam agir adequadamente e ter mecanismos para identificar responsabilidade e propriedade. Tenha mecanismos para solicitar adições, altera-

ções e exceções para que você não restrinja a inovação. Defina contratos entre equipes que descrevem como elas trabalham juntas para apoiar umas às outras e seus resultados de negócios.

Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados empresariais. A liderança sênior engajada deve definir expectativas e medir o sucesso. Ela deve ser patrocinadora, defensora e motivadora da adoção das melhores práticas e da evolução da organização. Capacite os membros da equipe a tomar medidas quando os resultados estiverem em risco para minimizar o impacto e os incentive a encaminhar para os tomadores de decisão e as partes interessadas quando acharem que há um risco para que isso possa ser resolvido e evitar incidentes. Forneça comunicações oportunas, claras e acionáveis de riscos conhecidos e eventos planejados para que os membros da equipe possam tomar as medidas apropriadas e oportunas.

Incentive a experimentação para acelerar o aprendizado e manter os membros da equipe interessados e envolvidos. As equipes devem aumentar os conjuntos de habilidades para adotar novas tecnologias e apoiar mudanças na demanda e nas responsabilidades. Dê apoio e incentivo a isso fornecendo um tempo de estrutura dedicado para o aprendizado. Garanta que os membros da equipe tenham os recursos, tanto ferramentas quanto pessoas, para serem bem-sucedidos e escalar para auxiliar os resultados empresariais. Aproveite a diversidade entre organizações para buscar várias perspectivas únicas. Use essa abordagem para aumentar a inovação, desafiar suas suposições e reduzir o risco de viés de confirmação. Aumente a inclusão, a diversidade e a acessibilidade em suas equipes para obter perspectivas benéficas.

Se houver requisitos externos de regulamentação ou conformidade aplicáveis à sua organização, use os recursos fornecidos pela Conformidade com a nuvem AWS para ajudar a instruir suas equipes de modo que elas possam determinar o impacto em suas prioridades. O Well-Architected Framework enfatiza o aprendizado, a medição e a melhoria. Ele fornece uma abordagem consistente para você avaliar arquiteturas e implementar projetos que aumentarão de escala ao longo do tempo. A AWS fornece o AWS Well-Architected Tool para ajudar você a analisar sua abordagem antes do desenvolvimento, o estado das cargas de trabalho antes da produção e o estado das cargas de trabalho na produção. Você pode compará-las com as melhores práticas de arquitetura da AWS mais recentes, monitorar o status geral de suas cargas de trabalho e obter insights sobre possíveis riscos. O AWS Trusted Advisor é uma ferramenta que fornece acesso a um conjunto principal de verificações que recomendam otimizações que podem ajudar a moldar suas prioridades. Os clientes Business e Enterprise Support recebem acesso a verificações adicionais com foco em segurança, confiabilidade, performance e otimização de custos que podem ajudar a moldar as prioridades.

A AWS pode ajudar a instruir suas equipes sobre a AWS e os serviços oferecidos por ela para aumentar o entendimento do impacto das opções na carga de trabalho. Você deve usar os recursos fornecidos pelo AWS Support (AWS Knowledge Center, AWS Discussion Forms e AWS Support Center) e pelo AWS Documentation para educar su-

as equipes. Entre em contato com o AWS Support pelo AWS Support Center para receber ajuda com suas perguntas da AWS. A AWS também compartilha melhores práticas e padrões que aprendemos durante a operação da AWS na Amazon Builders' Library. Uma variedade de outras informações úteis está disponível no blog da AWS e no podcast oficial da AWS. O AWS Training and Certification oferece treinamento gratuito por meio de cursos digitais autoguiados sobre os fundamentos da AWS. Você também pode se inscrever para um treinamento presencial com instrutor para apoiar ainda mais o desenvolvimento das habilidades de suas equipes com a AWS.

Você deve usar ferramentas ou serviços que permitam controlar centralmente seus ambientes em todas as contas, como o AWS Organizations, para ajudar a gerenciar seus modelos operacionais. Serviços como o AWS Control Tower expandem esse recurso de gerenciamento, permitindo que você defina esquemas (compatíveis com modelos operacionais) para a configuração de contas, aplique governança contínua usando o AWS Organizations e automatize o provisionamento de novas contas. Os provedores de serviços gerenciados, como o AWS Managed Services, o AWS Managed Services Partners ou provedores de serviços gerenciados na rede de parceiros da AWS, fornecem especialização na implementação de ambientes de nuvem e dão suporte aos seus requisitos de segurança e conformidade e objetivos empresariais. A adição de serviços gerenciados ao seu modelo operacional pode economizar tempo e recursos, além de permitir que você mantenha as equipes internas reduzidas e focadas em resultados estratégicos que diferenciarão seus negócios, em vez de desenvolver novas habilidades e recursos.

As perguntas a seguir se concentram nessas considerações para (pilar inferior). (Para uma lista de perguntas e melhores práticas sobre (pilar inferior), leia o Apêndice.).

OPS 1: Como você determina quais são suas prioridades?
Todos precisam entender seu papel no sucesso nos negócios. Tenha objetivos compartilhados para definir as prioridades dos recursos. Isso maximizará os benefícios de seus esforços.
OPS 2: Como você estrutura sua organização para dar suporte aos seus resultados comerciais?
Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes precisam entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas e ter objetivos compartilhados. Entender a responsabilidade, a propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes.
OPS 3: Como sua cultura organizacional oferece suporte aos resultados comerciais?
Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados comerciais.

Em determinado momento, talvez você deseje destacar um pequeno subconjunto de prioridades. Use uma abordagem equilibrada em longo prazo para garantir o desenvolvimento dos recursos necessários e o gerenciamento de riscos. Reveja as prioridades regularmente e as atualize conforme as necessidades mudam. Quando a respon-

sabilidade e a propriedade não foram definidas ou não são conhecidas, você corre o risco de não realizar as ações necessárias em tempo hábil e de despender esforços redundantes e possivelmente conflitantes para atender a essas necessidades. A cultura organizacional tem impacto direto na satisfação com a tarefa e na retenção dos membros da equipe. Incentive o envolvimento e as habilidades dos membros da equipe para promover o êxito da sua empresa. A experimentação é necessária para que a inovação ocorra e transforme ideias em resultados. Reconheça que um resultado indesejado é um experimento com êxito que identificou um caminho que não levará ao êxito.

Preparar

Para se preparar para a excelência operacional, você precisa entender suas cargas de trabalho e os comportamentos esperados. Você poderá projetá-los para fornecer insights sobre seu status e criar os procedimentos para apoiá-los.

Projete sua carga de trabalho para que as informações necessárias sejam fornecidas a fim de que você entenda seu estado interno (tais como métricas, logs, eventos e rastreamento) em todos os componentes, em apoio à capacidade de observação e à investigação de problemas. Itere para desenvolver a telemetria necessária para monitorar a integridade da carga de trabalho, identificar quando os resultados estão em risco e permitir respostas eficazes. Ao instrumentar sua carga de trabalho, colete um amplo conjunto de informações para permitir a percepção situacional (por exemplo, alterações de estado, atividade do usuário, acesso a privilégios, contadores de utilização), sabendo que é possível usar filtros para selecionar as informações mais úteis ao longo do tempo.

Adote abordagens que melhoram o fluxo de alterações na produção e permitem refatoração, comentários rápidos sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e permite a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação ou descobertos em seus ambientes.

Adote abordagens que forneçam feedback rápido sobre a qualidade e permitam recuperação rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças. Planeje alterações malsucedidas para que você possa responder mais rapidamente, se necessário, e testar e validar as alterações feitas. Esteja ciente das atividades planejadas em seus ambientes para que você possa gerenciar o risco de alterações que afetem as atividades planejadas. Enfatize alterações frequentes, pequenas e reversíveis para limitar o escopo das alterações. Isso resulta em solução de problemas mais fácil e correção mais rápida, com a opção de reverter uma alteração. Isso também significa que você pode conseguir o benefício de alterações valiosas com mais frequência.

Avalie a prontidão operacional de carga de trabalho, processos, procedimentos e pessoal para compreender os riscos operacionais relacionados à carga de trabalho. Você

deve usar um processo consistente (incluindo listas de verificação manuais ou automatizadas) para saber quando está pronto para trabalhar com sua carga de trabalho ou para fazer uma mudança. Isso também permitirá que você encontre as áreas que precisa abordar. Tenha runbooks que documentem suas atividades de rotina e playbooks que orientem seus processos para a resolução de problemas. Entenda os benefícios e os riscos para tomar decisões informadas para permitir que as alterações entrem na produção.

A AWS permite que você visualize toda a carga de trabalho (aplicativos, infraestrutura, política, governança e operações) como código. Tudo pode ser definido e atualizado usando o código. Isso significa que você pode aplicar a mesma disciplina de engenharia usada para o código do aplicativo a cada elemento da pilha e compartilhá-los entre equipes ou organizações para ampliar os benefícios dos esforços de desenvolvimento. Use operações como código na nuvem e a capacidade de experimentar com segurança para desenvolver sua carga de trabalho, procedimentos de operações e praticar falhas. O uso do AWS CloudFormation permite que você tenha ambientes consistentes, com modelos, desenvolvimento de sandbox, teste e produção de área restrita, com níveis crescentes de controle de operações.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

OPS 4: Como você projeta sua carga de trabalho para entender o estado dela?
Projete sua carga de trabalho para que as informações necessárias sejam fornecidas em todos os componentes (tais como métricas, logs e rastreamento) a fim de que você entenda seu estado interno. Isso permite que você forneça respostas efetivas quando for apropriado.
OPS 5: Como você reduz defeitos, facilita a correção e melhora o fluxo na produção?
Adote abordagens que melhoram o fluxo de alterações na produção, que permitem refatoração, feedback rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e permite a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação.
OPS 6: Como você reduz os riscos de implantação?
Adote abordagens que forneçam feedback rápido sobre a qualidade e permitam recuperação rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças.
OPS 7: Como você sabe que está pronto para oferecer suporte a uma carga de trabalho?
Avalie a prontidão operacional de sua carga de trabalho, processos/procedimentos e pessoal para entender os riscos operacionais relacionados.

Invista na implementação de atividades operacionais como código para maximizar a produtividade do pessoal de operações, minimizar taxas de erro e permitir respostas automatizadas. Use as estratégias “pre-mortem” para antecipar falhas e criar procedimentos, quando apropriado. Aplique metadados usando tags de recursos e AWS Resource Groups seguindo uma estratégia consistente de marcação para permitir a identificação de seus recursos. Identifique seus recursos para organização, contabilidade

de custos, controles de acesso e direcione a execução de atividades operacionais automatizadas. Adote práticas de implantação que aproveitem a elasticidade da nuvem para facilitar as atividades de desenvolvimento e a pré-implantação de sistemas para implementações mais rápidas. Ao fazer alterações nas listas de verificação usadas para avaliar suas cargas de trabalho, planeje o que você fará com sistemas ativos que não estejam mais em conformidade.

Operar

A operação bem-sucedida de uma carga de trabalho é medida pela obtenção de resultados de negócios e de clientes. Defina os resultados esperados, determine como o sucesso será medido e identifique as métricas que serão usadas nesses cálculos para determinar se a carga de trabalho e as operações foram bem-sucedidas. A integridade operacional inclui a integridade da carga de trabalho e a integridade e o sucesso de operações realizadas em apoio à carga de trabalho (por exemplo, implantação e resposta a incidentes). Estabeleça linhas de base de métricas para melhoria, investigação e intervenção, colete e analise as métricas e valide seu entendimento sobre o sucesso das operações e como elas mudam ao longo do tempo. Use as métricas coletadas para determinar se você está satisfazendo as necessidades do cliente e da empresa e identifique áreas para melhoria.

É necessário um gerenciamento eficiente e eficaz dos eventos operacionais para alcançar a excelência operacional. Isso se aplica a eventos operacionais planejados e não planejados. Use runbooks estabelecidos para eventos bem compreendidos e use manuais para ajudar na investigação e na resolução de problemas. Priorize respostas a eventos com base no impacto nos negócios e no cliente. Assegure que caso um alerta seja gerado em resposta a um evento, exista um processo associado a ser executado com um proprietário especificamente identificado. Defina com antecedência o pessoal necessário para resolver um evento e inclua acionadores de encaminhamento para envolver pessoal adicional, conforme necessário, com base na urgência e no impacto. Identifique e envolva indivíduos com autoridade para tomar uma decisão sobre cursos de ação em que haverá um impacto nos negócios resultante de uma resposta de evento não abordada anteriormente.

Comunique o status operacional das cargas de trabalho por meio de painéis e notificações adaptadas ao público-alvo (por exemplo, cliente, empresa, desenvolvedores, operações) para que eles possam tomar as ações adequadas, para que suas expectativas sejam gerenciadas e para que sejam informados quando as operações normais forem retomadas.

Na AWS, você pode gerar visualizações do painel de suas métricas coletadas de cargas de trabalho e nativamente da AWS. Você pode aproveitar o CloudWatch ou aplicativos de terceiros para agregar e apresentar visualizações em nível de operações de negócios, carga de trabalho e atividades operacionais. A AWS fornece informações sobre a carga de trabalho por meio de recursos de registro em log, incluindo o AWS X-Ray,

CloudWatch, CloudTrail e VPC Flow Logs, permitindo a identificação de problemas de carga de trabalho no suporte à análise e correção da causa raiz.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

OPS 8: Como você compreende a integridade da sua carga de trabalho?

Defina, capture e analise as métricas da carga de trabalho para obter visibilidade destes eventos, para que você possa tomar as ações apropriadas.

OPS 9: Como você compreende a integridade de suas operações?

Defina, capture e analise as métricas de operações para obter visibilidade dos eventos de operações, para que você possa tomar as ações apropriadas.

OPS 10: Como você gerencia os eventos de carga de trabalho e operações?

Prepare e valide procedimentos para responder a eventos, com o objetivo de minimizar a interrupção de sua carga de trabalho.

Todas as métricas coletadas devem estar alinhadas a uma necessidade comercial e aos resultados que elas auxiliam. Desenvolva respostas com script para eventos bem compreendidos e automatize a performance deles em resposta ao reconhecimento do evento.

Evoluir

Você deve aprender, compartilhar e melhorar continuamente para manter a excelência operacional. Dedique ciclos de trabalho para fazer melhorias incrementais contínuas. Execute uma análise pós-incidente de todos os eventos que afetam o cliente. Identifique os fatores que contribuem e a ação preventiva para limitar ou evitar a recorrência. Comunique fatores contribuintes às comunidades afetadas, conforme adequado. Avalie e priorize regularmente oportunidades de melhoria (por exemplo, solicitações de recursos, correção de problemas e requisitos de conformidade), incluindo a carga de trabalho e os procedimentos operacionais. Inclua ciclos de comentários nos procedimentos para identificar rapidamente áreas que podem ser melhoradas e aprender com a execução das operações.

Compartilhe as lições aprendidas entre as equipes para compartilhar os benefícios dessas lições. Analise as tendências nas lições aprendidas e execute análises retrospectivas entre as equipes de métricas de operações para identificar oportunidades e métodos de melhoria. Implemente alterações destinadas a trazer melhorias e avaliar os resultados para determinar o sucesso.

Na AWS, você pode exportar seus dados de log para o Amazon S3 ou enviar logs diretamente para o Amazon S3 para armazenamento de longo prazo. Usando o AWS Glue, você pode descobrir e preparar dados de log no Amazon S3 para estudo analítico, armazenando metadados associados no AWS Glue Data Catalog. O Amazon Athena, por meio da integração nativa com o Glue, pode ser usado para analisar dados de

log, consultando-os com o SQL padrão. Uma ferramenta de inteligência de negócios como o Amazon QuickSight permite visualizar, explorar e analisar dados. Descoberta de tendências e eventos de interesse que podem promover melhorias.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

OPS 11: Como você evolui as operações?
Dedique tempo e recursos para a melhoria incremental contínua, a fim de aumentar a eficácia e a eficiência de suas operações.

A evolução bem-sucedida das operações baseia-se em: pequenas melhorias frequentes; fornecer ambientes seguros e tempo para experimentar, desenvolver e testar melhorias; e ambientes em que o aprendizado com falhas é incentivado. O suporte de operações de ambientes de sandbox, desenvolvimento, teste e produção, com nível crescente de controles operacionais, facilita o desenvolvimento e aumenta a previsibilidade de resultados bem-sucedidos das alterações implementadas na produção.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para (pilar).

Documentação

- DevOps and AWS

Whitepaper

- Operational Excellence Pillar

Vídeo

- DevOps at Amazon

Segurança

O pilar (pilar) inclui (descrição)

O pilar Segurança apresenta uma visão geral dos princípios de design, melhores práticas e perguntas. Você pode encontrar orientações prescritivas sobre implementação no whitepaper sobre o pilar Segurança.

Princípios de design

Existem (contagem) princípios do projeto para (pilar inferior) na nuvem:

- **Implementar uma forte base de identidade**: Implemente o princípio do privilégio mínimo e separe as tarefas com a autorização apropriada para cada interação com os recursos da AWS. Centralize o gerenciamento de identidades e procure eliminar a dependência de credenciais estáticas de longo prazo.
- **Habilitar a rastreabilidade**: Monitore, alerte e audite ações e alterações em seu ambiente em tempo real. Integre a coleta de logs e métricas aos sistemas para investigar e executar ações automaticamente.
- **Aplicar segurança a todas as camadas**: Aplique uma abordagem de defesa detalhada com vários controles de segurança. Aplique a todas as camadas (por exemplo, borda da rede, VPC, balanceamento de carga, cada instância e serviço de computação, sistema operacional, aplicativo e código).
- **Automatizar as melhores práticas de segurança**: Mecanismos de segurança baseados em software automatizados melhoram sua capacidade de ajustar a escala de forma segura, mais rápida e com custos reduzidos. Crie arquiteturas seguras, incluindo a implementação de controles definidos e gerenciados como código em modelos controlados por versão.
- **Proteger dados em trânsito e em repouso**: Classifique seus dados em níveis de sensibilidade e use mecanismos, como criptografia, tokenização e controle de acesso, quando apropriado.
- **Manter as pessoas afastadas dos dados**: Use mecanismos e ferramentas para reduzir ou eliminar a necessidade de acesso direto ou processamento manual de dados. Isso reduz o risco de erros de processamento ou modificação e erro humano ao manipular dados confidenciais.
- **Preparar-se para eventos de segurança**: Prepare-se para um incidente tendo políticas e processos de gerenciamento e investigação de incidentes alinhados aos requisitos organizacionais. Execute simulações de resposta a incidentes e use ferramentas com automação para aumentar sua velocidade de identificação, investigação e recuperação.

Definição

Existem (contagem) melhores práticas para (pilar inferior) na nuvem:

- **Segurança**
- **Identity and Access Management**
- **Detecção**
- **Proteção de infraestrutura**
- **Proteção de dados**

- **Resposta a incidentes**

Antes de projetar qualquer carga de trabalho, estabeleça práticas que influenciem a segurança. Controle quem pode fazer o quê. Além disso, é útil conseguir identificar incidentes de segurança, proteger seus sistemas e serviços e manter a confidencialidade e a integridade dos dados por meio de proteção de dados. Você deve ter um processo bem definido e treinado para responder a incidentes de segurança. Essas ferramentas e técnicas são importantes porque apoiam objetivos como evitar perdas financeiras ou cumprir obrigações regulatórias.

O Modelo de Responsabilidade Compartilhada da AWS permite que as organizações que adotam a nuvem alcancem suas metas de segurança e conformidade. Como a AWS protege fisicamente a infraestrutura que suporta nossos serviços em nuvem, como cliente da AWS, você pode se concentrar no uso de serviços para atingir seus objetivos. A Nuvem AWS também oferece maior acesso aos dados de segurança e uma abordagem automatizada para responder a eventos de segurança.

Melhores práticas

Segurança

Para operar sua carga de trabalho com segurança, você deve aplicar as melhores práticas gerais a todas as áreas de segurança. Use os requisitos e os processos que você definiu em excelência operacional em nível de carga de trabalho e também organizacional e aplique-os a todas as áreas.

Manter-se atualizado com as recomendações da AWS e do setor e a inteligência de ameaças ajuda você a desenvolver seu modelo de ameaças e objetivos de controle. A automação de processos, testes e validação de segurança permite que você escale suas operações de segurança.

As perguntas a seguir se concentram nessas considerações para (pilar inferior). (Para uma lista de perguntas e melhores práticas sobre (pilar inferior), leia o Apêndice.).

SEC 1: Como você opera com segurança sua carga de trabalho?

Para operar sua carga de trabalho com segurança, você deve aplicar as melhores práticas gerais a todas as áreas de segurança. Use os requisitos e os processos que você definiu em excelência operacional em nível de carga de trabalho e também organizacional e aplique-os a todas as áreas. Manter-se atualizado com as recomendações da AWS e do setor e a inteligência de ameaças ajuda você a desenvolver seu modelo de ameaças e objetivos de controle. A automação de processos, testes e validação de segurança permite que você escale suas operações de segurança.

Na AWS, a segregação de cargas de trabalho diferentes por conta, com base na respectiva função e nos requisitos de conformidade ou confidencialidade de dados, é uma abordagem recomendada.

Identity and Access Management

O Identity and Access Management é parte essencial de um programa de segurança da informação, que garante que apenas usuários autorizados e autenticados possam acessar seus recursos e somente da forma que você pretender. Por exemplo, você deve definir entidades principais (ou seja, contas, usuários, funções e serviços que podem executar ações em sua conta), criar políticas alinhadas com essas entidades principais e implementar um gerenciamento forte de credenciais. Esses elementos de gerenciamento de privilégios formam o núcleo da autenticação e autorização.

Na AWS, o gerenciamento de privilégios é compatível principalmente com o serviço AWS Identity and Access Management (IAM), que permite controlar o acesso do usuário e programático a produtos e recursos da AWS. Você deve aplicar políticas granulares, que atribuem permissões a um usuário, grupo, função ou recurso. Você também pode exigir práticas de senha forte, como nível de complexidade, evitando reutilização e impondo multi-factor authentication (MFA). Você pode usar federação com seu serviço de diretório atual. Para cargas de trabalho que exigem que os sistemas tenham acesso à AWS, o IAM possibilita acesso seguro por meio de funções, perfis de instância, federação de identidades e credenciais temporárias.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

SEC 2: Como você gerencia identidades para pessoas e máquinas?

Há dois tipos de identidades que você precisa gerenciar para operar cargas de trabalho seguras da AWS. Entender o tipo de identidade de que você precisa para gerenciar e conceder acesso ajuda a garantir que as identidades corretas tenham acesso aos recursos certos nas condições certas. Identidades humanas: administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações da AWS. Eles são membros da sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da web, aplicação cliente ou ferramentas interativas de linha de comando. Identidades de máquina: aplicações de serviço, ferramentas operacionais e cargas de trabalho precisam de uma identidade para solicitar serviços da AWS; por exemplo, para ler dados. Essas identidades incluem máquinas em execução no seu ambiente da AWS, como instâncias do Amazon EC2 ou funções do AWS Lambda. Você também pode gerenciar identidades de máquina para partes externas que precisam de acesso. Além disso, você pode ter máquinas fora da AWS que precisam de acesso ao seu ambiente da AWS.

SEC 3: Como você gerencia permissões para pessoas e máquinas?

Gerencie permissões para controlar o acesso a identidades de pessoas e máquinas que precisam de acesso à AWS e à sua carga de trabalho. As permissões controlam quem pode acessar o quê e em quais condições.

As credenciais não devem ser compartilhadas entre usuários ou sistemas. O acesso do usuário deve ser concedido usando uma abordagem de privilégio mínimo, com melhores práticas que incluem requisitos de senha e imposição de MFA. O acesso programático, incluindo chamadas à API a produtos da AWS, deve ser realizado usando cre-

denciais de privilégio limitado e temporárias como aquelas emitidas pelo AWS Security Token Service.

A AWS fornece recursos que podem ajudá-lo no Identity and Access Management. Para conhecer as melhores práticas, verifique nossos experimentos práticos sobre gerenciamento de credenciais e autenticação, controle de acesso humano e controle de acesso programático.

Detecção

Você pode usar controles de detecção para identificar uma potencial ameaça ou incidente de segurança. Eles são uma parte essencial das estruturas de governança e podem ser usados para apoiar um processo de qualidade, uma obrigação legal ou de conformidade e para os esforços de identificação e resposta a ameaças. Existem diferentes tipos de controles de detecção. Por exemplo, a realização de um inventário de ativos e seus atributos detalhados promove tomadas de decisão mais eficazes (e controles de ciclo de vida) para ajudar a estabelecer linhas de base operacionais. Você também pode usar a auditoria interna, um exame dos controles relacionados aos sistemas de informação, para garantir que as práticas atendam às políticas e aos requisitos e que você tenha definido as notificações de alerta automatizadas corretas com base nas condições definidas. Esses controles são fatores reativos importantes que podem ajudar sua organização a identificar e entender o escopo da atividade anômala.

Na AWS, você pode implementar controles de detecção processando logs, eventos e monitoramento que possibilitam auditoria, análise automatizada e alarmes. Os logs do CloudTrail, as chamadas à API da AWS e o CloudWatch fornecem o monitoramento de métricas com alarmes, enquanto o AWS Config fornece o histórico de configuração. O Amazon GuardDuty é um serviço de detecção de ameaças gerenciado que monitora continuamente comportamentos mal-intencionados ou não autorizados para ajudar a proteger contas e cargas de trabalho da AWS. Logs em nível de serviço também estão disponíveis, por exemplo, você pode usar o Amazon Simple Storage Service (Amazon S3) para registrar solicitações de acesso em log.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

SEC 4: Como você detecta e investiga eventos de segurança?
Capture e analise eventos de logs e métricas para gerar visibilidade. Tome medidas em eventos de segurança e potenciais ameaças para ajudar a proteger sua carga de trabalho.

O gerenciamento de log é importante para uma carga de trabalho do Well-Architected por motivos que vão de segurança ou análise forense a requisitos regulatórios ou legais. É fundamental que você analise os logs e responda a eles para que possa identificar possíveis incidentes de segurança. A AWS fornece uma funcionalidade que torna o gerenciamento de log mais fácil de implementar possibilitando que você defina um ciclo de vida de retenção de dados ou defina em que local os dados serão preser-

vados, arquivados ou, por fim, excluídos. Isso torna o processamento de dados previsível e confiável mais simples e econômico.

Proteção de infraestrutura

A proteção de infraestrutura abrange metodologias de controle, como defesa em profundidade, necessárias para atender às melhores práticas e obrigações organizacionais ou regulatórias. O uso dessas metodologias é fundamental para operações contínuas bem-sucedidas na nuvem ou no local.

Na AWS, é possível implementar inspeção de pacote stateful e stateless, seja usando tecnologias nativas da AWS ou produtos e serviços de parceiros disponíveis por meio do AWS Marketplace. Você deve usar o Amazon Virtual Private Cloud (Amazon VPC) para criar um ambiente privado, protegido e escalável em que seja possível definir sua topologia, incluindo gateways, tabelas de roteamento e sub-redes públicas e privadas.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

SEC 5: Como você protege seus recursos de rede?

Qualquer carga de trabalho que tenha alguma forma de conectividade de rede, seja a Internet ou uma rede privada, exige várias camadas de defesa para ajudar a proteger contra ameaças externas e internas baseadas em rede.

SEC 6: Como você protege seus recursos de computação?

Os recursos de computação exigem várias camadas de defesa para ajudar na proteção contra ameaças externas e internas. Os recursos de computação incluem instâncias do EC2, contêineres, funções do AWS Lambda, serviços de banco de dados, dispositivos de IoT e muito mais.

É aconselhável usar várias camadas de defesa em qualquer tipo de ambiente. No caso de proteção de infraestrutura, muitos dos conceitos e métodos são válidos em modelos no local e em nuvem. Impor proteção de limites, monitorar pontos de entrada e saída e registro em log, monitoramento e geração de alertas abrangentes são medidas essenciais para um plano eficaz de segurança da informação.

Os clientes da AWS são capazes de personalizar, ou reforçar, a configuração de uma Amazon Elastic Compute Cloud (Amazon EC2), de um contêiner do Amazon EC2 Container Service (Amazon ECS) ou de uma instância do AWS Elastic Beanstalk, além de manter essa configuração em uma imagem de máquina da Amazon (AMI) imutável. Ao serem acionados pelo Auto Scaling ou iniciados manualmente, todos os novos servidores virtuais (instâncias) iniciados com esse AMI recebem a configuração reforçada.

Proteção de dados

Antes de criar a arquitetura de qualquer sistema, devem ser adotadas práticas fundamentais que influenciam a segurança. Por exemplo, a classificação de dados fornece

uma maneira de categorizar os dados organizacionais com base nos níveis de sensibilidade, e a criptografia protege os dados ao torná-los ininteligíveis ao acesso não autorizado. Essas ferramentas e técnicas são importantes porque apoiam objetivos como evitar perdas financeiras ou cumprir obrigações regulatórias.

Na AWS, as seguintes práticas facilitam a proteção de dados:

- como cliente da AWS, você mantém controle total sobre seus dados.
- A AWS torna mais fácil criptografar e gerenciar chaves, incluindo a rotação regular de chaves, que pode ser facilmente automatizada pela AWS ou mantida por você.
- O registro em log detalhado com conteúdo importante, como acesso e alterações a arquivo, está disponível.
- A AWS projetou sistemas de armazenamento para resiliência excepcional. Por exemplo, o Amazon S3 Standard, o S3 Standard–IA, o S3 One Zone-IA e o Amazon Glacier são todos projetados para oferecer 99,999999999% de durabilidade de objetos em determinado ano. Esse nível de durabilidade corresponde a uma perda anual média esperada de 0,000000001% dos objetos.
- O versionamento, que pode fazer parte de um processo maior de gerenciamento de ciclo de vida de dados, pode proteger contra substituições, exclusões e danos similares inadvertidos.
- A AWS nunca inicia a movimentação de dados entre regiões. O conteúdo colocado em uma região permanecerá nessa região, a menos que você explicitamente habilite um recurso ou utilize um serviço que forneça essa funcionalidade.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

SEC 7: Como classificar meus dados?
A classificação serve para categorizar os dados com base em criticidade e confidencialidade para ajudá-lo a determinar os controles de proteção e retenção apropriados.
SEC 8: Como você protege seus dados em repouso?
Proteja seus dados em repouso implementando vários controles para reduzir o risco de acesso não autorizado ou manuseio incorreto.
SEC 9: Como você protege seus dados em trânsito?
Proteja seus dados em trânsito implementando vários controles para reduzir o risco de acesso não autorizado ou perda.

A AWS oferece vários meios de criptografar dados em repouso e em trânsito. Integramos recursos em nossos serviços que tornam mais fácil criptografar seus dados. Por exemplo, implementamos criptografia no lado do servidor (SSE) para o Amazon S3 para tornar mais fácil para você armazenar seus dados em um formato criptografado. Você também pode providenciar que todo o processo de criptografia e descriptografia

HTTPS (geralmente conhecido como terminação SSL) seja processado por Elastic Load Balancing (ELB).

Resposta a incidentes

Mesmo com controles preventivos e de detecção consolidados, sua organização ainda deve implementar processos para responder e mitigar o impacto potencial de incidentes de segurança. A arquitetura de sua carga de trabalho afeta fortemente a capacidade de suas equipes de operar efetivamente durante um incidente, de isolar ou conter sistemas e de restaurar operações para um bom estado conhecido. Ter as ferramentas e o acesso prontos antes de um incidente de segurança e praticar rotineiramente a resposta a incidentes durante os dias de jogo ajudará a garantir que sua arquitetura possa acomodar investigações e recuperação oportunas.

Na AWS, as seguintes práticas facilitam a resposta eficaz a incidentes:

- o registro em log detalhado está disponível e contém conteúdo importante, como acesso a arquivos e alterações.
- Os eventos podem ser processados automaticamente e acionar ferramentas que automatizam respostas usando as APIs da AWS.
- Você pode pré-provisionar ferramentas e uma "sala limpa" usando o AWS CloudFormation. Isso permite que você realize análise forense em um ambiente seguro e isolado.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

SEC 10: Como você prevê, responde e se recupera de incidentes?
A preparação é essencial para investigação, resposta e recuperação oportunas e eficazes de incidentes de segurança para ajudar a minimizar interrupções na sua organização.

Garanta acesso rápido de sua equipe de segurança e automatize o isolamento de instâncias, bem como a captura de dados e estado para análise forense.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para (pilar).

Documentação

- AWS Cloud Security
- AWS Compliance
- AWS Security Blog

Whitepaper

- Security Pillar
- AWS Security Overview
- AWS Security Best Practices
- AWS Risk and Compliance

Vídeo

- AWS Security State of the Union
- Shared Responsibility Overview

Confiabilidade

O pilar (pilar) inclui (descrição)

O pilar Confiabilidade apresenta uma visão geral dos princípios de design, das melhores práticas e das perguntas. Você encontra orientações prescritivas sobre implementação no whitepaper Pilar Confiabilidade.

Princípios de design

Existem (contagem) princípios do projeto para (pilar inferior) na nuvem:

- **Recuperação automática de falhas**: Ao monitorar os Key Performance Indicators (KPIs – Indicadores-chave de performance) de uma carga de trabalho, você pode acionar a automação quando um limite é ultrapassado. Esses KPIs devem ser uma medida do valor empresarial, não dos aspectos técnicos da operação do serviço. Isso permite a notificação automática e o rastreamento de falhas, além de processos de recuperação automatizados que solucionam ou reparam a falha. Com uma automação mais sofisticada, é possível antecipar e corrigir falhas antes que elas ocorram.
- **Teste os procedimentos de recuperação**: Em um ambiente no local, geralmente realiza-se o teste para provar que a carga de trabalho funciona em um cenário específico. Normalmente, o teste não é usado para validar estratégias de recuperação. Na nuvem, você pode testar o comportamento de falha da carga de trabalho e validar os procedimentos de recuperação. É possível usar a automação para simular falhas diferentes ou para recriar cenários que levaram a falhas no passado. Essa abordagem expõe caminhos de falha que você pode testar e corrigir antes que ocorra um cenário de falha real, o que reduz os riscos.
- **Escale horizontalmente para aumentar a disponibilidade agregada da carga de trabalho**: Substitua um recurso grande por vários recursos pequenos para reduzir o

impacto de uma única falha na carga de trabalho geral. Distribua as solicitações por vários recursos menores para garantir que elas não compartilhem um ponto de falha comum.

- **Pare de tentar adivinhar sua capacidade**: Uma causa comum de falha nas cargas de trabalho no local é a saturação de recursos, quando as demandas impostas a uma carga de trabalho excedem a capacidade dela. Geralmente, esse é o objetivo dos ataques de negação de serviço. Na nuvem, você pode monitorar a demanda e a utilização da carga de trabalho e automatizar a adição ou a remoção de recursos para manter o nível ideal e atender à demanda, sem aprovisionamento em excesso ou subprovisionamento. Ainda há limites, mas algumas cotas podem ser controladas e outras podem ser gerenciadas. Consulte Gerencie cotas e restrições de serviço.
- **Gerencie as alterações na automação**: As alterações na sua infraestrutura devem ser feitas por meio de automação. Dentre aquelas que precisam ser gerenciadas estão as alterações na automação, que podem ser acompanhadas e analisadas.

Definição

Existem (contagem) melhores práticas para (pilar inferior) na nuvem:

- **Fundamentos**
- **Arquitetura da carga de trabalho**
- **Gerenciamento de alterações**
- **Gerenciamento de falhas**

Para atingir a confiabilidade, você deve começar com as bases: um ambiente em que as cotas de serviço e a topologia de rede acomodam a carga de trabalho. A arquitetura da carga de trabalho do sistema distribuído deve ser projetada para evitar e mitigar falhas. A carga de trabalho deve processar as alterações na demanda ou nos requisitos e ser projetada para detectar falhas e se reparar automaticamente.

Melhores práticas

Fundamentos

Os requisitos fundamentais são aqueles que têm um escopo que vai além de uma única carga de trabalho ou projeto. Antes de criar a arquitetura de um sistema, é necessário instaurar os requisitos fundamentais que influenciam a confiabilidade. Por exemplo, você deve ter largura de banda de rede suficiente no datacenter.

Com a AWS, a maioria desses requisitos fundamentais já está incorporada ou pode ser tratada conforme necessário. A nuvem foi projetada para ser praticamente ilimitada,

portanto, é responsabilidade da AWS atender ao requisito de capacidade suficiente de rede e de computação, deixando você livre para alterar o tamanho e as alocações de recursos sob demanda.

As perguntas a seguir se concentram nessas considerações para (pilar inferior). (Para uma lista de perguntas e melhores práticas sobre (pilar inferior), leia o Apêndice.).

REL 1: Como você gerencia as cotas e restrições de serviço?

Para arquiteturas de carga de trabalho baseadas na nuvem, há cotas de serviço, que também são conhecidas como limites de serviço. Essas cotas existem para evitar o aprovisionamento acidental de mais recursos do que o necessário e para limitar as taxas de solicitação nas operações de API para proteger os serviços contra abuso. Há também restrições de recursos, por exemplo, a taxa de envio de bits por um cabo de fibra óptica ou a quantidade de armazenamento em um disco físico.

REL 2: Como você planeja sua topologia de rede?

Muitas vezes, as cargas de trabalho estão presentes em vários ambientes. Dentre eles estão vários ambientes de nuvem (acessíveis publicamente e privados) e possivelmente sua infraestrutura de datacenter existente. Os planos devem incluir considerações de rede, como conectividade dentro dos sistemas e entre eles, gerenciamento de endereços IP públicos e privados e resolução de nomes de domínio.

Para arquiteturas de carga de trabalho baseadas na nuvem, há cotas de serviço, que também são conhecidas como limites de serviço. Essas cotas existem para evitar o aprovisionamento acidental de mais recursos do que o necessário e para limitar as taxas de solicitação em operações de API para proteger os serviços contra abuso. Muitas vezes, as cargas de trabalho estão presentes em vários ambientes. Você deve monitorar e gerenciar essas cotas para todos os ambientes de carga de trabalho. Eles incluem vários ambientes de nuvem (com acesso tanto público quanto privado) e podem incluir sua infraestrutura de datacenter existente. Os planos devem incluir considerações de rede, como conectividade dentro dos sistemas e entre eles, gerenciamento de endereços IP públicos e privados e resolução de nomes de domínio.

Arquitetura da carga de trabalho

Uma carga de trabalho confiável começa com decisões iniciais de projeto que envolvem tanto o software quanto a infraestrutura. Suas decisões de arquitetura afetarão o comportamento da carga de trabalho em todos os cinco pilares do Well-Architected. Para atingir a confiabilidade, há padrões específicos que você deve seguir.

Com a AWS, os desenvolvedores de carga de trabalho podem usar as linguagens e tecnologias que preferem. Os SDKs da AWS eliminam a complexidade da codificação por meio de APIs específicas à linguagem para os serviços da AWS. Esses SDKs e a possibilidade de escolher a linguagem permitem que os desenvolvedores implementem as melhores práticas de confiabilidade apresentadas neste documento. Os desen-

volvedores também podem ler e descobrir como a Amazon cria e opera softwares na Amazon Builders' Library.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

REL 3: Como você projeta sua arquitetura de serviços de carga de trabalho?
Use uma Service-Oriented Architecture (SOA – Arquitetura orientada por serviços) ou uma arquitetura de microsserviços para criar cargas de trabalho altamente escaláveis e confiáveis. A SOA é a prática de tornar componentes de software reutilizáveis por meio de interfaces de serviço. A arquitetura de microsserviços vai além para tornar os componentes menores e mais simples.
REL 4: Como você projeta interações em um sistema distribuído para evitar falhas?
Os sistemas distribuídos dependem das redes de comunicação para interconectar componentes, como servidores ou serviços. Sua carga de trabalho deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar sem afetar negativamente outros componentes ou a carga de trabalho. Essas melhores práticas evitam falhas e melhoram o Mean Time Between Failures (MTBF – Tempo médio entre falhas).
REL 5: Como você projeta interações em um sistema distribuído para mitigar ou resistir a falhas?
Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes (como servidores ou serviços). Sua carga de trabalho deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar sem afetar negativamente outros componentes ou a carga de trabalho. Essas melhores práticas permitem que as cargas de trabalho resistam a tensões ou falhas, recuperem-se mais rapidamente delas e reduzam o impacto de tais prejuízos. Como resultado, o Mean Time To Recovery (MTTR – Tempo médio até a recuperação) é melhorado.

Os sistemas distribuídos dependem das redes de comunicação para interconectar componentes, como servidores ou serviços. Sua carga de trabalho deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar sem afetar negativamente outros componentes ou a carga de trabalho.

Gerenciamento de alterações

As alterações na carga de trabalho ou no ambiente dela devem ser previstas e acomodadas para alcançar uma operação confiável da carga de trabalho. As alterações incluem aquelas impostas à sua carga de trabalho, como picos na demanda, bem como aquelas internas, como implantações de recursos e patches de segurança.

Por meio da AWS, você pode monitorar o comportamento de uma carga de trabalho e automatizar a resposta aos KPIs. Por exemplo, a carga de trabalho pode adicionar outros servidores à medida que recebe mais usuários. Você pode controlar quem tem permissão para fazer alterações na carga de trabalho e realizar auditorias no histórico dessas alterações.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

REL 6: Como você monitora recursos de carga de trabalho?

Os logs e as métricas são uma ferramenta poderosa para saber a integridade das suas cargas de trabalho. Você pode configurar sua carga de trabalho para monitorar logs e métricas e enviar notificações quando os limites forem ultrapassados ou em caso de eventos importantes. O monitoramento permite que sua carga de trabalho reconheça quando os limites de baixa performance são ultrapassados ou quando há falhas, para que ela possa se recuperar automaticamente em resposta.

REL 7: Como você projeta sua carga de trabalho para se adaptar às mudanças na demanda?

Uma carga de trabalho escalável oferece elasticidade para adicionar ou remover recursos automaticamente para que atendam melhor à demanda atual a qualquer momento.

REL 8: Como você implementa uma alteração?

As alterações controladas são necessárias para implantar novas funcionalidades e garantir que as cargas de trabalho e o ambiente operacional executem softwares conhecidos e possam ser corrigidos ou substituídos de maneira previsível. Se essas alterações forem descontroladas, será difícil prever o efeito ou resolver problemas decorrentes delas.

Quando você cria a arquitetura de uma carga de trabalho para adicionar e remover recursos automaticamente em resposta às alterações na demanda, isso não apenas aumenta a confiabilidade, mas também garante que o sucesso nos negócios não se torne um fardo. Com o monitoramento implantado, sua equipe será automaticamente alertada quando os KPIs se desviarem das normas esperadas. O registro automático de alterações em seu ambiente permite realizar auditorias e identificar rapidamente as ações que podem ter afetado a confiabilidade. Os controles do gerenciamento de alterações garantem que você possa impor as regras que oferecem a confiabilidade necessária.

Gerenciamento de falhas

Em qualquer sistema de complexidade razoável, espera-se que ocorram falhas. A confiabilidade exige que sua carga de trabalho reconheça as falhas no momento em que elas ocorrem e tome medidas para evitar que elas prejudiquem a disponibilidade. As cargas de trabalho devem ser capazes de resistir a falhas e reparar problemas automaticamente.

Com a AWS, você pode aproveitar a automação para reagir aos dados de monitoramento. Por exemplo, quando uma métrica específica ultrapassa um limite, você pode acionar uma ação automatizada para solucionar o problema. Além disso, em vez de tentar diagnosticar e corrigir um recurso com falha que faz parte do seu ambiente de produção, você pode substituí-lo por um novo e executar a análise do recurso com falha fora de banda. Como a nuvem permite que você suporte versões temporárias de um sistema inteiro a baixo custo, é possível usar testes automatizados para verificar os processos de recuperação completos.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

REL 9: Como você faz backup dos dados?
Faça backup de dados, aplicativos e configurações para atender aos seus requisitos de Recovery Time Objective (RTO – Objetivo do tempo de recuperação) e de Recovery Point Objective (RPO – Objetivo do ponto de recuperação).
REL 10: Como usar o isolamento de falhas para proteger sua carga de trabalho?
Os limites isolados de falhas restringem o efeito de uma falha em uma carga de trabalho a um número controlado de componentes. A falha não afeta os componentes fora do limite. Ao usar vários limites isolados de falhas, você pode restringir o impacto sobre sua carga de trabalho.
REL 11: Como você projeta sua carga de trabalho para resistir a falhas de componentes?
As cargas de trabalho que exigem alta disponibilidade e baixo Mean Time To Recovery (MTTR – Tempo médio até a recuperação) devem ser projetadas visando a resiliência.
REL 12: Como testar a confiabilidade?
Depois de projetar sua carga de trabalho para resiliência à pressão da produção, o teste é a única maneira de garantir que ela opere conforme projetado e com a resiliência esperada.
REL 13: Como você planeja a recuperação de desastres (DR)?
Implementar backups e componentes redundantes de carga de trabalho é o ponto de partida da sua estratégia de DR. O RTO e o RPO são os objetivos para restaurar a disponibilidade. Defina-os de acordo com suas necessidades de negócios. Implemente uma estratégia para atender a esses objetivos, considerando os locais e a função dos recursos e dos dados da carga de trabalho.

Faça backup regular dos seus dados e teste seus arquivos de backup para garantir a recuperação de erros tanto físicos quanto lógicos. Para gerenciar falhas, é essencial testar as cargas de trabalho com frequência e de maneira automatizada por meio da indução de falhas e da observação do processo de recuperação. Faça isso periodicamente e também após alterações significativas na carga de trabalho. Acompanhe ativamente os KPIs, como Recovery Time Objective (RTO – Objetivo do tempo de recuperação) e Recovery Point Objective (RPO – Objetivo do ponto de recuperação), para avaliar a resiliência de uma carga de trabalho, principalmente em cenários de teste de falhas. O acompanhamento dos KPIs ajudará você a identificar e mitigar os pontos únicos de falha. O objetivo é testar integralmente os processos de recuperação da carga de trabalho para ter certeza de que você pode recuperar todos os seus dados e continuar a atender os clientes, mesmo diante de problemas contínuos. Seus processos de recuperação devem ser tão bem trabalhados quanto os processos de produção normais.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para (pilar).

Documentação

- AWS Documentation
- AWS Global Infrastructure
- AWS Auto Scaling: How Scaling Plans Work
- What Is AWS Backup?

Whitepaper

- Reliability Pillar: AWS Well-Architected
- Implementing Microservices on AWS

Eficiência de performance

O pilar (pilar) inclui (descrição)

O pilar Eficiência de performance fornece uma visão geral dos princípios, melhores práticas e perguntas atinentes ao projeto. Você encontra orientações prescritivas sobre implementação no whitepaper Pilar Eficiência de performance.

Princípios de design

Existem (contagem) princípios do projeto para (pilar inferior) na nuvem:

- **Democratizar 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 sobre como 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.
- **Tornar-se global em minutos**: A implantação de sua carga de trabalho em várias regiões da AWS em todo o mundo permite oferecer menor latência e uma melhor experiência para seus clientes a um custo mínimo.
- **Usar 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 da web) e os serviços de eventos podem hospedar o código. Isso elimina o fardo operacional do ge-

renciamento de servidores físicos e pode reduzir os custos transacionais, pois os serviços gerenciados operam em escala de nuvem.

- **Experimentar com mais frequência**: Com recursos virtuais e automatizáveis, você pode executar rapidamente testes comparativos usando diferentes tipos de instâncias, armazenamento ou configurações.
- **Considere a simpatia mecânico**: Entenda como os serviços de nuvem são consumidos e use sempre a abordagem tecnológica mais alinhada às suas metas de carga de trabalho. Por exemplo, avalie padrões de acesso a dados ao selecionar abordagens de banco de dados ou armazenamento.

Definição

Existem (contagem) melhores práticas para (pilar inferior) na nuvem:

- **Seleção**
- **Análise**
- **Monitoramento**
- **Concessões**

Adote uma abordagem impulsionada por 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.

A avaliação periódica de suas escolhas garante que você esteja aproveitando a 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.

Melhores práticas

Seleção

A solução ideal para uma carga de trabalho específica varia e, muitas vezes, as soluções combinam várias abordagens. Cargas de trabalho bem arquitetadas usam várias soluções e habilitam 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 necessidades da sua carga de trabalho. Você também pode encontrar opções que não são facilmente obtidas com infraestrutura no local. 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.

As perguntas a seguir se concentram nessas considerações para (pilar inferior). (Para uma lista de perguntas e melhores práticas sobre (pilar inferior), leia o Apêndice.).

PERF 1: Como você seleciona a arquitetura de melhor performance?
Muitas vezes, é necessário empregar várias abordagens para obter a performance ideal em uma carga de trabalho. Os sistemas com boa arquitetura usam várias soluções e recursos para aprimorar a performance.

Use uma abordagem impulsionada por dados para selecionar os padrões e a implementação de sua arquitetura e, por fim, obter uma solução econômica. Os arquitetos de soluções da AWS, as arquiteturas de referência da AWS e os parceiros da Rede de parceiros da AWS (APN) podem ajudá-lo a selecionar uma arquitetura com base em conhecimento do setor, mas os dados obtidos por meio de benchmarking ou teste de carga serão necessários para otimizar sua arquitetura.

Sua arquitetura provavelmente combinará várias abordagens arquiteturais diferentes (por exemplo, orientada por eventos, ETL ou pipeline). A implementação de sua arquitetura usará os serviços da AWS que são específicos para a otimização da performance de sua arquitetura. Nas seções a seguir, analisamos os quatro principais tipos de recursos que você deve levar em consideração (computação, armazenamento, banco de dados e rede).

Computação

Selecionar recursos computacionais que atendam aos seus requisitos, necessidades de performance e fornecem grande eficiência de custo e esforço permitirá que você faça mais com o mesmo número de recursos. Ao avaliar opções de computação, esteja ciente dos requisitos de performance e custo da carga de trabalho e use isso para tomar decisões bem embasadas.

Na AWS, a computação está disponível de três formas: instâncias, contêineres e funções: as

- **instâncias** são servidores virtualizados, permitindo que você altere seus recursos com um botão ou uma chamada de API. Como as decisões de recursos na nuvem não são imutáveis, você pode testar diferentes tipos de servidores. Na AWS, essas instâncias de servidor virtual vêm em diferentes famílias e tamanhos e oferecem uma ampla variedade de capacidades, inclusive Solid-State Drives (SSD – Unidade de estado sólido) e Graphics Processing Units (GPU – Unidades de processamento gráfico). Os
- **contêineres** são um método de virtualização do sistema operacional que permite executar um aplicativo e suas dependências em processos isolados por recursos. O

AWS Fargate é um serviço de computação sem servidor para contêineres, ou também é possível usar o Amazon EC2 se você precisar de controle sobre a instalação, a configuração e o gerenciamento do seu ambiente de computação. Você também pode escolher entre várias plataformas de orquestração de contêineres: Amazon Elastic Container Service (ECS) ou Amazon Elastic Kubernetes Service (EKS). As

- **funções** abstraem o ambiente de execução do código que você deseja executar. Por exemplo, o AWS Lambda permite que você execute código sem executar uma instância.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 2: Como você seleciona sua solução de computação?
A solução de computação ideal para uma carga de trabalho varia conforme o design do aplicativo, os padrões de uso e as definições de configuração. As arquiteturas podem usar diferentes soluções de computação para vários componentes e podem habilitar diferentes recursos para melhorar a performance. Selecionar a solução de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Ao arquitetar o uso da computação, você deve aproveitar os mecanismos de elasticidade disponíveis para garantir que você tenha capacidade suficiente para sustentar a performance conforme a demanda muda.

Armazenamento

O armazenamento na nuvem é um componente essencial da computação em nuvem e mantêm as informações usadas pela sua carga de trabalho. Geralmente, o armazenamento na nuvem é mais confiável, escalável e seguro do que sistemas de armazenamento tradicionais no local. Escolha entre serviços de armazenamento de objetos, blocos e arquivos, bem como opções de migração de dados para a nuvem para sua carga de trabalho.

Na AWS, o armazenamento está disponível de três formas: objeto, bloco e arquivo: o

- **Armazenamento de objetos** fornece uma plataforma escalável e durável para tornar os dados acessíveis a partir de qualquer local da Internet para conteúdo gerado pelo usuário, arquivamento ativo, computação sem servidor, armazenamento de big data ou backup e recuperação. O Amazon Simple Storage Service (Amazon S3) é um serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e performance líderes do setor. O Amazon S3 foi projetado para oferecer 99,999999999% (11 noves) de durabilidade e armazena dados para milhões de aplicativos para empresas de todo o mundo. O
- **Armazenamento em bloco** oferece armazenamento em bloco altamente disponível, consistente e de baixa latência para cada host virtual e é semelhante ao armazenamento de conexão direta (DAS) ou a uma rede de área de armazenamento (SAN). O Amazon Elastic Block Store (Amazon EBS) foi projetado para cargas de

trabalho que exigem armazenamento persistente acessível por instâncias do EC2, o que ajuda você a ajustar aplicativos com o custo, a performance e a capacidade de armazenamento corretos. O

- **Armazenamento de arquivos** fornece acesso a um sistema de arquivos compartilhado entre vários sistemas. Soluções de armazenamento de arquivos, como o Amazon Elastic File System (EFS), ou são ideais para casos de uso como grandes repositórios de conteúdo, ambientes de desenvolvimento, armazenamentos de mídia ou diretórios iniciais de usuários. O Amazon FSx torna fácil e econômico iniciar e executar sistemas de arquivos populares para que você possa aproveitar os sofisticados conjuntos de recursos e a rápida performance de sistemas de arquivos de código aberto amplamente utilizados e licenciados comercialmente.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 3: Como você seleciona sua solução de armazenamento?

A solução de armazenamento ideal para um sistema varia conforme o tipo de método de acesso (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o rendimento necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. Os sistemas Well-Architected usam várias soluções de armazenamento e habilitam diferentes recursos para melhorar a performance e usar os recursos de modo eficiente.

Quando você seleciona uma solução de armazenamento, garantir que ela se alinhe com seus padrões de acesso será fundamental para alcançar a performance desejada.

Banco de dados

A nuvem oferece serviços de banco de dados específicos que abordam diferentes problemas apresentados por sua carga de trabalho. Você pode escolher entre vários mecanismos de banco de dados de finalidade específica, inclusive bancos de dados relacionais, de chave-valor, documentos, em memória, gráficos, séries temporais e livros contábeis. Ao escolher o melhor banco de dados para resolver um problema específico (ou um grupo de problemas), você pode se libertar de bancos de dados monolíticos genéricos restritivos e se concentrar na criação de aplicativos para atender às necessidades de performance dos seus clientes.

Na AWS, você pode escolher entre vários mecanismos de banco de dados de finalidade específica, inclusive bancos de dados relacionais, de chave-valor, documentos, em memória, gráficos, séries temporais e livros contábeis. Com os bancos de dados da AWS, você não precisa se preocupar com tarefas de gerenciamento de banco de dados, como provisionamento, aplicação de patches, instalação, configuração, backups ou recuperação de servidores. A AWS monitora continuamente seus clusters para manter suas cargas de trabalho funcionando com armazenamento com autorreparação e escalabilidade automatizada, para que você possa se concentrar no desenvolvimento de aplicativos de maior valor.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 4: Como você seleciona sua solução de banco de dados?

A solução de banco de dados ideal para um sistema varia conforme os requisitos de disponibilidade, consistência, tolerância da partição, latência, durabilidade, escalabilidade e capacidade de consulta. Muitos sistemas usam soluções de banco de dados diferentes para vários subsistemas e habilitam diferentes recursos para melhorar a performance. Selecionar a solução e os recursos de banco de dados incorretos para um sistema pode levar a uma menor eficiência.

A abordagem de banco de dados da carga de trabalho tem um impacto significativo na eficiência da performance. Muitas vezes, é uma área escolhida de acordo com padrões organizacionais, em vez de por meio de uma abordagem orientada por dados. Assim como no armazenamento, é essencial considerar os padrões de acesso da sua carga de trabalho e também se outras soluções que não são de banco de dados podem resolver o problema com mais eficiência (como usar gráficos, séries temporais ou um mecanismo de pesquisa ou banco de dados de armazenamento na memória).

Rede

Como a rede está entre todos os componentes da carga de trabalho, ela pode ter grandes impactos positivos e negativos sobre a performance e o comportamento da carga de trabalho. Também há cargas de trabalho que são altamente dependentes da performance da rede, como Computação de Alta Performance (HPC), para a qual é importante ter um entendimento profundo da rede a fim de aumentar a performance do cluster. É necessário determinar os requisitos de largura de banda, latência, instabilidade e throughput da carga de trabalho.

Na AWS, as redes são virtualizadas e estão disponíveis em vários tipos e configurações diferentes. Isso facilita fazer a correspondência entre os métodos de rede e suas necessidades. A AWS oferece recursos do produto (por exemplo, Rede aprimorada, instâncias otimizadas do Amazon EBS, Amazon S3 Transfer Acceleration e Amazon CloudFront dinâmico) para otimizar o tráfego da rede. A AWS também oferece recursos de rede (p. ex., roteamento de latência do Amazon Route 53, Amazon VPC endpoints, AWS Direct Connect e AWS Global Accelerator) para reduzir a distância ou a instabilidade da rede.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 5: Como você configura sua solução de redes?

A solução de rede ideal para uma carga de trabalho varia com base nos requisitos de latência, throughput, instabilidade e largura de banda. Restrições físicas, como recursos de usuário ou no local, determinam as opções de localização. Essas restrições podem ser compensadas com pontos de presença ou posicionamento de recursos.

Você deve considerar o local ao implantar sua rede e pode optar por colocar os recursos perto de onde eles serão usados para reduzir a distância. Use métricas de rede pa-

ra fazer alterações na configuração de rede conforme a carga de trabalho evolui. Ao aproveitar as regiões, grupos de canais e serviços de borda, você pode melhorar significativamente a performance. É possível recriar ou modificar as redes baseadas na nuvem rapidamente, portanto, é necessário evoluir sua arquitetura de rede ao longo do tempo para manter a eficiência da performance.

Análise

As tecnologias de nuvem evoluem rapidamente e você deve garantir que os componentes da carga de trabalho estejam usando novas tecnologias e abordagens para melhorar continuamente a performance. Você deve avaliar e considerar continuamente alterações nos componentes da carga de trabalho para garantir que está cumprindo seus objetivos de performance e custo. As novas tecnologias, como Machine Learning e inteligência artificial (IA), podem permitir que você reimagine as experiências do cliente e realize inovações em todas as cargas de trabalho de negócios.

Aproveite a inovação contínua na AWS, impulsionada pelas necessidades do cliente. Lançamos novas regiões, pontos de presença, serviços e recursos regularmente. Qualquer uma dessas versões pode aprimorar positivamente a eficiência da performance de sua arquitetura.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 6: Como você aprimora sua carga de trabalho para aproveitar novas versões?

As opções de arquitetura de carga de trabalho são limitadas. No entanto, ao longo do tempo novas tecnologias e abordagens ficam disponíveis e podem aprimorar a performance de sua carga de trabalho.

Geralmente arquiteturas com baixa performance são o resultado de um processo de análise de performance inexistente ou problemático. Caso sua arquitetura esteja apresentando uma performance insatisfatória, a implementação de um processo de análise de performance permitirá que você aplique o ciclo Plan-do-check-act (PDCA – Planejar-realizar-verificar-agir) de Deming para promover um aprimoramento iterativo.

Monitoramento

Após implementar sua carga de trabalho, é necessário monitorar a performance dela para que você possa corrigir todos os problemas antes que eles afetem seus clientes. As métricas de monitoramento devem ser usadas para gerar alarmes quando os limites são ultrapassados.

O Amazon CloudWatch é um serviço de monitoramento e observação que fornece dados e informações práticas para monitorar sua carga de trabalho, responder a alterações de performance em todo o sistema, otimizar a utilização de recursos e obter uma visão unificada da saúde operacional. O CloudWatch coleta dados operacionais e de

monitoramento na forma de logs, métricas e eventos de cargas de trabalho executadas na AWS e em servidores no local. O AWS X-Ray ajuda desenvolvedores a analisarem e depurarem aplicativos distribuídos de produção. Com o AWS X-Ray, você pode obter informações sobre a performance do aplicativo, descobrir causas raiz e identificar gargalos de performance. É possível usar esses insights para reagir rapidamente e manter sua carga de trabalho funcionando sem problemas.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 7: Como você monitora seus recursos para garantir que eles estejam funcionando?

A performance do sistema pode diminuir com o tempo. Monitore a performance do sistema para identificar degradações e corrigir fatores internos ou externos, como a carga do aplicativo ou o sistema operacional.

Garantir que você não veja falsos positivos é essencial para uma solução eficaz de monitoramento. Os triggers automatizados evitam erros humanos e podem reduzir o tempo necessário para corrigir problemas. Planeje dias de jogo, nos quais as simulações sejam conduzidas no ambiente de produção para testar sua solução de alarme e garantir que ela reconheça corretamente os problemas.

Concessões

Ao arquitetar soluções, pense nas concessões para garantir uma abordagem ideal. Dependendo de sua situação, você pode abrir mão de consistência, durabilidade e espaço por tempo ou latência para oferecer uma performance mais alta.

Com a AWS, você pode se tornar global em minutos e implantar recursos em vários locais do mundo para estar mais perto dos seus usuários finais. Você também pode adicionar dinamicamente réplicas somente leitura a repositórios de informações (como sistemas de banco de dados) a fim de reduzir a carga sobre o banco de dados principal.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

PERF 8: Como você usa concessões para melhorar a performance?

Ao elaborar soluções, determinar as concessões permite que você selecione uma abordagem ideal. Muitas vezes, você pode aumentar a performance trocando consistência, durabilidade e espaço por tempo e latência.

Conforme você altera a carga de trabalho, colete e avalie métricas para determinar o impacto dessas alterações. Meça os impactos ao sistema e também ao usuário final para entender como suas concessões afetam sua carga de trabalho. Use uma abordagem sistemática, como teste de carga, para explorar se a concessão aumenta a performance.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para (pilar).

Documentação

- Amazon S3 Performance Optimization
- Amazon EBS Volume Performance

Whitepaper

- Performance Efficiency Pillar

Vídeo

- AWS re:Invent 2019: Amazon EC2 foundations (CMP211-R2)
- AWS re:Invent 2019: Leadership session: Storage state of the union (STG201-L)
- AWS re:Invent 2019: Leadership session: AWS purpose-built databases (DAT209-L)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system (CMP303-R2)
- AWS re:Invent 2019: Scaling up to your first 10 million users (ARC211-R)

Otimização de custos

O pilar (pilar) inclui (descrição)

O pilar Otimização de custos fornece uma visão geral dos princípios de design, melhores práticas e perguntas. Você pode encontrar orientações prescritivas sobre implementação no whitepaper Pilar Otimização de custos.

Princípios de design

Existem (contagem) princípios do projeto para (pilar inferior) na nuvem:

- **Implementar o gerenciamento financeiro na nuvem**: Para obter sucesso financeiro e acelerar a realização de valor empresarial na nuvem, você precisa investir em gerenciamento financeiro na nuvem/otimização de custos. Sua organização precisa dedicar tempo e recursos para criar aptidão nesse novo domínio de tecnologia e

gerenciamento de uso. Semelhante à sua aptidão de Segurança ou Operações, você precisa criar aptidão por meio da criação de conhecimento, programas, recursos e processos para se tornar uma organização econômica.

- **Adotar um modelo de consumo**: Pague somente pelos recursos de computação necessários e aumente ou reduza o uso dependendo dos requisitos comerciais, não usando previsões elaboradas. Por exemplo, ambientes de desenvolvimento e teste são geralmente usados apenas por oito horas ao dia durante a semana de trabalho. Você pode desligar esses recursos quando eles não estiverem em uso para obter uma economia potencial de 75% (40 horas versus 168 horas).
- **Meça a eficiência geral**: Meça o resultado comercial da carga de trabalho e os custos associados com a sua entrega. Use essa medida para saber os ganhos obtidos com o aumento da saída e a redução de custos.
- **Pare de gastar dinheiro em tarefas pesadas genéricas**: A AWS faz o trabalho pesado das operações de datacenter, como o armazenamento em rack, o empilhamento e a alimentação de servidores. Ele também elimina a sobrecarga operacional do gerenciamento de sistemas operacionais e aplicativos com serviços gerenciados. Isso permite que você mantenha o foco em seus clientes e projetos de negócios e não na infraestrutura de TI.
- **Analisar e atribuir despesas**: A nuvem facilita a identificação precisa do uso e do custo dos sistemas, o que permite a atribuição transparente de custos de TI a proprietários de cargas de trabalho individuais. Isso ajuda a medir o retorno sobre o investimento (ROI) e oferece aos proprietários de cargas de trabalho a oportunidade de otimizar recursos e reduzir custos.

Definição

Existem (contagem) melhores práticas para (pilar inferior) na nuvem:

- **Praticar o gerenciamento financeiro na nuvem**
- **Reconhecimento de despesas e usos**
- **Recursos econômicos**
- **Gerenciar recursos de demanda e fornecimento**
- **Otimizar ao longo do tempo**

Como acontece com os outros pilares dentro do Well-Architected Framework, é preciso escolher, por exemplo, entre otimizar para aumentar a velocidade de entrada no mercado ou para reduzir custos. Em alguns casos, é melhor otimizar a velocidade, entrar no mercado rapidamente, enviar novos recursos ou simplesmente cumprir um prazo, em vez de investir na otimização de custos inicial. Às vezes, as decisões de projeto são tomadas com base na pressa e não em dados, já que sempre existe a tentação

de compensar “para garantir”, em vez de dedicar tempo a realizar testes comparativos da implantação mais econômica. Isso pode levar a implantações com provisionamento excessivo e subotimizadas. Porém, essa é uma escolha razoável quando você precisa transferir rapidamente recursos de seu ambiente no local para a nuvem e então otimizar posteriormente. Investir na quantidade certa de esforço em uma estratégia de otimização de custos com antecedência permite aproveitar os benefícios econômicos da nuvem de modo mais rápido, garantindo uma adesão consistente às melhores práticas e evitando provisionamento excessivo desnecessário. As seções a seguir fornecem técnicas e melhores práticas para a implementação inicial e contínua do gerenciamento financeiro na nuvem e otimização de custos de suas cargas de trabalho.

Melhores práticas

Praticar o gerenciamento financeiro na nuvem

Com a adoção da nuvem, as equipes de tecnologia inovam mais rapidamente devido à redução dos ciclos de implantação de aprovação, aquisição e infraestrutura. Uma nova abordagem para o gerenciamento financeiro na nuvem é necessária para obter valor empresarial e sucesso financeiro. Essa abordagem é o gerenciamento financeiro na nuvem, e ela cria recursos em toda a organização por meio da implementação de criação, programas, recursos e processos de conhecimento em toda a organização.

Muitas organizações são compostas por várias unidades diferentes com prioridades diferentes. A capacidade de alinhar sua organização a um conjunto combinado de objetivos financeiros e fornecer a ela os mecanismos para alcançá-los criará uma organização mais eficiente. Uma organização capaz inovará e criará mais rapidamente, será mais ágil e se ajustará a todos os fatores internos ou externos.

Na AWS, você pode usar o Cost Explorer e, opcionalmente, o Amazon Athena e o Amazon QuickSight com o Relatório de custos e uso (CUR) para fornecer reconhecimento de custos e uso em toda a organização. O Orçamentos da AWS fornece notificações proativas para custo e uso. Os blogs da AWS oferecem informações sobre novos serviços e recursos para garantir que você esteja atualizado com os novos lançamentos de serviços.

As perguntas a seguir se concentram nessas considerações para (pilar inferior). (Para uma lista de perguntas e melhores práticas sobre (pilar inferior), leia o Apêndice.).

COST 1: Como implementar o gerenciamento financeiro na nuvem?
A implementação da gestão financeira na nuvem permite que as organizações obtenham valor empresarial e sucesso financeiro à medida que otimizam o custo, o uso e a escala na AWS.

Ao criar uma função de otimização de custos, considere usar membros e também complementar a equipe com especialistas em CFM e CO. Os membros da equipe com-

preenderão como a organização funciona atualmente e como implementar melhorias com rapidez. Considere também incluir pessoas com conjuntos de habilidades complementares ou especializadas, como estudo analítico e gerenciamento de projetos.

Ao implementar o reconhecimento de custos em sua organização, considere melhorar programas e processos existentes ou desenvolver com base neles. É muito mais rápido adicionar ao que já existe do que criar processos e programas novos. Isso resultará em resultados de maneira muito mais rápida.

Reconhecimento de despesas e usos

A maior flexibilidade e agilidade que a nuvem permite incentiva a inovação, desenvolvimento e implantação em ritmo acelerado. Elimina os processos manuais e o tempo associado ao provisionamento da infraestrutura no local, incluindo a identificação de especificações de hardware, negociação de cotações de preços, gerenciamento de pedidos de compra, programação de remessas e implantação dos recursos. No entanto, a facilidade de uso e a capacidade sob demanda praticamente ilimitada exigem uma nova forma de pensar sobre as despesas.

Muitas empresas são compostas por vários sistemas executados por várias equipes. A capacidade de atribuir custos de recursos à organização individual ou aos proprietários do produto gera um comportamento eficiente do uso e ajuda a reduzir o desperdício. A atribuição precisa de custos permite saber quais produtos são realmente rentáveis e permite tomar decisões mais informadas sobre alocação de orçamento.

Na AWS, você cria uma estrutura de conta com o AWS Organizations ou o AWS Control Tower, o que fornece separação e ajuda na alocação de custos e uso. Você também pode usar a marcação em recursos para aplicar informações empresariais e da organização ao seu uso e custo. Use o AWS Cost Explorer para obter visibilidade do custo e do uso ou crie estudos analíticos e painéis personalizados com o Amazon Athena e o Amazon QuickSight. O controle do custo e do uso é feito por meio de notificações com o Orçamentos da AWS, além de controles com o AWS Identity and Access Management (IAM) e cotas de serviços.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

COST 2: Como você governa o uso?

Estabeleça políticas e mecanismos para garantir que os custos adequados sejam gerados enquanto os objetivos são alcançados. Ao empregar uma abordagem de verificação e equilíbrio, você pode inovar sem gastar demais.

COST 3: Como você monitora o uso e os custos?

Estabeleça políticas e procedimentos para monitorar e alocar adequadamente os custos. Isso permite medir e aprimorar a eficiência de custos dessa carga de trabalho.

COST 4: Como você desativa os recursos?

Implemente o controle de alterações e o gerenciamento de recursos, desde o início do projeto até o fim da vida útil. Isso garante o desligamento ou encerramento dos recursos não utilizados para reduzir o desperdício.

Você pode usar tags de alocação de custos para categorizar e acompanhar o uso e os custos da AWS. Quando você aplica tags aos recursos da AWS (como instâncias do EC2 ou buckets do S3), a AWS gera um relatório de custo e uso com seu uso e suas tags. Você pode aplicar tags que representam categorias da organização (como centros de custo, nomes de carga de trabalho ou proprietários) para organizar os custos em vários serviços.

Use o nível correto de detalhes e granularidade no monitoramento e nos relatórios de custo e uso. Para obter insights e tendências de alto nível, use a granularidade diária com o AWS Cost Explorer. Para análise e inspeção mais aprofundadas, use a granularidade por hora no AWS Cost Explorer ou no Amazon Athena e no Amazon QuickSight com o Relatório de custo e uso (CUR) em uma granularidade por hora.

A combinação de recursos marcados com o acompanhamento do ciclo de vida da entidade (funcionários, projetos) permite identificar recursos ou projetos órfãos que não estão mais gerando valor para a organização e devem ser desativados. Você pode configurar alertas de pagamento para notificá-lo sobre gastos excessivos previstos.

Recursos econômicos

Usar as instâncias e os recursos adequados para sua carga de trabalho é fundamental para economizar gastos. Por exemplo, um processo de criação de relatórios pode levar cinco horas para ser executado em um servidor pequeno, mas uma hora em um servidor grande que custa o dobro. Ambos os servidores fornecem o mesmo resultado, mas o servidor menor acarreta mais custos ao longo do tempo.

Uma carga de trabalho bem projetada usa os recursos com o melhor custo-benefício, o que pode ter um impacto econômico positivo e considerável. Você também pode usar serviços gerenciados para reduzir gastos. Por exemplo, em vez de manter servidores para entrega de e-mails, você pode usar um serviço que é pago individualmente por mensagem.

A AWS oferece diversas opções de definição de preço flexíveis e econômicas para você adquirir as instâncias do EC2 e de outros serviços que sejam mais adequados às suas necessidades. *Instâncias sob demanda* permitem que você pague pela capacidade computacional por hora, sem nenhum compromisso mínimo necessário. *Savings Plans e as instâncias reservadas* oferecem economias de até 75% em relação à definição de preço sob demanda. Com instâncias spot, você pode aproveitar a capacidade não utilizada do Amazon EC2 e ter economias de até 90% na definição de preço sob demanda. As *instâncias spot* são apropriadas para sistemas que aceitam o uso de uma frota de servidores em que os servidores individuais se movimentam dinamicamente, como servidores da Web sem estado, processamento de lotes ou ao usar HPC e big data.

A seleção do serviço adequado também pode reduzir o uso e os gastos, como o CloudFront para minimizar a transferência de dados ou eliminar gastos completamente e como ao usar o Amazon Aurora em RDS para remover gastos com licenças caras de banco de dados.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

COST 5: Como você avalia o custo ao selecionar serviços?

O Amazon EC2, Amazon EBS e Amazon S3 são produtos fundamentais da AWS. Os produtos gerenciados, como Amazon RDS e Amazon DynamoDB, são produtos da AWS de nível superior ou de aplicativo. Ao selecionar os produtos fundamentais e os serviços gerenciados adequados, você pode otimizar os custos dessa carga de trabalho. Por exemplo, usando serviços gerenciados, é possível reduzir ou remover grande parte da sobrecarga administrativa e operacional, liberando você para trabalhar em aplicativos e atividades relacionadas a negócios.

COST 6: Como você atinge as metas de custo ao selecionar tamanho, número e tipo de recurso?

Escolha o tamanho e o número de recursos apropriados para a tarefa em mãos. Ao selecionar o tipo, tamanho e número mais econômicos, você minimiza o desperdício.

COST 7: Como você usa os modelos de definição de preço para reduzir custos?

Use o modelo de definição de preço mais adequado nos recursos para minimizar as despesas.

COST 8: Como você planeja as cobranças de transferência de dados?

Certifique-se de planejar e monitorar as cobranças de transferência de dados para tomar decisões de arquitetura que minimizam custos. Uma mudança arquitetônica pequena, porém eficaz, pode reduzir drasticamente os custos operacionais ao longo do tempo.

Ao considerar os gastos durante a escolha do serviço e usar ferramentas como Cost Explorer e AWS Trusted Advisor para conferir regularmente seu uso da AWS, você pode monitorar ativamente a utilização e ajustar suas implantações de acordo com ela.

Gerenciar recursos de demanda e fornecimento

Quando você passa para a nuvem, paga apenas pelo que precisa. Você pode fornecer recursos para atender à demanda da carga de trabalho no momento em que eles são necessários, o que elimina a necessidade de um provisionamento em excesso que é caro e desperdiça recursos. Você também pode modificar a demanda usando um controle de utilização, buffer ou fila para suavizar a demanda e atendê-la com menos recursos, o que resulta em um custo menor, ou processá-la posteriormente com um serviço em lote.

Na AWS, você pode provisionar automaticamente os recursos para corresponderem à demanda da carga de trabalho. O auto scaling que usa abordagens baseadas em demanda e tempo permitem que você adicione e remova recursos conforme necessário. Se você conseguir prever alterações na demanda, poderá economizar mais dinheiro e garantir que os recursos sejam compatíveis com as necessidades da sua carga de trabalho. Você pode usar o Amazon API Gateway para implementar o controle de utiliza-

ção ou o Amazon SQS para implementar uma fila em sua carga de trabalho. Os dois permitirão que você modifique a demanda nos componentes da carga de trabalho.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

COST 9: Como você gerencia a demanda e fornece recursos?

Para uma carga de trabalho que tenha gasto e performance equilibrados, verifique se tudo o que você paga é usado e evite instâncias significativamente subutilizadas. Uma métrica de utilização distorcida tem um impacto adverso na organização, nos custos operacionais (performance degradada devido à superutilização) ou nos gastos da AWS (devido ao excesso de provisionamento).

Ao projetar para modificar a demanda e fornecer recursos, pense ativamente nos padrões de uso, no tempo necessário para provisionar novos recursos e na previsibilidade do padrão de demanda. Ao gerenciar a demanda, verifique se você tem uma fila ou um buffer corretamente dimensionado e se está respondendo à demanda da carga de trabalho no período necessário.

Otimizar ao longo do tempo

Quando a AWS lança novos serviços e recursos, é recomendável analisar as escolhas de estruturas existentes para garantir que elas continuem sendo as mais econômicas. Conforme seus requisitos mudam, seja incisivo na desativação de recursos, serviços completos e sistemas que não são mais necessários.

A implementação de novos recursos ou tipos de recursos pode otimizar sua carga de trabalho de modo incremental, minimizando o esforço necessário para implementar a alteração. Isso proporciona melhorias contínuas na eficiência ao longo do tempo e garante que você permaneça na tecnologia mais atualizada para reduzir custos operacionais. Você também pode substituir ou adicionar novos componentes à carga de trabalho por novos serviços. Isso pode fornecer aumentos significativos na eficiência. Portanto, é essencial revisar regularmente sua carga de trabalho e implementar novos serviços e recursos.

As perguntas a seguir se concentram nessas considerações para (pilar inferior).

COST 10: Como você avalia os novos serviços?

Como a AWS lança novos serviços e recursos, faz parte das melhores práticas analisar as decisões de arquitetura existentes para garantir que elas continuem sendo as mais econômicas.

Ao conferir regularmente suas implantações, analise como serviços mais novos podem ajudar você a economizar dinheiro. Por exemplo, o Amazon Aurora no RDS pode reduzir gastos com bancos de dados relacionados. O uso sem servidor, como o Lambda, pode remover a necessidade de operar e gerenciar instâncias para executar código.

Recursos

Consulte os seguintes recursos para saber mais sobre nossas melhores práticas para (pilar).

Documentação

- AWS Documentation

Whitepaper

- Cost Optimization Pillar

O processo de análise

A análise das arquiteturas precisa ser feita de maneira consistente, com uma abordagem sem culpa que incentive o aprofundamento. Deve ser um processo leve (horas, não dias) que seja uma conversa e não uma auditoria. O objetivo de analisar uma arquitetura é identificar quaisquer problemas críticos que possam precisar ser abordados ou áreas que possam ser melhoradas. O resultado da análise é um conjunto de ações que devem melhorar a experiência de um cliente usando a carga de trabalho.

Conforme discutido na seção “Sobre arquitetura”, cada membro da equipe deve assumir a responsabilidade pela qualidade de sua arquitetura. Recomendamos que os membros da equipe que criam uma arquitetura usem o Well-Architected Framework para analisar continuamente sua arquitetura, em vez de realizar uma reunião formal de análise. Uma abordagem contínua permite que os membros da equipe atualizem as respostas à medida que a arquitetura evolui e melhorem a arquitetura à medida que você fornece recursos.

O AWS Well-Architected está alinhado à forma como a AWS analisa sistemas e serviços internamente. Ele tem como premissa um conjunto de princípios do projeto que influenciam a abordagem arquitetônica e perguntas que garantem que as pessoas não negligenciem as áreas que aparecem com frequência na análise de causa-raiz (RCA). Sempre que houver um problema significativo com um sistema interno, um serviço da AWS ou um cliente, examinaremos a RCA para ver se podemos melhorar os processos de análise que usamos.

As revisões devem ser aplicadas às principais etapas do ciclo de vida do produto, logo no início da fase de projeto para evitar *portas unidirecionais*[1] que são difíceis de alterar e antes da data de ativação. Sua carga de trabalho continuará evoluindo após o lançamento à medida que você adicionar novos recursos e alterar as implementações de tecnologia. A arquitetura de uma carga de trabalho muda com o tempo. Você precisará seguir boas práticas de higiene para impedir as características arquitetônicas de se degradarem à medida que evoluírem. Ao fazer alterações significativas na arquitetura, você deve seguir um conjunto de processos de higiene, incluindo uma análise do Well-Architected.

Se você quiser usar a revisão como um snapshot único ou uma medida independente, precisará garantir a presença de todas as pessoas certas na conversa. Geralmente, descobrimos que é, nas análises, a primeira vez em que a equipe realmente compreende o que implementou. Uma abordagem que funciona bem ao analisar a carga de trabalho de outra equipe é ter uma série de conversas informais sobre sua arquitetura, nas quais se pode ter as respostas para a maioria das perguntas. Em seguida, você

[1] Muitas decisões são portas bidirecionais. Essas decisões podem usar um processo leve. As portas unidirecionais são difíceis ou impossíveis de reverter e exigem mais inspeção antecipada.

pode continuar com uma ou duas reuniões para se esclarecer ou aprofundar nas áreas de ambiguidade ou risco percebidas.

Aqui estão alguns itens sugeridos para facilitar suas reuniões:

- Uma sala de reuniões com quadros brancos
- Imprimir diagramas ou notas de projeto
- Lista de ações de perguntas que exigem pesquisas fora de banda para responder (por exemplo, “habilitamos ou não a criptografia?”)

Depois de fazer uma análise você deve ter uma lista de problemas que podem ser priorizados com base no contexto da sua empresa. Você também deve considerar o impacto desses problemas no trabalho diário de sua equipe. Se você resolver esses problemas com antecedência, poderá disponibilizar mais tempo para trabalhar na criação de valor empresarial, em vez de resolver problemas recorrentes. Ao solucionar problemas, é possível atualizar a análise para ver como a arquitetura está melhorando.

Embora o valor de uma análise seja claro após sua realização, você pode descobrir que uma nova equipe pode ser resistente a princípio. Aqui estão algumas objeções que podem ser tratadas por meio da instrução da equipe sobre os benefícios de uma análise:

- “Estamos muito ocupados!” (Geralmente dito quando a equipe está se preparando para um grande lançamento.)
  - Se você estiver se preparando para um grande lançamento, deseja que ele ocorra sem problemas. A análise permitirá que você entenda os problemas que pode ter perdido.
  - Recomendamos que você faça revisões no início do ciclo de vida do produto para descobrir riscos e desenvolver um plano de mitigação alinhado ao roteiro de entrega de recursos.
- “Não temos tempo para fazer nada com os resultados!” (Geralmente, quando há um evento que não pode ser adiado, como o Super Bowl, no qual estão focados.)
  - Esses eventos não podem ser adiados. Deseja realmente entrar nele sem conhecer os riscos em sua arquitetura? Mesmo se você não abordar todos esses problemas, ainda poderá ter playbooks para lidar com eles, caso ocorram
- “We don’t want others to know the secrets of our solution implementation!”
  - Se você apresentar as perguntas do Well-Architected Framework para a equipe, eles verão que nenhuma das perguntas revela qualquer informação de propriedade comercial ou técnica.

Ao realizar várias análises com as equipes da sua organização, é possível identificar problemas temáticos. Por exemplo, você pode ver que um grupo de equipes tem gru-

pos de problemas em um pilar ou tópico específico. Veja todas as análises de maneira holística e identifique quaisquer mecanismos, treinamento ou palestras de engenharia principal que possam ajudar a resolver esses problemas temáticos.

Conclusão

O AWS Well-Architected Framework fornece melhores práticas de arquitetura nos cinco pilares para projetar e operar sistemas confiáveis, seguros, eficientes e econômicos na nuvem. O Framework fornece um conjunto de perguntas que permite analisar uma arquitetura existente ou proposta. Ele também fornece um conjunto de melhores práticas da AWS para cada pilar. O uso do Framework em sua arquitetura o ajudará a produzir sistemas estáveis e eficientes, permitindo que você se concentre em seus requisitos funcionais.

Colaboradores

As pessoas e as organizações a seguir contribuíram com este documento:

- Rodney Lester: Sr. Gerente do Well-Architected, Amazon Web Services
- Brian Carlson: Líder de operações do Well-Architected, Amazon Web Services
- Ben Potter: Líder de segurança do Well-Architected, Amazon Web Services
- Eric Pullen: Líder de performance do Well-Architected, Amazon Web Services
- Seth Eliot: Líder de confiabilidade do Well-Architected, Amazon Web Services
- Nathan Besh: Líder de custos do Well-Architected, Amazon Web Services
- Jon Steele: Sr. Gerente técnico de contas, Amazon Web Services
- Ryan King: Gerente técnico de programas, Amazon Web Services
- Erin Rifkin: Gerente sênior de produtos, Amazon Web Services
- Max Ramsay: Arquiteto-chefe de soluções de segurança, Amazon Web Services
- Scott Paddock: Arquiteto de soluções de segurança, Amazon Web Services
- Callum Hughes: Arquiteto de soluções, Amazon Web Services

Leitura adicional

AWS Cloud Compliance

AWS Well-Architected Partner program

AWS Well-Architected Tool

AWS Well-Architected homepage

Cost Optimization Pillar whitepaper

Operational Excellence Pillar whitepaper

Performance Efficiency Pillar whitepaper

Reliability Pillar whitepaper

Security Pillar whitepaper

The Amazon Builders' Library

Revisões do documento

Tabela 2. Revisões principais:

Data	Descrição
Julho 2020	Revisão e reescrita da maioria das perguntas e respostas.
Julho de 2019	Adição do AWS Well-Architected Tool, links para o AWS Well-Architected Labs e parceiros do AWS Well-Architected, correções secundárias para possibilitar a versão da estrutura em vários idiomas.
Novembro de 2018	Revisão e reescrita da maioria das perguntas e respostas, para garantir que as perguntas se concentrem em um tópico de cada vez. Isso fez com que algumas perguntas anteriores fossem divididas em várias perguntas. Adição de termos comuns às definições (carga de trabalho, componente etc). Apresentação alterada da pergunta no corpo principal para incluir texto descritivo.
Junho de 2018	Atualizações para simplificar o texto de pergunta, padronizar respostas e melhorar a legibilidade.
Novembro de 2017	O trecho sobre excelência operacional foi movido para a frente dos pilares e reescrito para enquadrar outros pilares. Atualizamos outros pilares para refletir a evolução da AWS.
Novembro de 2016	Atualização do Framework para incluir o pilar de excelência operacional e revisão e atualização dos outros pilares para reduzir a duplicação e incorporar aprendizados da realização de análises com milhares de clientes.
Novembro de 2015	Atualização do apêndice com as informações atuais do Amazon CloudWatch Logs.
Outubro de 2015	Publicação original.

Apêndice: Perguntas e melhores práticas

Excelência operacional

Organização

OPS 1 Como você determina quais são suas prioridades?

Todos precisam entender seu papel no sucesso nos negócios. Tenha objetivos compartilhados para definir as prioridades dos recursos. Isso maximizará os benefícios de seus esforços.

Melhores práticas:

- **Avaliar as necessidades de clientes externos**: Envolva as principais partes interessadas, incluindo equipes corporativas, de desenvolvimento e operacionais, a fim de determinar onde concentrar os esforços nas necessidades de clientes externos. Isso garantirá que você tenha um entendimento completo do suporte às operações necessário para obter os resultados desejados nos negócios.
- **Avaliar as necessidades internas do cliente**: Envolva as principais partes interessadas, incluindo equipes corporativas, de desenvolvimento e operacionais, ao determinar onde concentrar os esforços nas necessidades de clientes internos. Isso garantirá que você tenha um entendimento completo do suporte às operações necessário para obter resultados nos negócios.
- **Avaliar os requisitos de governança**: Certifique-se de que você esteja ciente das diretrizes ou obrigações definidas pela sua organização que possam exigir ou enfatizar um foco específico. Avalie fatores internos, como política, padrões e requisitos da organização. Confirme se você tem os mecanismos para identificar alterações na governança. Se nenhum requisito de governança for identificado, certifique-se de ter aplicado a auditoria devida a essa determinação.
- **Avaliar os requisitos de conformidade**: Avalie os fatores externos, como requisitos de conformidade regulamentar e as normas do setor, a fim de garantir que você esteja ciente das diretrizes ou obrigações que possam exigir ou enfatizar um foco específico. Se nenhum requisito de conformidade for identificado, aplique a auditoria devida a essa determinação.
- **Avaliar o cenário de ameaças**: Avalie as ameaças à empresa (por exemplo, concorrência, risco e passivos empresariais, riscos operacionais e ameaças à segurança da informação) e mantenha as informações atuais em um registro de risco. Inclua o impacto dos riscos ao determinar onde concentrar os esforços.
- **Avaliar as concessões**: Avalie o impacto das compensações entre interesses concorrentes ou abordagens alternativas para ajudar a tomar decisões embasadas ao determinar onde concentrar os esforços ou escolher um plano de ação. Por exemplo, a aceleração da velocidade de entrada no mercado de novos recursos pode ser enfatizada em relação à otimização de custos, ou você pode escolher um banco de dados relacional para dados não re-

lacionais para simplificar o esforço de migração de um sistema, em vez de migrar para um banco de dados otimizado para seu tipo de dados e atualizar seu aplicativo.

- **Gerenciar benefícios e riscos**: Gerencie benefícios e riscos para tomar decisões informadas ao determinar onde concentrar os esforços. Pode ser benéfico, por exemplo, implantar uma carga de trabalho com problemas não resolvidos a fim de disponibilizar recursos novos e significativos aos clientes. Talvez seja possível mitigar os riscos associados ou talvez seja inaceitável permitir que um risco permaneça; nesse caso, você tomará as devidas medidas para resolver o risco.

OPS 2 Como você estrutura sua organização para dar suporte aos seus resultados comerciais?

Suas equipes devem compreender o papel delas na obtenção de resultados empresariais. As equipes precisam entender o papel delas no êxito de outras equipes e a função das outras equipes no êxito delas e ter objetivos compartilhados. Entender a responsabilidade, a propriedade, como as decisões são tomadas e quem tem autoridade para tomar decisões ajudará a concentrar os esforços e maximizar os benefícios das suas equipes.

Melhores práticas:

- **Recursos com identificação de proprietários**: Entenda quem tem a propriedade de cada componente de aplicativo, carga de trabalho, plataforma e infraestrutura, qual valor empresarial é fornecido por esse componente e por que essa propriedade existe. Entender o valor empresarial desses componentes individuais e como eles dão suporte aos resultados comerciais informa os processos e procedimentos aplicados a eles.
- **Processos e procedimentos com identificação de proprietários**: Entenda quem tem a propriedade da definição de processos e procedimentos individuais, por que esses processos e procedimentos específicos são usados e por que essa propriedade existe. Entender os motivos pelos quais processos e procedimentos específicos são usados permite identificar oportunidades de melhoria.
- **Atividades de operações com identificação de proprietários responsáveis pela performance**: Entenda quem tem a responsabilidade de realizar atividades específicas em cargas de trabalho definidas e por que essa responsabilidade existe. Entender quem tem a responsabilidade de realizar atividades informa quem realizará a atividade, validará o resultado e fornecerá feedback ao proprietário da atividade.
- **Os membros da equipe sabem pelo que são responsáveis**: Entender as responsabilidades de sua função e como você contribui para resultados comerciais informa a priorização de suas tarefas e por que sua função é importante. Isso permite que os membros da equipe reconheçam as necessidades e respondam adequadamente.
- **Existem mecanismos para identificar responsabilidade e propriedade**: Quando nenhum indivíduo ou equipe é identificado, há caminhos de escalonamento definidos para alguém com autoridade para atribuir propriedade ou plano para o que precisa ser abordado.
- **Existem mecanismos para solicitar adições, alterações e exceções**: Você pode fazer solicitações aos proprietários de processos, procedimentos e recursos. Tome decisões embasadas para aprovar solicitações quando elas forem viáveis e foram consideradas apropriadas após uma avaliação de benefícios e riscos.

- **As responsabilidades entre as equipes são predefinidas ou negociadas**: Há acordos definidos ou negociados entre as equipes que descrevem como elas trabalham e oferecem suporte entre si (por exemplo, tempos de resposta, objetivos de nível de serviço ou acordos de nível de serviço). Ao entender o impacto do trabalho das equipes nos resultados de negócios e nos resultados de outras equipes e organizações, você sabe a priorização de tarefas e permite que elas respondam adequadamente.

OPS 3 Como sua cultura organizacional oferece suporte aos resultados comerciais?

Forneça suporte aos membros da equipe para que eles possam ser mais eficazes na tomada de ações e no suporte aos resultados comerciais.

Melhores práticas:

- **Patrocinador executivo**: A liderança sênior define claramente as expectativas para a organização e avalia o êxito. A liderança sênior é patrocinadora, defensora e motivadora da adoção das melhores práticas e da evolução da organização
- **Os membros da equipe são capacitados a executar ações quando os resultados estão em risco**: O proprietário da carga de trabalho definiu orientação e escopo, permitindo que os membros da equipe respondam quando os resultados estiverem em risco. Mecanismos de escalonamento são usados para obter orientação quando os eventos estão fora do escopo definido.
- **Incentivamos o escalonamento**: Os membros da equipe têm mecanismos e são incentivados a escalar as preocupações para os tomadores de decisão e as partes interessadas se acharem que os resultados estão em risco. O escalonamento deve ser realizado de maneira antecipada e frequente para que os riscos possam ser identificados e isso evite incidentes.
- **As comunicações são oportunas, claras e acionáveis**: Mecanismos existem e são usados para fornecer avisos oportunos aos membros da equipe acerca de riscos conhecidos e eventos planejados. Contexto, detalhes e tempo necessários (quando possível) são fornecidos para ajudar a determinar se há necessidade de uma ação e qual ação é necessária e a tomar as medidas necessárias em tempo hábil. Por exemplo, a notificação de vulnerabilidades de software para que a aplicação de patches possa ser expressa ou o aviso de promoções de vendas planejadas para que um congelamento de alterações possa ser implementado para evitar o risco de interrupção do serviço.
- **Incentivamos a experimentação**: A experimentação acelera o aprendizado e mantém os membros da equipe interessados e envolvidos. Um resultado indesejado é um experimento bem-sucedido que identificou um caminho que não levará ao êxito. Os membros da equipe não são punidos por experimentos bem-sucedidos com resultados indesejados. A experimentação é necessária para que a inovação ocorra e transforme ideias em resultados.
- **Os membros da equipe são capacitados e incentivados a manter e ampliar os conjuntos de habilidades**: As equipes devem aumentar os conjuntos de habilidades para adotar novas tecnologias e apoiar mudanças na demanda e responsabilidades no apoio às suas cargas de trabalho. O desenvolvimento das habilidades em novas tecnologias costuma ser uma fonte de satisfação dos membros da equipe e apoia a inovação. Ofereça apoio aos membros da equipe na busca e atualização de certificações do setor que validem e reconheçam as suas habilidades crescentes. Treine profissionais em diferentes funções para promover a transferência de conhecimento e reduzir o risco de impacto significativo quan-

do você perde membros da equipe qualificados e experientes com conhecimento institucional. Reserve tempo estruturado e dedicado para o aprendizado.

- **Forneça recursos adequados às equipes**: Mantenha a capacidade dos membros da equipe e forneça ferramentas e recursos para dar suporte às suas necessidades de carga de trabalho. A sobrecarga de membros da equipe aumenta o risco de incidentes resultantes de erros humanos. Os investimentos em ferramentas e recursos (por exemplo, fornecendo automação para atividades executadas com frequência) podem escalar a eficácia da equipe, permitindo que ela apoie atividades adicionais.
- **Diversas opiniões são incentivadas e procuradas dentro e entre equipes**: Aproveite a diversidade entre organizações para buscar várias perspectivas únicas. Use essa abordagem para aumentar a inovação, desafiar suas suposições e reduzir o risco de viés de confirmação. Aumente a inclusão, a diversidade e a acessibilidade em suas equipes para obter perspectivas benéficas.

Preparar

OPS 4 Como você projeta sua carga de trabalho para entender o estado dela?

Projete sua carga de trabalho para que as informações necessárias sejam fornecidas em todos os componentes (tais como métricas, logs e rastreamento) a fim de que você entenda seu estado interno. Isso permite que você forneça respostas efetivas quando for apropriado.

Melhores práticas:

- **Implemente a telemetria de aplicativos**: Use o código dos aplicativos para emitir informações sobre seu estado interno, status e obtenção de resultados comerciais. Tamanho da fila, mensagens de erro e tempos de resposta são alguns exemplos. Use essas informações para determinar quando uma resposta é necessária.
- **Implementar e configure a telemetria da carga de trabalho**: Projete e configure sua carga de trabalho para emitir informações sobre o estado interno e o status atual. Volume de chamadas da API, códigos de status HTTP e eventos de dimensionamento são alguns exemplos. Use essas informações para auxiliá-lo na determinação de quando uma resposta é necessária.
- **Implementar a telemetria das atividades do usuário**: Instrumente o código do aplicativo para emitir informações sobre a atividade do usuário, tais como streams de cliques ou transações iniciadas, abandonadas e concluídas. Use essas informações para ajudar a entender como o aplicativo é usado, padrões de uso e determinar quando uma resposta é necessária.
- **Implementar a telemetria de dependência**: Projete e configure sua carga de trabalho para emitir informações sobre o status (por exemplo, acessibilidade ou tempo de resposta) dos recursos dos quais depende. Exemplos de dependências externas podem incluir bancos de dados externos, DNS e conectividade de rede. Use essas informações para determinar quando uma resposta é necessária.
- **Implementar a rastreabilidade de transação**: Implemente o código do aplicativo e configure os componentes da carga de trabalho para emitir informações sobre o fluxo de tran-

sações na carga de trabalho. Use essas informações para determinar quando uma resposta é necessária e para identificar a causa raiz dos problemas.

OPS 5 Como você reduz defeitos, facilita a correção e melhora o fluxo na produção?

Adote abordagens que melhoram o fluxo de alterações na produção, que permitem refatoração, feedback rápido sobre a qualidade e correção de erros. Isso acelera as alterações benéficas que entram na produção, limita os problemas implantados e permite a rápida identificação e correção dos problemas introduzidos pelas atividades de implantação.

Melhores práticas:

- **Usar controle de versão**: Use o controle de versão para habilitar o rastreamento de alterações e liberações.
- **Testar e validar alterações**: Teste e valide as alterações para ajudar a limitar e detectar erros. Automatize os testes para reduzir erros causados por processos manuais e reduzir o nível de esforço para testar.
- **Usar sistemas de gerenciamento de configurações**: Use sistemas de gerenciamento de configurações para fazer e rastrear alterações nas configurações. Esses sistemas reduzem os erros causados pelos processos manuais e o nível de esforço para implantar as alterações.
- **Usar sistemas de gerenciamento de compilação e implantação**: Usar sistemas de gerenciamento de compilação e implantação. Esses sistemas reduzem os erros causados pelos processos manuais e o nível de esforço para implantar as alterações.
- **Executar gerenciamento de patches**: Execute o gerenciamento de patches para obter recursos, solucionar problemas e manter a conformidade com a governança. Automatize o gerenciamento de patches para reduzir erros causados por processos manuais e reduzir o nível de esforço para corrigir.
- **Compartilhar padrões de projetos**: Compartilhe as melhores práticas entre as equipes para aumentar a conscientização e maximizar os benefícios dos esforços de desenvolvimento.
- **Implementar práticas para aprimorar a qualidade do código**: Implemente práticas para aprimorar a qualidade do código e minimizar os defeitos. Por exemplo, desenvolvimento orientado por testes, análises de código e adoção de padrões.
- **Usar vários ambientes**: Use vários ambientes para experimentar, desenvolver e testar a carga de trabalho. Use níveis crescentes de controles à medida que os ambientes se aproximam da produção para adquirir confiança de que sua carga de trabalho operará conforme pretendido quando implantada.
- **Fazer alterações frequentes, pequenas e reversíveis**: Alterações frequentes, pequenas e reversíveis reduzem o escopo e o impacto de uma alteração. Isso facilita a solução de problemas, permite uma correção mais rápida e oferece a opção de reverter uma alteração.
- **Automatize totalmente a integração e a implantação**: Automatize a compilação, implantação e o teste da carga de trabalho. Isso reduz os erros causados pelos processos manuais e reduz o esforço para implantar alterações.

OPS 6 Como você reduz os riscos de implantação?

Adote abordagens que forneçam feedback rápido sobre a qualidade e permitam recuperação rápida de alterações que não têm os resultados desejados. O uso dessas práticas reduz o impacto dos problemas introduzidos pela implantação de mudanças.

Melhores práticas:

- **Planeje-se para eventuais alterações sem êxito**: Planeje reverter para um bom estado anterior ou a realização de reparos no ambiente de produção se uma mudança não tiver o resultado desejado. Esta preparação reduz o tempo de recuperação através de respostas mais rápidas.
- **Testar e validar alterações**: Teste as alterações e valide os resultados em todas as etapas do ciclo de vida, para confirmar novos recursos e minimizar o risco e o impacto de implementações com falha.
- **Use sistemas de gerenciamento para implantação**: Use sistemas de gerenciamento para implantação a fim de rastrear e implementar mudanças. Isso reduz os erros causados pelos processos manuais e reduz o esforço para implantar alterações.
- **Teste usando implantações limitadas**: Teste implantações limitadas junto com os sistemas existentes para confirmar os resultados desejados antes da implantação em grande escala. Use testes para implantação canário ou implantações individuais, por exemplo.
- **Implante usando ambientes paralelos**: Implemente alterações em ambientes paralelos e faça a transição para o novo ambiente. Mantenha o ambiente anterior até que haja confirmação de uma implantação bem-sucedida. Ao fazer isso, o tempo de recuperação é minimizado, permitindo assim a reversão para o ambiente anterior.
- **Implante mudanças frequentes, pequenas e reversíveis**: Use alterações frequentes, pequenas e reversíveis para reduzir o escopo de uma alteração. Isso resulta em solução de problemas mais fácil e correção mais rápida, com a opção de reverter uma alteração.
- **Automatize totalmente a integração e a implantação**: Automatize a construção, implantação e o teste da carga de trabalho. Isso reduz os erros causados pelos processos manuais e reduz o esforço para implantar alterações.
- **Automatize testes e reversões**: Automatize os testes dos ambientes implantados para confirmar os resultados desejados. Automatize a reversão para o bom estado anterior conhecido quando os resultados não forem alcançados para minimizar o tempo de recuperação e reduzir os erros causados por processos manuais.

OPS 7 Como você sabe que está pronto para oferecer suporte a uma carga de trabalho?

Avalie a prontidão operacional de sua carga de trabalho, processos/procedimentos e pessoal para entender os riscos operacionais relacionados.

Melhores práticas:

- **Garanta a capacidade de pessoal**: Tenha um mecanismo para validar que você tem o número adequado de pessoal treinado para fornecer suporte às necessidades operacionais. Treine e ajuste a capacidade de pessoal conforme necessário para manter o suporte eficiente.

- **Garanta uma análise consistente da prontidão operacional**: Verifique se você tem uma análise consistente de sua prontidão para operar uma carga de trabalho. As análises devem incluir, no mínimo, a prontidão operacional das equipes e da carga de trabalho e as considerações de segurança. Implemente atividades de análise em código e acione a análise automatizada em resposta a eventos, quando adequado, para garantir consistência, velocidade de execução e reduzir erros causados por processos manuais.
- **Use runbooks para executar procedimentos**: Os runbooks são os procedimentos documentados para alcançar resultados específicos. Habilite respostas consistentes e rápidas para eventos bem conhecidos, documentando procedimentos nos runbooks. Implemente runbooks como código e acione a execução de runbooks em resposta a eventos, quando adequado, para garantir consistência, agilizar as respostas e reduzir erros causados por processos manuais.
- **Usar playbooks para investigar problemas**: Habilite respostas consistentes e rápidas a problemas que não são bem compreendidos, documentando o processo de investigação nos playbooks. Playbooks são as etapas predefinidas executadas para identificar os fatores que contribuem para um cenário de falha. Os resultados de qualquer etapa do processo são usados para determinar as próximas etapas a serem seguidas até que o problema seja identificado ou encaminhado.
- **Tome decisões informadas para implantar sistemas e mudanças**: Avalie os recursos da equipe para oferecer suporte à carga de trabalho e à conformidade da carga de trabalho com a governança. Avalie isso em relação aos benefícios da implantação ao determinar se deseja fazer a transição para um sistema ou mudar para produção. Compreenda os benefícios e riscos para tomar decisões informadas.

Operar

OPS 8 Como você compreende a integridade da sua carga de trabalho?

Defina, capture e analise as métricas da carga de trabalho para obter visibilidade destes eventos, para que você possa tomar as ações apropriadas.

Melhores práticas:

- **Identifique os indicadores-chave de performance**: Identifique os indicadores-chave de performance (KPIs) com base nos resultados de negócios desejados (por exemplo, taxa de pedidos, taxa de retenção do cliente e lucro versus despesa operacional) e resultados do cliente (por exemplo, satisfação do cliente). Avalie os KPIs para determinar o sucesso da carga de trabalho.
- **Defina as métricas de carga de trabalho**: Defina métricas de carga de trabalho para medir a realização de KPIs (por exemplo, carrinhos de compras abandonados, pedidos feitos, custo, preço e despesas de carga de trabalho alocadas). Defina métricas de carga de trabalho para medir a integridade da carga de trabalho (por exemplo, tempo de resposta da interface, taxa de erros, solicitações feitas, solicitações concluídas e utilização). Avalie as métricas para determinar se a carga de trabalho está alcançando os resultados desejados e para entender a sua integridade.

- **Colete e analise as métricas de carga de trabalho.**: Faça revisões proativas regulares das métricas para identificar tendências e determine onde as respostas apropriadas são necessárias.
- **Estabeleça as linhas de base de métricas de carga de trabalho.**: Estabeleça as linhas de base das métricas para fornecer valores esperados como base para comparação e identificação de componentes com performance inferior e superior. Identificar limites para melhoria, investigação e intervenção.
- **Aprenda os padrões esperados de atividade para carga de trabalho.**: Estabeleça padrões de atividade de carga de trabalho para identificar comportamentos anômalos para que você possa responder adequadamente, se necessário.
- **Atente para quando os resultados da carga de trabalho estiverem em risco**: Emita um alerta quando os resultados da carga de trabalho estiverem em risco, para que você possa responder adequadamente, se necessário.
- **Atente para quando anomalias de carga de trabalho forem detectadas**: Emita um alerta quando forem detectadas anomalias na carga de trabalho, para que você possa responder adequadamente, se necessário.
- **Valide a obtenção de resultados e a eficácia de KPIs e métricas.** : Crie uma visualização em nível de negócios de suas operações de carga de trabalho para ajudá-lo a determinar se você está satisfazendo estas necessidades e para identificar áreas que precisam de melhorias para atingir as metas de negócios. Valide a eficácia dos KPIs e métricas e revise-os, se necessário.

OPS 9 Como você compreende a integridade de suas operações?

Defina, capture e analise as métricas de operações para obter visibilidade dos eventos de operações, para que você possa tomar as ações apropriadas.

Melhores práticas:

- **Identifique os indicadores-chave de performance**: Identifique os indicadores-chave de performance (KPIs) com base nos negócios desejados (por exemplo, novos recursos entregues) e nos resultados do cliente (por exemplo, casos de suporte ao cliente). Avalie KPIs para determinar o sucesso das operações.
- **Defina as métricas de operações**: Defina métricas de operações para medir a realização de KPIs (por exemplo, implantações com êxito e implantações com falha). Defina métricas de operações para medir a integridade das atividades de operações (por exemplo, tempo médio para detectar um incidente (MTTD) e tempo médio para recuperação (MTTR) de um incidente). Avalie as métricas para determinar se as operações estão alcançando os resultados desejados e para entender a integridade das atividades operacionais.
- **Colete e analise as métricas de operações**: Faça revisões proativas regulares das métricas para identificar tendências e determine onde as respostas apropriadas são necessárias.
- **Estabeleça as linhas de base das métricas de operações**: Estabeleça as linhas de base das métricas para fornecer valores esperados como base para comparação e identificação de atividades operacionais com performance inferior e superior.

- **Aprenda os padrões esperados de atividade para operações**: Estabeleça padrões de atividades de operações para identificar atividades anômalas para poder responder adequadamente, se necessário.
- **Atente para quando os resultados das operações estiverem em risco**: Emita um alerta quando os resultados das operações estiverem em risco para que você possa responder adequadamente, se necessário.
- **Atente para quando anomalias de operações forem detectadas**: Emita um alerta quando forem detectadas anomalias de operações para que você possa responder adequadamente, se necessário.
- **Valide a obtenção de resultados e a eficácia de KPIs e métricas.** : Crie uma visualização em nível de negócios de suas atividades operacionais para ajudá-lo a determinar se você está satisfazendo estas necessidades e para identificar áreas que precisam de melhorias para atingir as metas de negócios. Valide a eficácia dos KPIs e métricas e revise-os, se necessário.

OPS 10 Como você gerencia os eventos de carga de trabalho e operações?

Prepare e valide procedimentos para responder a eventos, com o objetivo de minimizar a interrupção de sua carga de trabalho.

Melhores práticas:

- **Use processos para gerenciamento de eventos, incidentes e problemas**: Tenha processos para tratar de eventos observados, eventos que exijam intervenção (incidentes) e eventos que exijam intervenção e que se repitam ou que não possam ser resolvidos no momento (problemas). Use esses processos para mitigar o impacto desses eventos nos negócios e em seus clientes, garantindo respostas oportunas e apropriadas.
- **Ter um processo por alerta**: Tenha uma resposta bem-definida (runbook ou playbook), com um proprietário especificamente identificado, para qualquer evento para o qual você acione um alerta. Isso garante respostas eficazes e rápidas aos eventos de operações e evita que eventos acionáveis sejam ocultados por notificações menos valiosas.
- **Priorizar eventos operacionais com base no impacto nos negócios**: Quando vários eventos demandarem intervenção, aborde primeiro os mais significativos para os negócios. Os impactos, por exemplo, podem incluir perda de vidas ou ferimentos, perda financeira ou danos à reputação ou confiança.
- **Defina caminhos de escalação**: Defina caminhos de escalação em seus runbooks e playbooks, incluindo o que aciona a escalação e os procedimentos para escalação. Identifique especificamente os proprietários de cada ação para garantir respostas eficazes e rápidas aos eventos de operações.
- **Habilitar notificações por push**: Comunique-se diretamente com seus usuários (e-mail ou SMS, por exemplo) quando os serviços que eles usam são afetados e novamente quando os serviços retornam às condições operacionais normais, para permitir que os usuários tomem as medidas apropriadas.

- **Comunique o status por meio de painéis**: Forneça painéis personalizados para os públicos-alvo (por exemplo, equipes técnicas internas, liderança e clientes) para comunicar o status operacional atual dos negócios e fornecer métricas de interesse.
- **Automatizar respostas a eventos**: Automatize as respostas aos eventos para reduzir os erros causados por processos manuais e garantir respostas rápidas e consistentes.

Evoluir

OPS 11 Como você evolui as operações?

Dedique tempo e recursos para a melhoria incremental contínua, a fim de aumentar a eficácia e a eficiência de suas operações.

Melhores práticas:

- **Tenha um processo para melhoria contínua.**: Avalie e priorize regularmente oportunidades de melhorias para concentrar os esforços onde eles possam oferecer os maiores benefícios.
- **Executar análise pós-incidente**: Analise os eventos que afetam o cliente e identifique os fatores que contribuem e as ações preventivas. Use essas informações para desenvolver mitigações para limitar ou evitar recorrência. Desenvolva procedimentos para respostas rápidas e eficazes. Comunique os fatores contribuintes e as ações corretivas conforme apropriado, de acordo com o público-alvo.
- **Implementar ciclos de comentários**: Inclua ciclos de comentários em procedimentos e cargas de trabalho para ajudar a identificar problemas e áreas que precisam de melhorias.
- **Executar o gerenciamento de conhecimento**: Existem mecanismos para que os membros da equipe descubram as informações que estão procurando em tempo hábil, acessem essas informações e identifiquem que são atuais e completas. Mecanismos estão presentes para identificar o conteúdo necessário, o conteúdo que precisa de atualização e o conteúdo que deve ser arquivado para que não seja mais referenciado.
- **Definir os condutores de melhoria**: Identifique os condutores de melhoria para ajudá-lo a avaliar e priorizar as oportunidades.
- **Validar os insights**: Revise os resultados e as respostas da análise com equipes multifuncionais e proprietários de negócios. Use essas revisões para estabelecer um entendimento comum, identificar impactos adicionais e determinar cursos de ação. Ajuste as respostas conforme apropriado.
- **Fazer análises de métricas de operações**: Realize regularmente análises retrospectivas das métricas de operações com participantes de equipes cruzadas de diferentes áreas do negócio. Use essas análises para identificar oportunidades de melhorias e possíveis ações e compartilhar as lições aprendidas.
- **Documentar e compartilhar as lições aprendidas**: Documente e compartilhe as lições aprendidas com a execução de atividades operacionais, para que você possa usá-las internamente e entre equipes.

- **Alocar tempo para fazer melhorias**: Dedique tempo e recursos em seus processos para possibilitar melhorias incrementais contínuas.

Segurança

Segurança

SEC 1 Como você opera com segurança sua carga de trabalho?

Para operar sua carga de trabalho com segurança, você deve aplicar as melhores práticas gerais a todas as áreas de segurança. Use os requisitos e os processos que você definiu em excelência operacional em nível de carga de trabalho e também organizacional e aplique-os a todas as áreas. Manter-se atualizado com as recomendações da AWS e do setor e a inteligência de ameaças ajuda você a desenvolver seu modelo de ameaças e objetivos de controle. A automação de processos, testes e validação de segurança permite que você escale suas operações de segurança.

Melhores práticas:

- **Separar as cargas de trabalho usando contas**: Organize as cargas de trabalho em contas separadas e contas de grupo com base na função ou em um conjunto comum de controles, em vez de espelhar a estrutura de comunicação da empresa. Tenha em mente a segurança e a infraestrutura ao começar para que sua organização possa definir proteções comuns à medida que as cargas de trabalho aumentam.
- **Proteger a conta da AWS**: Proteja o acesso às suas contas, por exemplo, habilitando a MFA, restrinja a utilização do usuário raiz e configure os contatos da conta.
- **Identificar e validar objetivos de controle**: Com base em seus requisitos de conformidade e riscos identificados no modelo de ameaça, derive e valide os objetivos de controle e os controles que você precisa aplicar à carga de trabalho. A validação contínua de objetivos de controle e controles ajuda a medir a eficácia da mitigação de riscos.
- **Manter-se atualizado sobre as ameaças à segurança**: Reconheça vetores de ataque mantendo-se a par das ameaças de segurança mais recentes para definir e implementar os controles adequados.
- **Manter-se atualizado com as recomendações de segurança**: Mantenha-se atualizado com as recomendações de segurança da AWS e do setor para desenvolver a postura de segurança da sua carga de trabalho.
- **Automatizar testes e validação de controles de segurança em pipelines**: Estabeleça linhas de base e modelos seguros para mecanismos de segurança que são testados e validados como parte de sua compilação, pipelines e processos. Use ferramentas e automação para testar e validar todos os controles de segurança continuamente. Por exemplo, verifique itens, como imagens de máquina e infraestrutura, como modelos de código, para detectar vulnerabilidades de segurança, irregularidades e desvios da linha de base estabelecida em cada estágio.
- **Identificar e priorizar riscos usando um modelo de ameaça**: Use um modelo de ameaça para identificar e manter um registro atualizado de potenciais ameaças. Priorize as amea-

ças e adapte os controles de segurança para prevenir, detectar e responder. Revise e mantenha essas informações no contexto do cenário de segurança em evolução.

- **Avaliar e implementar regularmente novos serviços e recursos de segurança**: Os parceiros da AWS e do APN lançam constantemente novos recursos e serviços que permitem que você desenvolva a postura de segurança da sua carga de trabalho.

Identity and Access Management

SEC 2 Como você gerencia identidades para pessoas e máquinas?

Há dois tipos de identidades que você precisa gerenciar para operar cargas de trabalho seguras da AWS. Entender o tipo de identidade de que você precisa para gerenciar e conceder acesso ajuda a garantir que as identidades corretas tenham acesso aos recursos certos nas condições certas. Identidades humanas: administradores, desenvolvedores, operadores e usuários finais precisam de uma identidade para acessar seus ambientes e aplicações da AWS. Eles são membros da sua organização ou usuários externos com quem você colabora e que interagem com seus recursos da AWS por meio de um navegador da web, aplicação cliente ou ferramentas interativas de linha de comando. Identidades de máquina: aplicações de serviço, ferramentas operacionais e cargas de trabalho precisam de uma identidade para solicitar serviços da AWS; por exemplo, para ler dados. Essas identidades incluem máquinas em execução no seu ambiente da AWS, como instâncias do Amazon EC2 ou funções do AWS Lambda. Você também pode gerenciar identidades de máquina para partes externas que precisam de acesso. Além disso, você pode ter máquinas fora da AWS que precisam de acesso ao seu ambiente da AWS.

Melhores práticas:

- **Usar mecanismos de login forte**: Imponha o tamanho mínimo da senha e instrua os usuários a evitar senhas comuns ou reutilizadas. Aplique a multi-factor authentication (MFA) com mecanismos de software ou hardware para fornecer controle de acesso adicional.
- **Usar credenciais temporárias**: Exija que as identidades adquiram credenciais temporárias dinamicamente. Para identidades da força de trabalho, use o AWS Single Sign-On ou federação com funções do IAM para acessar contas da AWS. Para identidades de máquina, exija o uso de funções do IAM em vez de chaves de acesso de longo prazo.
- **Armazenar e usar segredos com segurança**: As identidades de força de trabalho e de máquinas que precisam de segredos, como senhas para aplicações de terceiros, devem ser armazenadas com rotação automática, segundo os padrões mais recentes do setor em um serviço especializado.
- **Contar com um provedor de identidade centralizado**: Para identidades da força de trabalho, conte com um provedor de identidade que permita a você gerenciar identidades em um local centralizado. Dessa forma, você pode criar, gerenciar e revogar o acesso em um único local, o que facilita o gerenciamento do acesso. Esse procedimento reduz a necessidade de várias credenciais e oferece uma oportunidade de integração com processos de RH.
- **Fazer a auditoria e a rotação periódica das credenciais**: Quando você não puder contar com credenciais temporárias e exigir credenciais de longo prazo, faça uma auditoria das

credenciais para garantir que os controles definidos (por exemplo, MFA) sejam aplicados, alternados regularmente e que tenham o nível de acesso apropriado.

- **Utilizar grupos e atributos de usuários**: Coloque usuários com requisitos de segurança comuns em grupos definidos pelo provedor de identidade e implemente mecanismos para garantir que os atributos de usuário que podem ser usados para controle de acesso (por exemplo, departamento ou localização) estejam corretos e atualizados. Use esses grupos e atributos, em vez de usuários individuais, para controlar o acesso. Com isso, você pode gerenciar o acesso centralmente. Basta alterar uma vez a associação ou os atributos do grupo de um usuário. Ou seja, não será preciso atualizar muitas políticas individuais quando as necessidades de acesso de um usuário mudarem.

SEC 3 Como você gerencia permissões para pessoas e máquinas?

Gerencie permissões para controlar o acesso a identidades de pessoas e máquinas que precisam de acesso à AWS e à sua carga de trabalho. As permissões controlam quem pode acessar o quê e em quais condições.

Melhores práticas:

- **Definir requisitos de acesso**: Cada componente ou recurso da carga de trabalho precisa ser acessado por administradores, usuários finais ou outros componentes. É necessário ter uma definição clara de quem ou do que deve ter acesso a cada componente ou recurso e, em seguida, escolher o tipo de identidade apropriado e o método de autenticação e autorização.
- **Conceder menos privilégio**: Conceda somente o acesso de que as identidades precisam, permitindo acesso a ações específicas em recursos específicos da AWS em condições específicas. Conte com grupos e atributos de identidade para definir permissões dinamicamente em grande escala, em vez de definir permissões para usuários individuais. Por exemplo, você pode permitir o acesso de um grupo de desenvolvedores para gerenciar apenas recursos de seu próprio projeto. Dessa forma, quando um desenvolvedor é removido do grupo, seu acesso é revogado em todos os lugares em que esse grupo foi usado para controle de acesso, sem precisar efetuar qualquer alteração nas políticas de acesso.
- **Estabelecer processo de acesso de emergência**: Um processo que permite o acesso de emergência à carga de trabalho no caso improvável de um problema no processo automatizado ou no pipeline. Isso ajudará você a confiar no acesso de privilégio mínimo e garantirá que os usuários possam obter o nível certo de acesso quando precisarem. Por exemplo, estabeleça um processo para que os administradores verifiquem e aprovem sua solicitação.
- **Reduzir as permissões continuamente**: À medida que as equipes e as cargas de trabalho determinam o acesso de que precisam, remova as permissões que eles não usam mais e estabeleça processos de análise para obter permissões de privilégio mínimo. Monitore e reduza continuamente identidades e permissões não utilizadas.
- **Definir proteções de permissões para sua organização**: Estabeleça controles comuns que restrinjam o acesso a todas as identidades na organização. Por exemplo, você pode restringir o acesso a regiões específicas da AWS ou impedir que os operadores excluam recursos comuns, como uma função do IAM usada pela equipe de segurança central.

- **Gerenciar o acesso com base no ciclo de vida**: Integre controles de acesso ao ciclo de vida do operador e do aplicativo e ao seu provedor de federação centralizado. Por exemplo, remova o acesso do usuário que sair da organização ou mudar de funções.
- **Analisar o acesso público e entre contas**: Monitore continuamente as descobertas que destacam o acesso público e entre contas. Reduza o acesso público e o acesso entre contas aos recursos que exigem esse tipo de acesso.
- **Compartilhar recursos com segurança**: Controle o consumo de recursos compartilhados entre contas ou dentro da organização da AWS. Monitore recursos compartilhados e revise o acesso a recursos compartilhados.

Detecção

SEC 4 Como você detecta e investiga eventos de segurança?

Capture e analise eventos de logs e métricas para gerar visibilidade. Tome medidas em eventos de segurança e potenciais ameaças para ajudar a proteger sua carga de trabalho.

Melhores práticas:

- **Configurar registro em log de serviço e aplicativo**: Configure o registro em log em toda a carga de trabalho, incluindo logs de aplicativos, logs de recursos e logs de serviços da AWS. Por exemplo, verifique se o AWS CloudTrail, o Amazon CloudWatch Logs, o Amazon GuardDuty e o AWS Security Hub estão habilitados para todas as contas da organização.
- **Analisar logs, descobertas e métricas de forma centralizada**: Todos os logs, métricas e telemetria devem ser coletados centralmente e analisados automaticamente para detectar anomalias e indicadores de atividade não autorizada. Um painel pode fornecer informações sobre a integridade fáceis de acessar em tempo real. Por exemplo, certifique-se de que os logs do Amazon GuardDuty e do Security Hub sejam enviados para um local central para fins de alertas e análises.
- **Automatizar a resposta a eventos**: O uso de automação para investigar e corrigir eventos reduz o esforço humano e erros e permite escalar recursos de investigação. Análises regulares ajudarão você a ajustar ferramentas de automação e iterar continuamente. Por exemplo, automatize respostas a eventos do Amazon GuardDuty automatizando a primeira etapa de investigação e, em seguida, itere para remover gradualmente o esforço humano.
- **Implementar eventos de segurança acionáveis**: Crie alertas para serem enviados à sua equipe para ação. Certifique-se de que os alertas incluam informações relevantes para a equipe agir. Por exemplo, certifique-se de que os alertas do Amazon GuardDuty e do AWS Security Hub sejam enviados à equipe para ação ou enviados a ferramentas de automação de resposta que mantêm a equipe informada por meio de mensagens da estrutura de automação.

Proteção de infraestrutura

SEC 5 Como você protege seus recursos de rede?

Qualquer carga de trabalho que tenha alguma forma de conectividade de rede, seja a Internet ou uma rede privada, exige várias camadas de defesa para ajudar a proteger contra ameaças externas e internas baseadas em rede.

Melhores práticas:

- **Criar camadas de rede**: Agrupe componentes que compartilham requisitos de acessibilidade em camadas. Por exemplo, um cluster de banco de dados em uma VPC sem necessidade de acesso à Internet deve ser colocado em sub-redes sem nenhuma rota para/da Internet. Em uma carga de trabalho sem servidor operando sem uma VPC, camadas e segmentação semelhantes com microsserviços podem atingir o mesmo objetivo.
- **Controlar tráfego de todas as camadas**: Aplique controles com uma abordagem de defesa detalhada para tráfego de entrada e saída. Por exemplo, para a Amazon Virtual Private Cloud (VPC), isso inclui grupos de segurança, ACLs de rede e sub-redes. Para o AWS Lambda, considere executar em sua VPC privada com controles baseados em VPC.
- **Automatizar proteção de rede**: Automatize os mecanismos de proteção para fornecer uma rede de autodefesa com base em inteligência de ameaças e detecção de anomalias. Por exemplo, ferramentas de detecção e prevenção de intrusão que podem se adaptar proativamente às ameaças atuais e reduzir seu impacto.
- **Implementar inspeção e proteção**: Inspecione e filtre o tráfego em cada camada. Por exemplo, use um firewall de aplicação web para proteger contra o acesso acidental na camada de rede do aplicativo. Para as funções do Lambda, ferramentas de terceiros podem adicionar firewalls de camada de aplicativo ao ambiente de tempo de execução.

SEC 6 Como você protege seus recursos de computação?

Os recursos de computação exigem várias camadas de defesa para ajudar na proteção contra ameaças externas e internas. Os recursos de computação incluem instâncias do EC2, contêineres, funções do AWS Lambda, serviços de banco de dados, dispositivos de IoT e muito mais.

Melhores práticas:

- **Executar o gerenciamento de vulnerabilidades**: Verifique e corrija com frequência vulnerabilidades no código, nas dependências e na infraestrutura para proteger-se contra novas ameaças.
- **Reduzir superfície de ataque**: Reduza a superfície de ataque fortalecendo sistemas operacionais, minimizando componentes, bibliotecas e serviços consumíveis externamente em uso.
- **Implementar serviços gerenciados**: Implemente serviços que gerenciam recursos, como Amazon RDS, AWS Lambda e Amazon ECS, para reduzir as tarefas de manutenção de segurança como parte do modelo de responsabilidade compartilhada.
- **Automatizar proteção de computação**: Automatize seus mecanismos de computação de proteção, incluindo gerenciamento de vulnerabilidades, redução da superfície de ataque e gerenciamento de recursos.

- **Permitir que as pessoas executem ações a uma distância**: A remoção da capacidade de acesso interativo reduz o risco de erro humano e o potencial de configuração ou gerenciamento manual. Por exemplo, use um fluxo de trabalho de gerenciamento de alterações para implantar instâncias do EC2 usando infraestrutura como código e, em seguida, gerencie instâncias do EC2 usando ferramentas em vez de permitir acesso direto ou um bastion host.
- **Validar a integridade do software**: Implemente mecanismos (por exemplo, assinatura de código) para validar se o software, o código e as bibliotecas usados na carga de trabalho são de fontes confiáveis e não foram adulterados.

Proteção de dados

SEC 7 Como classificar meus dados?

A classificação serve para categorizar os dados com base em criticidade e confidencialidade para ajudá-lo a determinar os controles de proteção e retenção apropriados.

Melhores práticas:

- **Identificar os dados em sua carga de trabalho**: Isso inclui o tipo e a classificação dos dados, os processos de negócios associados, o proprietário dos dados, os requisitos legais e de conformidade aplicáveis, onde são armazenados e os controles resultantes que devem ser aplicados. Isso pode incluir classificações para indicar se os dados devem ser disponibilizados publicamente, se os dados são apenas de uso interno, como informações de identificação pessoal do cliente (PII) ou se os dados são para acesso mais restrito, como propriedade intelectual, dados legalmente privilegiados ou marcados como confidenciais, e muito mais.
- **Definir controles de proteção de dados**: Proteja os dados de acordo com seu nível de classificação. Por exemplo, proteja dados classificados como públicos usando recomendações relevantes enquanto protege dados confidenciais com controles adicionais.
- **Automatizar identificação e classificação**: Automatize a identificação e a classificação dos dados para reduzir o risco de erro humano.
- **Definir o gerenciamento do ciclo de vida de dados**: Sua estratégia de ciclo de vida definida deve ser baseada no nível de confidencialidade, bem como nos requisitos legais e organizacionais. Aspectos como o tempo da retenção dos dados, processos de destruição de dados, gerenciamento de acesso a dados, transformação de dados e compartilhamento de dados devem ser considerados.

SEC 8 Como você protege seus dados em repouso?

Proteja seus dados em repouso implementando vários controles para reduzir o risco de acesso não autorizado ou manuseio incorreto.

Melhores práticas:

- **Implementar gerenciamento de chaves seguro**: As chaves de criptografia devem ser armazenadas em segurança, com um rigoroso controle de acesso; por exemplo, usando um serviço de gerenciamento de chaves, como o AWS KMS. Considere o uso de chaves dife-

rentes e o controle de acesso às chaves, combinado com as políticas de recursos e IAM da AWS, para alinhamento com os níveis de classificação de dados e requisitos de segregação.

- **Aplicar criptografia em repouso**: Aplique seus requisitos de criptografia definidos com base nos mais recentes padrões e recomendações para proteger os dados em repouso.
- **Automatizar a proteção de dados em repouso**: Use ferramentas automatizadas para validar e aplicar controles de dados em repouso continuamente, por exemplo, verificar se há apenas recursos de armazenamento criptografados.
- **Aplicar controle de acesso**: Aplique controle de acesso com privilégios mínimos e mecanismos, incluindo backups, isolamento e versionamento, para ajudar a proteger seus dados ociosos. Impeça que os operadores concedam acesso público aos seus dados.
- **Usar mecanismos para evitar que as pessoas acessem os dados**: Impeça que os usuários acessem dados e sistemas confidenciais diretamente em circunstâncias operacionais normais. Por exemplo, ofereça um painel em vez de acesso direto a um armazenamento de dados para executar consultas. Quando os pipelines de CI/CD não forem usados, determine quais controles e processos são necessários para fornecer adequadamente um mecanismo de acesso break-glass normalmente desabilitado.

SEC 9 Como você protege seus dados em trânsito?

Proteja seus dados em trânsito implementando vários controles para reduzir o risco de acesso não autorizado ou perda.

Melhores práticas:

- **Implementar o gerenciamento seguro de chaves e certificados**: Armazene chaves e certificados de criptografia com segurança e alterne-os em intervalos regulares com rigoroso controle de acesso; por exemplo, com um serviço de gerenciamento de certificados como o AWS Certificate Manager (ACM).
- **Aplique a criptografia em trânsito**: Usar os requisitos de criptografia definidos com base em padrões e recomendações apropriados para conseguir cumprir os requisitos organizacionais, legais e de conformidade.
- **Automatizar a detecção de acesso não intencional a dados**: Use ferramentas como o GuardDuty para detectar automaticamente tentativas de mover dados para fora de limites definidos com base no nível de classificação dos dados, por exemplo, para detectar um cavalo de Troia que esteja copiando dados para uma rede desconhecida ou não confiável usando o protocolo DNS.
- **Autenticar as comunicações de rede**: Verifique a identidade das comunicações usando protocolos que oferecem suporte à autenticação, como Transport Layer Security (TLS) ou IPsec.

Resposta a incidentes

SEC 10 Como você prevê, responde e se recupera de incidentes?

A preparação é essencial para investigação, resposta e recuperação oportunas e eficazes de incidentes de segurança para ajudar a minimizar interrupções na sua organização.

Melhores práticas:

- **Identificar o pessoal-chave e os recursos externos**: Identifique o pessoal, as obrigações legais e os recursos internos e externos que ajudariam sua organização a responder a um incidente.
- **Desenvolver planos de gerenciamento de incidentes**: Crie planos para ajudar a responder, a se comunicar e a se recuperar de um incidente. Por exemplo, você pode começar com um plano de resposta a incidentes com os cenários mais prováveis para sua carga de trabalho e organização. Inclua como você se comunicaria e escalaria interna e externamente.
- **Preparar recursos forenses**: Identifique e prepare recursos de investigação forense adequados, incluindo especialistas externos, ferramentas e automação.
- **Automatizar a capacidade de contenção**: Automatize os recursos de contenção e recuperação de incidentes para reduzir o tempo de resposta e o impacto organizacional.
- **Pré-provisionar o acesso**: Certifique-se de que os respondentes a incidentes tenham o acesso correto pré-provisionado na AWS para reduzir o tempo de investigação até a recuperação.
- **Pré-implantar ferramentas**: Garanta que o pessoal de segurança tenha as ferramentas certas pré-implantadas na AWS para reduzir o tempo de investigação até a recuperação.
- **Promova dias de jogo**: Pratique dias de jogo de resposta a incidentes (simulações) regularmente, incorpore as lições aprendidas aos planos de gerenciamento de incidentes e melhore continuamente.

Confiabilidade

Fundamentos

REL 1 Como você gerencia as cotas e restrições de serviço?

Para arquiteturas de carga de trabalho baseadas na nuvem, há cotas de serviço, que também são conhecidas como limites de serviço. Essas cotas existem para evitar o aprovisionamento acidental de mais recursos do que o necessário e para limitar as taxas de solicitação nas operações de API para proteger os serviços contra abuso. Há também restrições de recursos, por exemplo, a taxa de envio de bits por um cabo de fibra óptica ou a quantidade de armazenamento em um disco físico.

Melhores práticas:

- **Conhecimento das cotas e restrições de serviço**: Você está ciente das suas cotas padrão e das solicitações de aumento de cota referentes à sua arquitetura de carga de trabalho. Você também sabe quais restrições de recursos, como disco ou rede, podem gerar impactos.
- **Gerencie cotas de serviço de várias contas e regiões**: Se você estiver usando várias contas ou regiões da AWS, solicite as cotas adequadas em todos os ambientes nos quais suas cargas de trabalho de produção são executadas.

- **Acomode as cotas e as restrições fixas de serviço por meio da arquitetura**: Tenha conhecimento das cotas de serviço e dos recursos físicos imutáveis e elabore um plano para evitar que eles afetem a confiabilidade.
- **Monitore e gerencie cotas**: Avalie seu uso potencial e aumente suas cotas adequadamente, permitindo o crescimento planejado do uso.
- **Automatize o gerenciamento de cotas**: Implemente ferramentas para alertar você quando os limites estiverem perto de serem atingidos. Ao usar as APIs das Cotas de serviços da AWS, você pode automatizar as solicitações de aumento de cota.
- **Verifique se existe uma lacuna suficiente entre as cotas atuais e o uso máximo para acomodar o failover**: Quando um recurso falha, ele ainda pode ser incluído na cotas até ser encerrado com êxito. Certifique-se de que suas cotas compensem a sobreposição de todos os recursos que falharam com substituições antes do encerramento desses recursos. Você deve considerar uma falha na zona de disponibilidade ao calcular essa lacuna.

REL 2 Como você planeja sua topologia de rede?

Muitas vezes, as cargas de trabalho estão presentes em vários ambientes. Dentre eles estão vários ambientes de nuvem (acessíveis publicamente e privados) e possivelmente sua infraestrutura de datacenter existente. Os planos devem incluir considerações de rede, como conectividade dentro dos sistemas e entre eles, gerenciamento de endereços IP públicos e privados e resolução de nomes de domínio.

Melhores práticas:

- **Use conectividade de rede altamente disponível em seus endpoints públicos de carga de trabalho**: Esses endpoints e o roteamento para eles devem ser altamente disponíveis. Para que isso seja possível, use DNS altamente disponível, Content Delivery Networks (CDNs – Redes de entrega de conteúdo), API Gateway, balanceamento de carga ou proxies reversos.
- **Provisione conectividade redundante entre as redes privadas na nuvem e nos ambientes no local**: Use várias conexões do AWS Direct Connect (DX) ou túneis VPN entre as redes privadas implantadas separadamente. Use vários locais do DX para alta disponibilidade. Se estiver usando várias regiões da AWS, garanta a redundância em pelo menos duas delas. Você pode avaliar os appliances do AWS Marketplace que encerram as VPNs. Se você usa appliances do AWS Marketplace, implante instâncias redundantes em zonas de disponibilidade diferentes para alta disponibilidade.
- **Garanta contas de alocação de sub-rede IP para expansão e disponibilidade**: Os intervalos de endereços IP do Amazon VPC devem ser grandes o suficiente para acomodar os requisitos da carga de trabalho, incluindo a futura expansão e alocação de endereços IP para sub-redes nas zonas de disponibilidade. Isso inclui load balancers, instâncias do EC2 e aplicativos baseados em contêiner.
- **Prefira topologias hub-and-spoke em vez da malha muitos-para-muitos**: Se mais de dois espaços de endereço de rede (por exemplo, VPCs e redes no local) estiverem conectados por meio do emparelhamento de VPC, do AWS Direct Connect ou da VPN, use um modelo hub-and-spoke, como o fornecido pelo AWS Transit Gateway.

- **Aplique intervalos de endereços IP privados não sobrepostos a todos os espaços de endereços privados em que estão conectados**: Os intervalos de endereços IP de cada uma das suas VPCs não devem se sobrepor quando emparelhados ou conectados por VPN. Você deve evitar conflitos de endereço IP da mesma forma entre uma VPC e ambientes no local ou com outros provedores de nuvem que você usa. Você também deve ter uma maneira de alocar intervalos de endereços IP privados quando necessário.

Arquitetura da carga de trabalho

REL 3 Como você projeta sua arquitetura de serviços de carga de trabalho?

Use uma Service-Oriented Architecture (SOA – Arquitetura orientada por serviços) ou uma arquitetura de microsserviços para criar cargas de trabalho altamente escaláveis e confiáveis. A SOA é a prática de tornar componentes de software reutilizáveis por meio de interfaces de serviço. A arquitetura de microsserviços vai além para tornar os componentes menores e mais simples.

Melhores práticas:

- **Escolha como segmentar a carga de trabalho**: A arquitetura monolítica deve ser evitada. Em vez dela, escolha entre SOA e microsserviços. Ao fazer cada escolha, analise os benefícios em relação às complexidades. O que é ideal para um novo produto a caminho do seu primeiro lançamento não se aplica a uma carga de trabalho que foi criada para escalabilidade a partir das necessidades iniciais. Os benefícios de usar segmentos menores incluem maior agilidade, flexibilidade organizacional e escalabilidade. As complexidades incluem maior latência potencial, depuração mais complexa e carga operacional aumentada
- **Crie serviços voltados a domínios e funcionalidades de negócios específicos**: A SOA cria serviços com funções bem delineadas que seguem as necessidades dos negócios. Os microsserviços usam modelos de domínio e contexto controlado para maior limitação de modo que cada serviço execute apenas uma ação. O foco na funcionalidade específica permite diferenciar os requisitos de confiabilidade de serviços diferentes e direcionar os investimentos de forma mais distinta. Um problema de negócio conciso e uma equipe pequena associada a cada serviço também facilitam a escalabilidade organizacional.
- **Forneça contratos de serviço por API**: Os contratos de serviço são acordos documentados entre as equipes que envolvem a integração dos serviços e incluem uma definição de API legível por máquina, limites de taxa e expectativas de performance. Uma estratégia de versionamento permite que os clientes continuem usando a API existente e migrem seus aplicativos para a API mais recente quando estiverem prontos. A implantação pode acontecer a qualquer momento, desde que o contrato não seja violado. A equipe do provedor de serviços pode usar a pilha de tecnologia de sua preferência para cumprir o contrato de API. Da mesma forma, o consumidor do serviço pode usar sua própria tecnologia.

REL 4 Como você projeta interações em um sistema distribuído para evitar falhas?

Os sistemas distribuídos dependem das redes de comunicação para interconectar componentes, como servidores ou serviços. Sua carga de trabalho deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar sem afetar negativamente outros componentes ou a carga de trabalho. Essas

melhores práticas evitam falhas e melhoram o Mean Time Between Failures (MTBF – Tempo médio entre falhas).

Melhores práticas:

- **Identifique qual tipo de sistema distribuído é necessário**: Os sistemas distribuídos em tempo real rígidos exigem respostas síncronas e rápidas, enquanto os sistemas em tempo real flexíveis têm uma janela de tempo para resposta maior, de minutos ou mais. Os sistemas off-line gerenciam as respostas por meio do processamento em lote ou assíncrono. Os sistemas distribuídos em tempo real rígidos têm os requisitos de confiabilidade mais rigorosos.
- **Implementar dependências com acoplamento fraco**: As dependências, como sistemas de enfileiramento, sistemas de streaming, fluxos de trabalho e load balancers, têm acoplamento fraco. O baixo acoplamento ajuda a isolar o comportamento de um componente dos outros componentes que dependem dele, o que aumenta a resiliência e a agilidade
- **Faça com que todas as respostas sejam idempotentes**: Um serviço idempotente garante que cada solicitação seja concluída exatamente uma vez, de modo que fazer várias solicitações idênticas tem o mesmo efeito de uma única solicitação. Um serviço idempotente facilita para um cliente implementar novas tentativas sem o receio de que uma solicitação seja processada erroneamente várias vezes. Para fazer isso, os clientes podem emitir solicitações de API com um token de idempotência. O mesmo token é usado sempre que a solicitação é repetida. Uma API de serviço idempotente usa o token para retornar uma resposta idêntica à resposta que foi retornada na primeira vez que a solicitação foi concluída.
- **Faça um trabalho constante**: Os sistemas podem falhar quando há alterações grandes e rápidas na carga. Por exemplo, um sistema de verificação de integridade que monitora a integridade de milhares de servidores deve sempre enviar a carga útil com o mesmo tamanho (um snapshot completo do estado atual). Se houver uma falha em todos os servidores ou se não houver falha alguma, o sistema de verificação de integridade realizará um trabalho constante sem alterações grandes e rápidas.

REL 5 Como você projeta interações em um sistema distribuído para mitigar ou resistir a falhas?

Os sistemas distribuídos dependem de redes de comunicação para interconectar componentes (como servidores ou serviços). Sua carga de trabalho deve operar de forma confiável, apesar da perda de dados ou da latência nessas redes. Os componentes do sistema distribuído devem operar sem afetar negativamente outros componentes ou a carga de trabalho. Essas melhores práticas permitem que as cargas de trabalho resistam a tensões ou falhas, recuperem-se mais rapidamente delas e reduzam o impacto de tais prejuízos. Como resultado, o Mean Time To Recovery (MTTR – Tempo médio até a recuperação) é melhorado.

Melhores práticas:

- **Implementar uma degradação simples para transformar dependências rígidas aplicáveis em dependências flexíveis**: Quando as dependências de um componente não estão íntegras, o próprio componente ainda pode funcionar, embora de maneira prejudicada. Por exemplo, quando há falha em uma chamada de dependência, faça o failover para uma resposta estática predeterminada.

- **Solicitações de controle de utilização**: Esse é um padrão de mitigação para responder a um aumento inesperado na demanda. Algumas solicitações são atendidas, mas aquelas que ultrapassam um limite definido são rejeitadas e retornam uma mensagem indicando que foram limitadas. A expectativa dos clientes é que eles recuem e abandonem a solicitação ou tentem novamente com uma taxa mais lenta.
- **Controle e limite as chamadas de repetição**: Use o recuo exponencial para tentar novamente após intervalos progressivamente mais longos. Introduza uma variação para tornar esses intervalos de repetição aleatórios e limite o número máximo de novas tentativas.
- **Falha rápida e filas limitadas**: Se a carga de trabalho não puder responder a uma solicitação com êxito, gere uma falha rápida. Isso permite a liberação dos recursos associados a uma solicitação e permite que o serviço se recupere se estiver ficando sem recursos. Se a carga de trabalho puder responder com êxito, mas a taxa de solicitações for muito alta, use uma fila para armazenar as solicitações em buffer. No entanto, não permita filas longas que possam levar ao fornecimento de solicitações obsoletas que o cliente já tinha descartado.
- **Defina tempos limite do cliente**: Defina tempos limite adequados, verifique-os sistematicamente e não use valores padrão, já que eles costumam ser muito altos
- **Crie serviços sem estado sempre que possível**: Os serviços não devem exigir estado ou devem descarregar o estado de modo que não haja dependência entre solicitações de clientes diferentes em relação aos dados armazenados localmente no disco ou na memória. Isso permite que os servidores sejam substituídos quando necessário sem prejudicar a disponibilidade. O Amazon ElastiCache ou o Amazon DynamoDB é um bom destino para o estado descarregado.
- **Implementar medidas emergenciais**: Trata-se de processos rápidos que podem atenuar o impacto da disponibilidade sobre a carga de trabalho. Eles podem ser operados na ausência de uma causa raiz. Uma medida emergencial ideal reduz a carga cognitiva dos resolvedores a zero ao fornecer critérios de ativação e de desativação totalmente determinísticos. Alguns exemplos de medidas são o bloqueio de todo o tráfego de robô ou o fornecimento de uma resposta estática. Geralmente, as medidas são manuais, mas também podem ser automatizadas.

Gerenciamento de alterações

REL 6 Como você monitora recursos de carga de trabalho?

Os logs e as métricas são uma ferramenta poderosa para saber a integridade das suas cargas de trabalho. Você pode configurar sua carga de trabalho para monitorar logs e métricas e enviar notificações quando os limites forem ultrapassados ou em caso de eventos importantes. O monitoramento permite que sua carga de trabalho reconheça quando os limites de baixa performance são ultrapassados ou quando há falhas, para que ela possa se recuperar automaticamente em resposta.

Melhores práticas:

- **Monitore todos os componentes da carga de trabalho (geração)**: Monitore os componentes da carga de trabalho com o Amazon CloudWatch ou ferramentas de terceiros. Monitore os serviços da AWS com o Personal Health Dashboard
- **Defina e calcule as métricas (agregação)**: Armazene os dados de log e aplique filtros quando necessário para calcular métricas como contagens de um evento de log específico ou latência calculada com base na data e hora dos eventos de log
- **Envie notificações (processamento e emissão de alarmes em tempo real)**: As organizações que precisam estar a par de tudo, recebem notificações quando ocorrem eventos importantes
- **Automatize respostas (processamento e emissão de alarmes em tempo real)**: Use a automação para executar uma ação quando um evento é detectado, por exemplo, para substituir componentes com falha
- **Armazenamento e estudo analítico**: Colete arquivos de log e históricos de métricas e analise-os para obter tendências mais abrangentes e informações sobre a carga de trabalho
- **Faça revisões regularmente**: Revise frequentemente a implementação do monitoramento da carga de trabalho e atualize-a com base em eventos e alterações significativos
- **Monitore o rastreamento completo das solicitações por meio do seu sistema**: Use o AWS X-Ray ou ferramentas de terceiros para que os desenvolvedores possam analisar e depurar mais facilmente os sistemas distribuídos para entender a performance dos aplicativos e dos serviços subjacentes deles

REL 7 Como você projeta sua carga de trabalho para se adaptar às mudanças na demanda?

Uma carga de trabalho escalável oferece elasticidade para adicionar ou remover recursos automaticamente para que atendam melhor à demanda atual a qualquer momento.

Melhores práticas:

- **Use a automação ao obter ou escalar recursos**: Ao substituir recursos danificados ou escalar sua carga de trabalho, automatize o processo por meio dos serviços gerenciados pela AWS, como o Amazon S3 e o AWS Auto Scaling. Você também pode usar ferramentas de terceiros e os SDKs da AWS para automatizar a escalabilidade.
- **Obtenha recursos após a detecção de danos em uma carga de trabalho**: Escale recursos de modo reativo quando necessário, se a disponibilidade for afetada, para restaurar a disponibilidade da carga de trabalho.
- **Obtenha recursos após a detecção de que mais recursos são necessários para uma carga de trabalho**: Escale os recursos proativamente para atender à demanda e evitar impacto na disponibilidade.
- **Fazer o teste de carga da sua carga de trabalho**: Adote uma metodologia de teste de carga para avaliar se a ação de escalabilidade atende aos requisitos da carga de trabalho.

REL 8 Como você implementa uma alteração?

As alterações controladas são necessárias para implantar novas funcionalidades e garantir que as cargas de trabalho e o ambiente operacional executem softwares conhecidos e possam ser corrigidos ou substituídos de maneira previsível. Se essas alterações forem descontroladas, será difícil prever o efeito ou resolver problemas decorrentes delas.

Melhores práticas:

- **Use runbooks para atividades padrão, como implantação**: Os runbooks são as etapas predefinidas usadas para atingir resultados específicos. Use-os para executar atividades padrão, sejam elas feitas manualmente ou automaticamente. Os exemplos incluem a implantação de uma carga de trabalho, a aplicação de patches a ela ou a realização de modificações de DNS.
- **Integre testes funcionais como parte da sua implantação**: Os testes funcionais são executados como parte da implantação automatizada. Se os critérios de êxito não forem atendidos, o pipeline será interrompido ou revertido.
- **Integre testes de resiliência como parte da sua implantação**: Os testes de resiliência (como parte da engenharia do caos) são executados como parte do pipeline de implantação automatizado em um ambiente de pré-produção.
- **Faça a implantação com uma infraestrutura imutável**: Esse é um modelo que não requer atualizações, patches de segurança ou alterações de configuração nas cargas de trabalho de produção. Quando uma alteração é necessária, a arquitetura é criada em uma nova infraestrutura e implantada na produção.
- **Implante alterações com automação**: As implantações e a aplicação de patches são automatizadas para eliminar o impacto negativo.

Gerenciamento de falhas

REL 9 Como você faz backup dos dados?

Faça backup de dados, aplicativos e configurações para atender aos seus requisitos de Recovery Time Objective (RTO – Objetivo do tempo de recuperação) e de Recovery Point Objective (RPO – Objetivo do ponto de recuperação).

Melhores práticas:

- **Identifique e faça backup de todos os dados que precisam ser incluídos no backup ou reproduza os dados das fontes**: O Amazon S3 pode ser usado como destino de backup para várias fontes de dados. Os serviços da AWS, como Amazon EBS, Amazon RDS e Amazon DynamoDB, têm recursos integrados para criar backups. É possível também usar um software de backup de terceiros. Por outro lado, se os dados puderem ser reproduzidos de outras fontes para atender ao RPO, talvez você não precise de um backup
- **Proteja e criptografe backups**: Use a autenticação e a autorização, como o AWS IAM, para detectar acessos e use a criptografia para detectar o comprometimento da integridade dos dados.
- **Execute o backup de dados automaticamente**: Configure os backups para serem feitos automaticamente com base em uma programação periódica ou de acordo com as altera-

ções feitas no conjunto de dados. É possível configurar instâncias do RDS, volumes do EBS, tabelas do DynamoDB e objetos do S3 para backup automático. É possível também usar soluções do AWS Marketplace ou de terceiros.

- **Execute a recuperação periódica dos dados para verificar a integridade e os processos de backup**: Execute um teste de recuperação para confirmar se a implementação do processo de backup atende aos seus objetivos do tempo de recuperação e de ponto de recuperação.

REL 10 Como usar o isolamento de falhas para proteger sua carga de trabalho?

Os limites isolados de falhas restringem o efeito de uma falha em uma carga de trabalho a um número controlado de componentes. A falha não afeta os componentes fora do limite. Ao usar vários limites isolados de falhas, você pode restringir o impacto sobre sua carga de trabalho.

Melhores práticas:

- **Implante a carga de trabalho em vários locais**: Distribua os dados e os recursos da carga de trabalho por várias zonas de disponibilidade ou, quando necessário, por regiões da AWS. A diversidade dos locais pode variar conforme a necessidade.
- **Automatize a recuperação de componentes restritos a um único local**: Se os componentes da carga de trabalho só puderem ser executados em uma única zona de disponibilidade ou datacenter no local, você deverá implementar capacidade suficiente para fazer uma recompilação completa da carga de trabalho em conformidade com os objetivos de recuperação definidos.
- **Use arquiteturas de anteparo**: Assim como os anteparos de um navio, esse padrão garante que uma falha seja contida em um pequeno subconjunto de solicitações ou usuários de modo que o número de solicitações prejudicadas seja limitado, e a maioria possa continuar sem erros. Geralmente, os anteparos de dados são chamados de partições ou fragmentos, enquanto os anteparos de serviços são conhecidos como células.

REL 11 Como você projeta sua carga de trabalho para resistir a falhas de componentes?

As cargas de trabalho que exigem alta disponibilidade e baixo Mean Time To Recovery (MTTR – Tempo médio até a recuperação) devem ser projetadas visando a resiliência.

Melhores práticas:

- **Monitore todos os componentes da carga de trabalho para detectar falhas**: Monitore constantemente a integridade da carga de trabalho para que você e seus sistemas automatizados detectem degradações ou falhas completas assim que elas ocorrerem. Monitore os Key Performance Indicators (KPIs – Indicadores-chave de performance) com base no valor empresarial.
- **Realize failover para recursos íntegros em locais intactos**: Se ocorrer uma falha de local, verifique se os dados e os recursos dos locais íntegros podem continuar processando as solicitações. Isso é mais fácil para as cargas de trabalho multizona porque os serviços da AWS, como o Elastic Load Balancing e o AWS Auto Scaling, ajudam a distribuir a carga entre as zonas de disponibilidade. Para as cargas de trabalho multirregionais, o procedimen-

to é mais complicado. Por exemplo, as réplicas de leitura entre as regiões permitem implantar os dados em várias regiões da AWS, mas você ainda deve promover a réplica de leitura a mestre e apontar o tráfego para ela em caso de falha no local principal. O Amazon Route 53 e o AWS Global Accelerator também podem ajudar a rotear o tráfego entre as regiões da AWS.

- **Automatize a reparação em todas as camadas**: Após a detecção de uma falha, use recursos automatizados para executar ações de correção.
- **Use a estabilidade estática para evitar o comportamento bimodal**: O comportamento bimodal é quando a carga de trabalho apresenta um comportamento diferente nos modos normal e de falha, por exemplo, depender da execução de novas instâncias se uma zona de disponibilidade falhar. Em vez disso, você deve criar cargas de trabalho que sejam estaticamente estáveis e que operem em apenas um modo. Nesse caso, provisione instâncias suficientes em cada zona de disponibilidade para processar a carga de trabalho se uma AZ foi removida e use as verificações de integridade do Elastic Load Balancing ou do Amazon Route 53 para remover a carga das instâncias danificadas.
- **Envie notificações quando os eventos afetarem a disponibilidade**: As notificações são enviadas após a detecção de eventos significativos, mesmo que o problema causado pelo evento tenha sido resolvido automaticamente.

REL 12 Como testar a confiabilidade?

Depois de projetar sua carga de trabalho para resiliência à pressão da produção, o teste é a única maneira de garantir que ela opere conforme projetado e com a resiliência esperada.

Melhores práticas:

- **Usar playbooks para investigar falhas**: Faça a documentação do processo de investigação em playbooks para permitir respostas consistentes e rápidas em cenários de falha. Os playbooks são as etapas predefinidas executadas para identificar os fatores que contribuem para um cenário de falha. Os resultados de qualquer etapa do processo são usados para determinar as próximas etapas a serem seguidas até que o problema seja identificado ou encaminhado.
- **Executar análise pós-incidente**: Analise os eventos que afetam o cliente e identifique os fatores contribuintes e os itens de ação preventiva. Use essas informações para desenvolver mitigações para limitar ou evitar recorrência. Desenvolva procedimentos para respostas rápidas e eficazes. Comunique os fatores contribuintes e as ações corretivas conforme apropriado, de acordo com o público-alvo. Tenha um método para comunicar essas causas a outras pessoas, conforme necessário.
- **Teste os requisitos funcionais**: Esse procedimento inclui testes de unidade e de integração que validam a funcionalidade necessária.
- **Teste os requisitos de escalabilidade e performance**: Esse procedimento inclui o teste de carga para validar se a carga de trabalho atende aos requisitos de escalabilidade e performance.
- **Teste a resiliência por meio da engenharia do caos**: Execute testes que injetam falhas regularmente em ambientes de pré-produção e de produção. Especule como sua carga de trabalho reagirá à falha, depois compare sua hipótese com os resultados do teste e reafir-

me se elas não corresponderem. Certifique-se de que os testes de produção não afetem os usuários.

- **Conduza dias de jogo regularmente**: Use os dias de jogo para praticar regularmente seus procedimentos de falha o mais próximo possível da produção (inclusive em ambientes de produção) com as pessoas que estarão envolvidas nos cenários de falha reais. Os dias de jogo aplicam medidas para garantir que os testes de produção não afetem os usuários.

REL 13 Como você planeja a recuperação de desastres (DR)?

Implementar backups e componentes redundantes de carga de trabalho é o ponto de partida da sua estratégia de DR. O RTO e o RPO são os objetivos para restaurar a disponibilidade. Defina-os de acordo com suas necessidades de negócios. Implemente uma estratégia para atender a esses objetivos, considerando os locais e a função dos recursos e dos dados da carga de trabalho.

Melhores práticas:

- **Defina os objetivos de recuperação para tempo de inatividade e perda de dados**: A carga de trabalho tem um Recovery Time Objective (RTO – Objetivo do tempo de recuperação) e um Recovery Point Objective (RPO – Objetivo do ponto de recuperação).
- **Use estratégias de recuperação definidas para atingir os objetivos de recuperação**: Uma estratégia de Disaster Recovery (DR – Recuperação de desastres) foi definida para atingir os objetivos.
- **Teste a implementação de recuperação de desastres para validá-la**: Teste regularmente o failover para DR para garantir que o RTO e o RPO sejam cumpridos.
- **Gerencie o desvio de configuração para o local ou a região de DR**: Certifique-se de que a infraestrutura, os dados e a configuração estejam conforme necessário no local ou na região de DR. Por exemplo, verifique se as AMIs e as cotas de serviço estão atualizadas.
- **Automatize a recuperação**: Use ferramentas da AWS ou de terceiros para automatizar a recuperação do sistema e rotear o tráfego para o local ou a região de DR.

Eficiência de performance

Seleção

PERF 1 Como você seleciona a arquitetura de melhor performance?

Muitas vezes, é necessário empregar várias abordagens para obter a performance ideal em uma carga de trabalho. Os sistemas com boa arquitetura usam várias soluções e recursos para aprimorar a performance.

Melhores práticas:

- **Compreenda os serviços e os recursos disponíveis**: Conheça e compreenda a ampla gama de serviços e recursos disponíveis na nuvem. Identifique os serviços e opções de configuração relevantes para sua carga de trabalho e entenda como alcançar a performance ideal.

- **Defina um processo para opções de arquitetura**: Use a experiência interna e o conhecimento da nuvem, ou recursos externos, como casos de uso publicados, documentação relevante ou whitepapers para definir um processo para escolher recursos e serviços. Você deve definir um processo que incentive a experimentação e o benchmarking com os serviços que poderiam ser usados em sua carga de trabalho.
- **Leve em conta os requisitos de custo ao tomar decisões** : Muitas vezes, as cargas de trabalho têm requisitos de custo para operação. Use controles internos de custo para selecionar tipos e tamanhos de recursos com base na necessidade prevista dos respectivos recursos.
- **Use políticas ou arquiteturas de referência**: Maximize a performance e a eficiência avaliando políticas internas e arquiteturas de referência existentes, usando sua análise a fim de selecionar serviços e configurações para sua carga de trabalho.
- **Use as orientações do seu provedor de nuvem ou de um parceiro apropriado**: Use recursos da empresa de nuvem, como arquitetos de soluções, serviços profissionais ou um parceiro apropriado para orientar suas decisões. Esses recursos podem ajudar a analisar e melhorar sua arquitetura para alcançar uma performance ideal.
- **Realize testes comparativos de cargas de trabalho**: Faça um teste comparativo de uma carga de trabalho para entender a performance dela na nuvem. Use os dados coletados em benchmarks para direcionar as decisões de arquitetura.
- **Fazer o teste de carga da sua carga de trabalho**: Implante sua arquitetura de carga de trabalho mais recente na nuvem usando recursos de diferentes tipos e tamanhos. Monitore a implantação a fim de capturar métricas de performance que identificam gargalos ou excessos de capacidade. Use essas informações de performance para projetar ou aprimorar a seleção de sua arquitetura e dos respectivos recursos.

PERF 2 Como você seleciona sua solução de computação?

A solução de computação ideal para uma carga de trabalho varia conforme o design do aplicativo, os padrões de uso e as definições de configuração. As arquiteturas podem usar diferentes soluções de computação para vários componentes e podem habilitar diferentes recursos para melhorar a performance. Selecionar a solução de computação incorreta para uma arquitetura pode levar a uma menor eficiência de performance.

Melhores práticas:

- **Avalie as opções de computação disponíveis**: Entenda as características de performance das opções relacionadas a computação disponíveis. Saiba como instâncias, contêineres e funções funcionam, e quais vantagens ou desvantagens elas agregam à sua carga de trabalho.
- **Compreenda as opções de configuração de computação disponíveis**: Compreenda como diferentes opções complementam sua carga de trabalho e que opções de configuração são melhores para seu sistema. Exemplos dessas opções incluem família de instâncias, tamanhos, recursos (GPU, E/S), tamanhos de função, instâncias de contêiner e modelo de um ou vários locatários.
- **Colete métricas relacionadas à computação**: Uma das melhores maneiras de entender a performance de seus sistemas de computação é registrar e acompanhar a verdadeira uti-

lização de vários recursos. Esses dados podem ser usados para fazer determinações mais precisas sobre os requisitos de recursos.

- **Determine a configuração necessária realizando o dimensionamento correto**: Analise as várias características de performance de sua carga de trabalho e como elas se relacionam a uso de memória, rede e CPU. Use esses dados para escolher os recursos mais adequados ao perfil da sua carga de trabalho. Por exemplo, a melhor maneira de atender a uma carga de trabalho com uso intenso de memória, como um banco de dados, pode ser usando a família r de instâncias. No entanto, uma carga de trabalho com intermitência pode se beneficiar mais de um sistema de contêiner elástico.
- **Use a elasticidade de recursos disponível**: A nuvem fornece a flexibilidade de expandir ou reduzir seus recursos dinamicamente por meio de diversos mecanismos para atender a mudanças na demanda. Combinada com métricas relacionadas à computação, uma carga de trabalho pode responder automaticamente a mudanças e utilizar um conjunto ideal de recursos para atingir sua meta.
- **Reavalie as necessidades de computação conforme as métricas**: Use as métricas no nível do sistema para identificar o comportamento e os requisitos de sua carga de trabalho ao longo do tempo. Avalie as necessidades de sua carga de trabalho, comparando os recursos disponíveis com esses requisitos, e faça alterações em seu ambiente de computação para melhor atender ao perfil de sua carga de trabalho. Por exemplo, ao longo do tempo, pode-se observar que um sistema consome mais memória do que inicialmente previsto, assim, a adoção de uma família ou tamanho de instância diferente pode melhorar tanto a performance quanto a eficiência.

PERF 3 Como você seleciona sua solução de armazenamento?

A solução de armazenamento ideal para um sistema varia conforme o tipo de método de acesso (bloco, arquivo ou objeto), os padrões de acesso (aleatório ou sequencial), o rendimento necessário, a frequência de acesso (online, offline, arquivamento), a frequência de atualização (WORM, dinâmica) e as restrições de disponibilidade e durabilidade. Os sistemas Well-Architected usam várias soluções de armazenamento e habilitam diferentes recursos para melhorar a performance e usar os recursos de modo eficiente.

Melhores práticas:

- **Compreenda as características e os requisitos de armazenamento**: Compreenda as diferentes características (p. ex., compartilhamento, tamanho de arquivo, tamanho do cache, padrões de acesso, latência, throughput e persistência de dados) necessárias para selecionar os serviços mais adequados à sua carga de trabalho, como armazenamento de objetos, armazenamento em bloco, armazenamento de arquivos ou armazenamento de instâncias.
- **Avalie as opções de configuração disponíveis**: Avalie as diversas características e opções de configuração e como se relacionam ao armazenamento. Entenda onde e como usar IOPS provisionadas, SSDs, armazenamento magnético, armazenamento de objeto, armazenamento em repositório ou armazenamento temporário para otimizar o espaço de armazenamento e a performance para sua carga de trabalho.
- **Tome decisões com base nos padrões de acesso e nas métricas**: Escolha sistemas de armazenamento com base nos padrões de acesso de sua carga de trabalho e configure-os determinando como a carga de trabalho acessa os dados. Aumente a eficiência do arma-

zenamento escolhendo armazenamento de objetos em vez de armazenamento em bloco. Configure as opções de armazenamento escolhidas para corresponder a seus padrões de acesso a dados.

PERF 4 Como você seleciona sua solução de banco de dados?

A solução de banco de dados ideal para um sistema varia conforme os requisitos de disponibilidade, consistência, tolerância da partição, latência, durabilidade, escalabilidade e capacidade de consulta. Muitos sistemas usam soluções de banco de dados diferentes para vários subsistemas e habilitam diferentes recursos para melhorar a performance. Selecionar a solução e os recursos de banco de dados incorretos para um sistema pode levar a uma menor eficiência.

Melhores práticas:

- **Entenda as características dos dados**: Entenda as diferentes características dos dados em sua carga de trabalho. Determine se a carga de trabalho requer transações, como ela interage com dados e quais são as demandas de performance dela. Use esses dados para selecionar a abordagem de melhor performance para seu banco de dados (p. ex., bancos de dados relacionais, de chave-valor em NoSQL, documentos, coluna ampla, gráficos, série temporal ou armazenamento em memória).
- **Avalie as opções disponíveis**: Avalie os serviços e as opções de armazenamento disponíveis como parte do processo de seleção para os mecanismos de armazenamento de sua carga de trabalho. Entenda como e quando usar um determinado serviço ou sistema para armazenamento de dados. Conheça as opções de configuração disponíveis que podem otimizar a performance ou a eficiência do banco de dados, como IOPS provisionadas, recursos de computação e memória, além de armazenamento em cache.
- **Colete e registre métricas de performance do banco de dados**: Use ferramentas, bibliotecas e sistemas que registram as medidas de performance relacionadas ao banco de dados. Por exemplo, meça transações por segundo, consultas lentas ou latência do sistema introduzida ao acessar o banco de dados. Use esses dados para entender a performance de seus sistemas de banco de dados.
- **Escolha o armazenamento de dados conforme os padrões de acesso**: Use os padrões de acesso da carga de trabalho para decidir que serviços e tecnologias usar. Por exemplo, utilize um banco de dados relacional para cargas de trabalho que exigem transações, ou um repositório de chave-valor que forneça um throughput maior, mas que seja eventualmente consistente quando aplicável.
- **Otimize o armazenamento de dados conforme as métricas e os padrões de acesso**: Use características de performance e padrões de acesso que otimizem o modo como os dados são armazenados ou consultados para obter a melhor performance possível. Meça como otimizações, p. ex., indexação, distribuição de chave, design do data warehouse ou estratégias de armazenamento em cache afetam a performance do sistema ou a eficiência geral.

PERF 5 Como você configura sua solução de redes?

A solução de rede ideal para uma carga de trabalho varia com base nos requisitos de latência, throughput, instabilidade e largura de banda. Restrições físicas, como recursos de usuário ou

no local, determinam as opções de localização. Essas restrições podem ser compensadas com pontos de presença ou posicionamento de recursos.

Melhores práticas:

- **Entenda como as redes afetam a performance**: Analise e entenda como decisões relacionadas à rede afetam a performance da carga de trabalho. Por exemplo, a latência da rede costuma afetar a experiência do usuário, e usar os protocolos incorretos pode esgotar a capacidade da rede devido à sobrecarga excessiva.
- **Avalie 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, aproveite os recursos de rede que estão disponíveis para reduzir a latência, a distância ou a instabilidade da rede.
- **Escolha VPN ou conectividade dedicada dimensionada adequadamente para cargas de trabalho híbridas**: Quando houver um requisito de comunicação no local, verifique se você tem largura de banda adequada para a performance da carga de trabalho. Com base nos requisitos de largura de banda, uma única conexão dedicada ou uma única VPN pode não ser suficiente, e você precisa habilitar o balanceamento de carga de tráfego em várias conexões.
- **Aproveite o balanceamento de carga e o descarregamento de criptografia**: Distribua o tráfego entre vários recursos e serviços para permitir que sua carga de trabalho aproveite a elasticidade que a nuvem oferece. Também é possível usar o balanceamento de carga para descarregar a terminação de criptografia a fim de melhorar a performance e gerenciar e rotear o tráfego de maneira eficaz.
- **Escolha os protocolos de rede para aumentar a performance**: Tome decisões sobre protocolos de comunicação entre sistemas e redes com base no impacto na performance da carga de trabalho.
- **Escolha o local da sua carga de trabalho com base nos requisitos de rede**: Use as opções de localização de nuvem disponíveis para reduzir a latência de rede ou aprimorar o throughput. Utilize regiões da AWS, zonas de disponibilidade, grupos de posicionamento e pontos de presença, como Outposts, regiões locais e Wavelength para reduzir a latência da rede ou melhorar o throughput.
- **Otimize 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. Meça o impacto dessas mudanças e use as medições de impacto para tomar decisões futuras.

Análise

PERF 6 Como você aprimora sua carga de trabalho para aproveitar novas versões?

As opções de arquitetura de carga de trabalho são limitadas. No entanto, ao longo do tempo novas tecnologias e abordagens ficam disponíveis e podem aprimorar a performance de sua carga de trabalho.

Melhores práticas:

- **Mantenha-se atualizado sobre novos recursos e serviços**: Avalie maneiras de aumentar a performance conforme surgem novos serviços, padrões de design e ofertas de produto. Determine quais deles poderiam aprimorar a performance ou aumentar a eficiência da carga de trabalho por meio de avaliações ad hoc, discussões internas ou análises externas.
- **Defina um processo para melhorar a performance da carga de trabalho**: Defina um processo para avaliar novos serviços, padrões de design, tipos de recursos e configurações conforme ficarem disponíveis. Por exemplo, execute testes de performance existentes em novas ofertas de instância para determinar o potencial delas de aprimorar sua carga de trabalho.
- **Aprimore a performance da carga de trabalho ao longo do tempo**: Como uma organização, use as informações coletadas por meio do processo de avaliação para promover ativamente a adoção de novos serviços ou recursos quando eles ficarem disponíveis.

Monitoramento

PERF 7 Como você monitora seus recursos para garantir que eles estejam funcionando?

A performance do sistema pode diminuir com o tempo. Monitore a performance do sistema para identificar degradações e corrigir fatores internos ou externos, como a carga do aplicativo ou o sistema operacional.

Melhores práticas:

- **Registrar métricas relacionadas à performance**: Use um serviço de monitoramento e observação para registrar métricas relacionadas à performance. Por exemplo, registre transações do banco de dados, consultas lentas, latência de E/S, taxa de transferência de solicitação HTTP, latência de serviço ou outros dados importantes.
- **Analisar as métricas quando ocorrem eventos ou incidentes**: Em resposta a (ou durante) um evento ou incidente, use painéis ou relatórios de monitoramento para entender e diagnosticar o impacto. Essas visualizações fornecem insights sobre quais partes da carga de trabalho não estão apresentando os níveis de performance esperados.
- **Estabelecer indicadores-chave de performance (KPIs) para medir a performance da carga de trabalho**: Identifique os KPIs que indicam se a performance da carga de trabalho está de acordo com o esperado. Por exemplo, uma carga de trabalho baseada em API pode usar latência de resposta geral como uma indicação da performance geral, e um site de comércio eletrônico pode optar por usar o número de compras efetuadas como KPI.
- **Usar monitoramento para gerar notificações baseadas em alarme**: Usando os indicadores-chave de performance (KPIs) relacionados à performance que você definiu, use um sistema de monitoramento que gere alarmes automaticamente quando essas medidas estiverem fora dos limites esperados.
- **Analisar as métricas regularmente**: Como manutenção de rotina, ou em resposta a eventos ou incidentes, analise as métricas que são coletadas. Use essas análises para identificar quais métricas foram essenciais para lidar com problemas e quais métricas adicionais ajudariam a identificar, resolver ou prevenir problemas caso estivessem sendo acompanhadas.

- **Monitorar e emitir alarmes de maneira proativa**: Use os indicadores-chave de performance (KPIs), aliados a sistemas de monitoramento e alerta, para abordar proativamente problemas relacionados à performance. 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.

Concessões

PERF 8 Como você usa concessões para melhorar a performance?

Ao elaborar soluções, determinar as concessões permite que você selecione uma abordagem ideal. Muitas vezes, você pode aumentar a performance trocando consistência, durabilidade e espaço por tempo e latência.

Melhores práticas:

- **Entenda as áreas em que a performance é mais importante**: Entenda e identifique áreas em que aumentar a performance de sua carga de trabalho 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 serviços de borda para aproximar a entrega de conteúdo dos clientes.
- **Aprenda sobre serviços e padrões de design**: Pesquise e entenda os vários padrões de design e serviços que ajudam a aumentar a performance da carga de trabalho. Como parte da análise, identifique o que você poderia dispensar para obter maior performance. Por exemplo, o uso de um serviço de cache pode ajudar a reduzir a carga imposta sobre sistemas de banco de dados; no entanto, isso requer alguma engenharia para implementar o armazenamento seguro em cache ou a possível introdução de consistência eventual em algumas áreas.
- **Identifique como as concessões afetam os clientes e a eficiência**: Ao avaliar melhorias relacionadas à performance, determine quais escolhas afetarão os clientes e a eficiência da carga de trabalho. Por exemplo, se o uso de um repositório de dados de chave-valor aumentar a performance do sistema, é importante avaliar como a natureza eventualmente consistente dele afetará os clientes.
- **Meça o impacto de melhorias de performance**: À medida que as alterações são feitas para melhorar a performance, avalie as métricas e os dados coletados. Use essas informações para determinar o impacto que o aprimoramento de performance teve sobre a carga de trabalho, os componentes da carga de trabalho e seus clientes. Essa medição ajuda a entender os aprimoramentos resultantes da concessão e a determinar se houve a introdução de algum efeito colateral negativo.
- **Use várias estratégias relacionadas à performance**: Quando aplicável, utilize várias estratégias para aumentar a performance. Por exemplo, o uso de estratégias como armazenar dados em cache para prevenir chamadas excessivas à rede ou ao banco de dados, o uso de réplicas de leitura para mecanismos de banco de dados visando aprimorar as taxas de lei-

tura, a fragmentação ou compactação de dados (quando possível) para reduzir os volumes de dados e o armazenamento em buffer e o streaming dos resultados conforme eles ficam disponíveis para evitar bloqueios.

Otimização de custos

Praticar o gerenciamento financeiro na nuvem

COST 1 Como implementar o gerenciamento financeiro na nuvem?

A implementação da gestão financeira na nuvem permite que as organizações obtenham valor empresarial e sucesso financeiro à medida que otimizam o custo, o uso e a escala na AWS.

Melhores práticas:

- **Estabelecer uma função de otimização de custos**: Crie uma equipe responsável por estabelecer e manter o reconhecimento de custos em toda a organização. A equipe exige pessoas de funções financeiras, de tecnologia e de negócios em toda a organização.
- **Estabelecer uma parceria entre finanças e tecnologia**: Envolva equipes financeiras e de tecnologia em discussões sobre custo e uso em todas as etapas da jornada para a nuvem. As equipes se reúnem e discutem regularmente assuntos como objetivos e metas organizacionais, o estado atual de custo e uso e práticas financeiras e contábeis.
- **Estabelecer previsões e orçamentos de nuvem**: Ajuste os processos de previsão e orçamento organizacional existentes para que sejam compatíveis com a natureza altamente variável dos custos e uso da nuvem. Os processos devem ser dinâmicos, usando algoritmos baseados em tendências ou em motivadores de negócios ou uma combinação deles.
- **Implementar o reconhecimento de custos em seus processos organizacionais**: Implemente o reconhecimento de custos em processos novos ou existentes que afetem o uso e aproveite os processos existentes para reconhecimento de custos. Implemente o reconhecimento de custos no treinamento de funcionários.
- **Relatar e notificar sobre a otimização de custos**: Configure os Orçamentos da AWS para fornecer notificações sobre custos e usos em relação às metas. Realize reuniões regulares para analisar a eficiência de custos dessa carga de trabalho e promover a cultura que reconhece os custos.
- **Monitorar custos proativamente**: Implemente ferramentas e painéis para monitorar os custos proativamente para a carga de trabalho. Não analise apenas os custos e as categorias ao receber notificações. Isso ajuda a identificar tendências positivas e promovê-las em toda a organização.
- **Manter-se atualizado com os novos lançamentos de serviço**: Consulte regularmente especialistas ou parceiros do APN para considerar quais serviços e recursos oferecem menor custo. Analise os blogs da AWS e outras fontes de informação.

Reconhecimento de despesas e usos

COST 2 Como você governa o uso?

Estabeleça políticas e mecanismos para garantir que os custos adequados sejam gerados enquanto os objetivos são alcançados. Ao empregar uma abordagem de verificação e equilíbrio, você pode inovar sem gastar demais.

Melhores práticas:

- **Desenvolver políticas baseadas nos requisitos da organização**: Desenvolva políticas que definam como os recursos são gerenciados por sua organização. As políticas devem cobrir aspectos de custos de recursos e cargas de trabalho, incluindo criação, modificação e desativação ao longo da vida útil do recurso.
- **Implementar objetivos e metas**: Implemente metas de custo e uso para sua carga de trabalho. As metas fornecem orientação para sua organização quanto ao custo e uso, e os objetivos oferecem resultados mensuráveis para suas cargas de trabalho.
- **Implementar uma estrutura de conta**: Implemente uma estrutura de contas que mapeie para sua organização. Isso auxilia na alocação e no gerenciamento de custos em toda a organização.
- **Implementar grupos e funções**: Implemente grupos e funções que se alinhem com as políticas e controle quem pode criar, modificar ou desativar instâncias e recursos em cada grupo. Por exemplo, implemente grupos de desenvolvimento, teste e produção. Isso se aplica aos serviços da AWS e às soluções de terceiros.
- **Implementar controles de custos**: Implemente controles baseados nas políticas da organização e nas funções e grupos definidos. Isso garante que os custos sejam gerados somente como definido pelos requisitos da organização: por exemplo, controle o acesso a regiões ou tipos de recursos com políticas de IAM.
- **Acompanhar o ciclo de vida do projeto**: Acompanhe, meça e realize auditorias no ciclo de vida dos projetos, equipes e ambientes para evitar o uso e pagamento de recursos desnecessários.

COST 3 Como você monitora o uso e os custos?

Estabeleça políticas e procedimentos para monitorar e alocar adequadamente os custos. Isso permite medir e aprimorar a eficiência de custos dessa carga de trabalho.

Melhores práticas:

- **Configurar fontes de informações detalhadas**: Configure o Relatório de custos e uso da AWS e a granularidade por hora do Cost Explorer para fornecer informações detalhadas de custos e uso. Configure sua carga de trabalho para ter entradas de log para cada resultado comercial entregue.
- **Identificar categorias de atribuição de custos**: Identifique as categorias de organização que podem ser usadas para alocar custos dentro da organização.

- **Estabelecer métricas da organização**: Estabeleça as métricas da organização que são necessárias para esta carga de trabalho. Exemplo de métricas de uma carga de trabalho são relatórios de clientes produzidos ou páginas da Web veiculadas aos clientes.
- **Configurar as ferramentas de faturamento e gerenciamento de custos**: Configure o AWS Cost Explorer e o Orçamentos da AWS de acordo com as políticas da organização.
- **Adicionar informações da organização ao custo e ao uso**: Defina um esquema de marcação baseado na organização, nos atributos da carga de trabalho e nas categorias de alocação de custos. Implemente a marcação em todos os recursos. Use o Cost Categories para agrupar custos e uso de acordo com atributos da organização.
- **Alocar custos baseados nas métricas de trabalho**: Aloque os custos da carga de trabalho por métricas ou resultados de negócios para medir a eficiência de custos da carga de trabalho. Implemente um processo para analisar o Relatório de custos e uso da AWS com o Amazon Athena, que pode fornecer informações e recurso de cobrança retroativa.

COST 4 Como você desativa os recursos?

Implemente o controle de alterações e o gerenciamento de recursos, desde o início do projeto até o fim da vida útil. Isso garante o desligamento ou encerramento dos recursos não utilizados para reduzir o desperdício.

Melhores práticas:

- **Acompanhar os recursos ao longo da vida útil**: Defina e implemente um método para acompanhar os recursos e as associações com sistemas ao longo da vida útil. Você pode usar a marcação para identificar a carga de trabalho ou a função do recurso.
- **Implementar um processo de desativação**: Implemente um processo para identificar e desativar recursos órfãos.
- **Desativar recursos**: Desative recursos acionados por eventos, como auditorias periódicas ou alterações no uso. Normalmente, a desativação é realizada periodicamente e é manual ou automatizada.
- **Desativar recursos automaticamente**: Projete a carga de trabalho para lidar normalmente com o encerramento de recursos ao identificar e desativar recursos não críticos, que não são necessários ou com baixa utilização.

Recursos econômicos

COST 5 Como você avalia o custo ao selecionar serviços?

O Amazon EC2, Amazon EBS e Amazon S3 são produtos fundamentais da AWS. Os produtos gerenciados, como Amazon RDS e Amazon DynamoDB, são produtos da AWS de nível superior ou de aplicativo. Ao selecionar os produtos fundamentais e os serviços gerenciados adequados, você pode otimizar os custos dessa carga de trabalho. Por exemplo, usando serviços gerenciados, é possível reduzir ou remover grande parte da sobrecarga administrativa e operacional, liberando você para trabalhar em aplicativos e atividades relacionadas a negócios.

Melhores práticas:

- **Identificar requisitos da organização para custos**: Trabalhe com os membros da equipe para definir o equilíbrio entre otimização de custos e outros pilares, como performance e confiabilidade, para essa carga de trabalho.
- **Analisar todos os componentes dessa carga de trabalho**: Garanta que todos os componentes da carga de trabalho sejam analisados, independentemente do tamanho atual ou dos custos atuais. O esforço da análise deve refletir o benefício potencial, como custos atuais e projetados.
- **Executar uma análise completa de cada componente**: Observe o custo geral para a organização de cada componente. Observe o custo total de propriedade considerando o custo de operações e gerenciamento, especialmente ao usar serviços gerenciados. O esforço de análise deve refletir o benefício potencial; por exemplo, o tempo gasto na análise é proporcional ao custo do componente.
- **Selecionar software com licenciamento econômico**: O software de código aberto eliminará os custos de licenciamento de software, o que pode contribuir com custos significativos para as cargas de trabalho. Quando for necessário um software licenciado, evite licenças vinculadas a atributos arbitrários, como CPUs, e procure aquelas que estejam vinculadas à saída ou aos resultados. O custo dessas licenças é mais próximo do benefício que elas oferecem.
- **Selecionar os componentes dessa carga de trabalho para otimizar o custo alinhado com as prioridades da organização**: Considere o custo ao selecionar todos os componentes. Isso inclui o uso de nível de aplicativo e serviços gerenciados, como o Amazon RDS, Amazon DynamoDB, Amazon SNS e Amazon SES, para reduzir o custo geral da organização. Use serviços de contêineres e sem servidor para computação, como o AWS Lambda, Amazon S3 para sites estáticos e Amazon ECS. Minimize os custos de licença usando software de código aberto ou software sem taxas de licença: por exemplo, Amazon Linux para cargas de trabalho de computação ou migração de bancos de dados para o Amazon Aurora.
- **Realizar análises de custos para diferentes usos ao longo do tempo**: As cargas de trabalho podem mudar ao longo do tempo. Alguns serviços ou recursos são mais econômicos em diferentes níveis de uso. Ao executar a análise em cada componente ao longo do tempo e no uso projetado, você garante que essa carga de trabalho permaneça econômica ao longo da vida útil.

COST 6 Como você atinge as metas de custo ao selecionar tamanho, número e tipo de recurso?

Escolha o tamanho e o número de recursos apropriados para a tarefa em mãos. Ao selecionar o tipo, tamanho e número mais econômicos, você minimiza o desperdício.

Melhores práticas:

- **Executar modelagem de custos**: Identifique os requisitos da organização e execute a modelagem de custos da carga de trabalho e de cada um dos componentes. Realize atividades de referência para a carga de trabalho sob diferentes cargas previstas e compare os custos. O esforço de modelagem deve refletir o benefício potencial. Por exemplo, o tempo gasto é proporcional ao custo do componente.

- **Selecionar tipo e tamanho do recurso com base nos dados**: Selecione o tamanho ou tipo de recurso com base nos dados sobre a carga de trabalho e nas características do recurso. Por exemplo, computação, memória, throughput ou gravação intensiva. Essa seleção geralmente é feita usando uma versão anterior da carga de trabalho (como uma versão no local), a documentação ou outras fontes de informações sobre a carga de trabalho.
- **Selecionar o tipo e o tamanho do recurso automaticamente com base nas métricas**: Use métricas da carga de trabalho em execução no momento para selecionar o tamanho e o tipo certos para otimizar o custo. Forneça adequadamente throughput, dimensionamento e armazenamento para serviços como Amazon EC2, Amazon DynamoDB, Amazon EBS (PIOPS), Amazon RDS, Amazon EMR e redes. Isso pode ser feito com um ciclo de comentários, como escalabilidade automática ou por código personalizado na carga de trabalho.

COST 7 Como você usa os modelos de definição de preço para reduzir custos?

Use o modelo de definição de preço mais adequado nos recursos para minimizar as despesas.

Melhores práticas:

- **Executar análise de modelo de definição de preço**: Analise cada componente da carga de trabalho. Determine se o componente e os recursos serão executados por períodos estendidos (para descontos de compromisso) ou dinâmicos e curtos (para spot ou sob demanda). Execute uma análise da carga de trabalho usando o recurso Recomendações no AWS Cost Explorer.
- **Implementar regiões com base nos custos**: A definição de preço dos recursos pode ser diferente em cada região. A consideração do custo da região garante que você pague o menor preço geral por essa carga de trabalho
- **Selecionar contratos de terceiros com termos econômicos**: Acordos e termos econômicos garantem que o custo desses serviços seja dimensionado de acordo com os benefícios oferecidos. Selecione contratos e definição de preço que escalem quando oferecerem benefícios adicionais à sua organização.
- **Implementar modelos de definição de preço para todos os componentes dessa carga de trabalho**: Os recursos em execução permanente devem utilizar capacidade reservada, como Savings Plans ou instâncias reservadas. A capacidade de curto prazo está configurada para usar instâncias spot ou frota spot. A demanda é usada somente para cargas de trabalho de curto prazo que não podem ser interrompidas e não executam o tempo suficiente para a capacidade reservada, entre 25 e 75% do período, dependendo do tipo de recurso.
- **Executar a análise do modelo de definição de preço no nível da conta mestre**: Use recomendações de instâncias reservadas e Savings Plans do Cost Explorer para executar análises regulares no nível da conta mestre e obter descontos de compromisso.

COST 8 Como você planeja as cobranças de transferência de dados?

Certifique-se de planejar e monitorar as cobranças de transferência de dados para tomar decisões de arquitetura que minimizam custos. Uma mudança arquitetônica pequena, porém eficaz, pode reduzir drasticamente os custos operacionais ao longo do tempo.

Melhores práticas:

- **Executar modelagem de transferência de dados**: Reúna os requisitos da organização e execute a modelagem de transferência de dados da carga de trabalho e de cada um dos componentes. Isso identifica o menor ponto de custo para os requisitos atuais de transferência de dados.
- **Selecionar componentes para otimizar o custo de transferência de dados**: Todos os componentes são selecionados, e a arquitetura é projetada para reduzir os custos de transferência de dados. Isso inclui o uso de componentes como otimização de WAN e configurações de Multi-AZ
- **Implementar serviços para reduzir custos de transferência de dados**: Implemente serviços para reduzir a transferência de dados. Por exemplo, usar uma CDN como o Amazon CloudFront para fornecer conteúdo aos usuários finais, armazenar em cache camadas usando o Amazon ElastiCache ou usar o AWS Direct Connect em vez da VPN para conectividade com a AWS.

Gerenciar recursos de demanda e fornecimento

COST 9 Como você gerencia a demanda e fornece recursos?

Para uma carga de trabalho que tenha gasto e performance equilibrados, verifique se tudo o que você paga é usado e evite instâncias significativamente subutilizadas. Uma métrica de utilização distorcida tem um impacto adverso na organização, nos custos operacionais (performance degradada devido à superutilização) ou nos gastos da AWS (devido ao excesso de provisionamento).

Melhores práticas:

- **Executar uma análise sobre a demanda de carga de trabalho**: Analise a demanda da carga de trabalho ao longo do tempo. Garanta que a análise cubra tendências sazonais e represente com precisão as condições operacionais durante toda a vida útil da carga de trabalho. O esforço de análise deve refletir o benefício potencial. Por exemplo, se o tempo gasto é proporcional ao custo da carga de trabalho.
- **Implementar um buffer ou controle de utilização para gerenciar a demanda**: O armazenamento em buffer e o controle de utilização modificam a demanda na carga de trabalho, suavizando todos os picos. Implemente o controle de utilização quando seus clientes realizarem novas tentativas. Implemente o armazenamento em buffer para armazenar a solicitação e adiar o processamento até um momento posterior. Os controles de utilização e buffers devem ser projetados para que os clientes recebam uma resposta no tempo necessário.
- **Fornecer recursos dinamicamente**: Os recursos são provisionados de maneira planejada. Isso pode ser baseado na demanda, como por meio da escalabilidade automática, ou no tempo, em que a demanda é previsível e os recursos são fornecidos com base no tempo. Esses métodos resultam na menor quantidade de sobreprovisionamento ou subprovisionamento.

Otimizar ao longo do tempo

COST 10 Como você avalia os novos serviços?

Como a AWS lança novos serviços e recursos, faz parte das melhores práticas analisar as decisões de arquitetura existentes para garantir que elas continuem sendo as mais econômicas.

Melhores práticas:

- **Desenvolver um processo de análise da carga de trabalho**: Desenvolva um processo que defina os critérios e o processo para análise da carga de trabalho. O esforço de análise deve refletir o benefício potencial: por exemplo, cargas de trabalho principais ou cargas de trabalho com valor superior a 10% da fatura são analisadas trimestralmente, enquanto cargas de trabalho abaixo de 10% são analisadas anualmente.
- **Revise e analise essa carga de trabalho regularmente**: As cargas de trabalho existentes são analisadas regularmente de acordo com os processos definidos.